Genomic Biosurveillance for Insect Pests

Submitted by

Alexander Medway Piper

Bachelor of Science (Biology)

A thesis submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

School of Applied Systems Biology
College of Science, Health and Engineering
La Trobe University
Victoria, Australia

March 2021

Table of contents

List of Abbreviations

2DSFS	Two-Dimensional Site Frequency Spectrum
ACV	Apple Cider Vinegar
ALR	Additive Log-Ratio
ANOVA	Analysis of Variance
ASV	Amplicon Sequence Variant
AUD	Australian Dollar
AUS	Australia
BLAST	Basic Local Alignment Search Tool
BOLD	Barcode of Life Data System
bp	Base pairs
BQSR	Base Quality Score Recalibration
BSA	Bovine Serum Albumin
CI	Confidence Interval
CLR	Centred Log-Ratio
CODA	Compositional Data Analysis
COI	Cytochrome c oxidase subunit I
Contig	Contiguous sequence
CV	Cross Validation
D	Tajima's D test for neutrality
ddPCR	Digital droplet Polymerase Chain Reaction
DNA	Deoxyribonucleic Acid
EDRR	Early Detection Rapid Response
EM	Expectation–Maximization
FAO	United Nations Food and Agriculture Organization
FF	Fruit crush and Floatation
FST	Fixation Index
Gbp / Gb	Gigabase pairs
GTR	General Time Reversible
H_E	Heterozygosity
HTS	High-Throughput Sequencing
HWE	Hardy-Weinberg Equilibrium
IBD	Isolation by Distance
ID	Identification
INDEL	Insertion or Deletion
IPM	Integrated Pest Management
IPPC	International Plant Protection Convention
Kip / kb	Kilobase pairs
LAMP	Loop-mediated Isothermal Amplification
LASSO	Least Absolute Shrinkage and Selection Operator
lcWGS	Low coverage Whole Genome Sequencing
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MAP	Minimum Average Partial
Mbp / Mb	Megabase pairs

NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
NSTI	Nearest Sequenced Taxon Index
NSW	New South Wales
NT	Northern Territory
NZ	New Zealand
OIE	Office International des Epizooties (World Organisation of Animal Health)
ONT	Oxford Nanopore Technologies
OTT	Open Tree of Life Taxonomy
OTU	Operational Taxonomic Unit
PacBio	Pacific Biosciences
PC / PCA	Principal Component / Principal Component Analysis
PCR	Polymerase Chain Reaction
PERMANOVA	Permutational Multivariate Analysis of Variance
PFA	Pest Free Area
PHMM	Profile Hidden Markov Model
QLD	Queensland
qPCR	Quantitative PCR
RFLP	Restriction Fragment Length Polymorphism
RMSE	Root Mean Squared Error
RNA	Ribonucleic Acid
rRNA	Ribosomal Ribonucleic Acid
SFS	Site Frequency Spectrum
SNP	Single Nucleotide Polymorphism
SPD	Synthetic lure + Propylene glycol + Dichlorvos insecticide
SPS	Agreement on the Application of Sanitary and Phytosanitary measures
SRA	Sequence Read Archive
SVM	Support Vector Machine
Syn	Synthetic lure
TAS	Tasmania
USA	United States of America
USD	United States Dollar
VCF	Variant Call Format
VIC	Victoria
WA	Western Australia
WGS	Whole Genome Sequencing
WTO	World Trade Organisation
zOTU	zero-radius Operational Taxonomic Unit
θ_w	Total number of segregating sites
θ_π	Average number of pairwise differences between sequences

Abstract

Biosurveillance systems aiming to detect and monitor populations of insect pests often employ traps that also catch a wide range of bycatch species. The resulting mixed specimens require extensive sorting by taxonomic experts before target pests can be identified, creating a major diagnostic bottleneck. This thesis explores how genomic techniques can be used to rapidly identify multiple species within unsorted trap samples, and then trace the geographic origins of detected populations, thereby increasing the scale and resolution at which insect biosurveillance can be conducted. First, short subregions of the cytochrome oxidase subunit I (COI) barcode were compared in-silico for their ability to act as broad-spectrum diagnostic markers for invasive insect pests. Second, four high performing mini-barcodes were applied in a non-destructive metabarcoding assay aimed at detecting spotted wing drosophila (*Drosophila suzukii*), a high priority exotic pest for Australia. In accordance with in-silico predictions, metabarcoding successfully detected *D. suzukii* and its close relatives spiked into mixed trap samples. Both field collection and DNA extraction protocols will, however, require optimisation to minimise sample and replicate dropouts. Third, the use of predictive models to correct for taxonomic bias inherent to metabarcoding assays was evaluated in order to expand their scope to quantitative population monitoring. All six evaluated models significantly improved the correlation between expected and observed relative abundances; and results could be transformed back to counts of insects using an independent measurement of absolute abundance, providing benefits for interpretability. Finally, low-coverage whole genome resequencing was used to compare trapped Queensland fruit fly (*Bactrocera tryoni*) specimens from recent outbreaks against a genomic reference panel of endemic populations. Despite weak concordance between genetic and geographic structure, outbreak specimens were successfully assigned to major populations, ruling out certain introduction pathways. These findings demonstrate how genomic biosurveillance can enhance management response to invasive insect pests, and practical integration into surveillance programmes is discussed.

Statement of Authorship

This thesis includes work by the author that has been published or accepted for publication as described in the text. Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis accepted for the award of any other degree or diploma. No other person's work has been used without due acknowledgment in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution

Alexander Medway Piper

31/03/2021

Scholarship acknowledgement

This work was supported by an Australian Government Research Training Program Scholarship. A complement to this scholarship was provided by Agriculture Victoria Research through the Improved Market Access for Horticulture program.

Acknowledgements

Throughout the writing of this thesis I have received a great deal of support and assistance from friends and colleagues. First and foremost, I would like to thank my principal supervisor Paul Cunningham for sculpting the mind of a young lad from regional Queensland into the scientist I am today, for treating me as a colleague from the very beginning, and always keeping my thinking sharp. I must also thank my co-supervisor Noel Cogan for his incredible generosity in supervision, and for providing the constant push to do it bigger and better. I extend this thanks to my other co-supervisor Mark Blacket for his patient guidance, encouragement, and complete dependability whenever I needed it.

I have had the great pleasure of working alongside many supportive colleagues in the Invertebrate and Weed Sciences research group and wider community at AgriBio. Completing this research would not have been possible without the assistance of Rebecca Baillie, Preeti Thakur, and Michelle Drayton who dealt with many naïve questions about genomic sequencing protocols, Jessica Henneken, Farrukh Baig, and Shahidul Islam who assisted with field work, and Lea Rako, Linda Semeraro, and Isabel Valenzuela who provided taxonomic support throughout. Special thanks must also go to Jana Batovska for being a pioneer and constant inspiration, Kevin Farnier for his ongoing moral support, and Elisse Nogarotto for her guidance in all thing's lab and life. I am also grateful to Francesco Martoni, Conrad Trollip, and Brendan Rodoni for many constructive discussions around sequencing and biosecurity, as well as David Williams for sharing his unparalleled knowledge of insect pest management. I extend this gratitude to Agriculture Victoria Research as an organisation for providing such incredible resources and stimulating environment to pursue research in.

I wish to thank all the friends I have made in my adopted home of Melbourne for their company throughout this journey, in particular Ben van Dorsselaer, to whom I owe all remaining sanity after the experience of writing up a thesis in a global pandemic. Finally, I'd like to thank my parents Susan and Terry, and my brothers Toby and Arlo for their unwavering support and encouragement for everything I have ever done.

Thesis Preface

This thesis is comprised of seven chapters, with the original experimental content formatted as self-contained manuscripts either published, submitted, or in preparation to submit to peer-reviewed scientific journals. The first chapter provides a general introduction to biological invasions, biosecurity, and insect biosurveillance. The second chapter consists of an in-depth review of the relevant literature, focussing on the application of metabarcoding to diagnostics of invasive insect pests. The following four chapters detail the original experimental work, with each containing its own separate introduction, methodology, results, and discussion sections. Each experimental chapter includes a preface which describes how the research links to the other chapters and lists the publication details, including a statement from a co-author confirming the contribution of the PhD candidate. The final chapter comprises a general discussion which integrates the major themes from each experimental chapter and identifies avenues for future research. As chapters 2-6 correspond to separate scientific manuscripts, some redundancy of content has arisen between the introduction and methods sections of each respective chapter. In the case of published or submitted manuscripts, each employs the distinct referencing and citation styles of the corresponding journal, with the bibliography and any supplementary material included at the end of the chapter. The remaining chapters employ a single referencing and citation style with the bibliography provided at the end of each chapter.

1
General Introduction

1.1 Biological invasions & biosecurity

Biological invasions have become symptomatic of our increasingly globalised world, as an unintended consequence of international trade and tourism, agricultural and urban development, and changing climates (Chown et al., 2015; Elton, 1958; Mack et al., 2000). Insects form a dominant component of the global spread of invasive species, with more than 4,900 species listed as invasive worldwide (Seebens et al., 2018). To date, the economic impact of insect invasions has primarily been considered through the lens of agricultural production, where the combined global costs of forgone output, pest control efforts, and loss of trade opportunities run into the tens of billions annually (Bradshaw et al., 2016). In Australia, new incursions of invasive insects threaten the productivity of a $14.4 billion horticultural sector (Hort Innovation, 2020), as well as the enviable market access position attained through historical freedom from many of the world's major pests and diseases (Beale et al., 2008). Beyond agroecosystems, invasive insects can have pervasive effects on the natural environment, altering food webs and outcompeting endemic species (Ehrenfeld, 2010; Kenis et al., 2009). In turn, both the newly introduced species and the practices required to control it impact the goods and services provided by ecosystems as well as the communities that depend upon them for recreation, cultural practices, and human amenity (Binimelis et al., 2007; Kenis et al., 2009; Pejchar & Mooney, 2009). Ultimately, this highlights the importance of preventing the introduction and spread of biological invaders to protect economic prosperity, food security, and human wellbeing.

Biosecurity

Biosecurity is a multidisciplinary field that encompasses the use of science, policy, and regulation to protect agriculture, food, and the environment from biological risk (FAO, 2003). This rebranding of the centuries-old practice of controlling pests and diseases is a direct reflection of globalisation and the required shift towards nation interdependence for managing introduction pathways (Hulme, 2011; Waage & Mumford, 2008). A primary aim of biosecurity programmes is to prevent the introduction of new pests or diseases in the first place, through a combination of risk analysis, regulatory measures, and quarantine inspection (Epanchin-Niell & Liebhold, 2015; Leung et al., 2002). This follows the precautionary principal, meaning a lack of scientific certainty about the risk posed by a potentially invasive species should not be used as a reason for not taking preventative action against its introduction (Cooney, 2004). When preventing introduction fails, eradication may be an option as long as incipient populations remain relatively small and localised (Liebhold et al., 2016; Pluess et al., 2012). If, however, the introduced species is detected too late for eradication to be successful, populations may be suppressed on a long-term basis in order to minimise impact on production (Kogan, 1988; Stenberg, 2017), but the return on investment quickly diminishes the longer a pest has had time to establish (Figure 1) (Finnoff et al., 2007; Leung et al., 2002; Rout et al., 2011). Due to the critical importance of detecting pests as early as possible, effective surveillance forms a key component of modern biosecurity programmes (Kalaris et al., 2014; Quinlan et al., 2015).

Biosecurity surveillance

Biosecurity surveillance covers a continuum of pre-border, at-border, and post-border activities (Beale et al., 2008; Kalaris et al., 2014). Pre-border surveillance includes monitoring of international outbreaks and conducting risk assessments for newly emerging threats (Andersen et al., 2004; MacLeod, 2015). At-border surveillance involves quarantine inspection of commodities and baggage, generally targeted towards identifying those pests highlighted by prior risk assessment (Martin et al., 2016; Whattam et al., 2014). Post-border surveillance is considered the last line of defence, aiming to detect newly-introduced populations as early as possible to increase the likelihood of successful eradication (Reaser et al., 2020; Sharma et al., 2014). This is commonly achieved

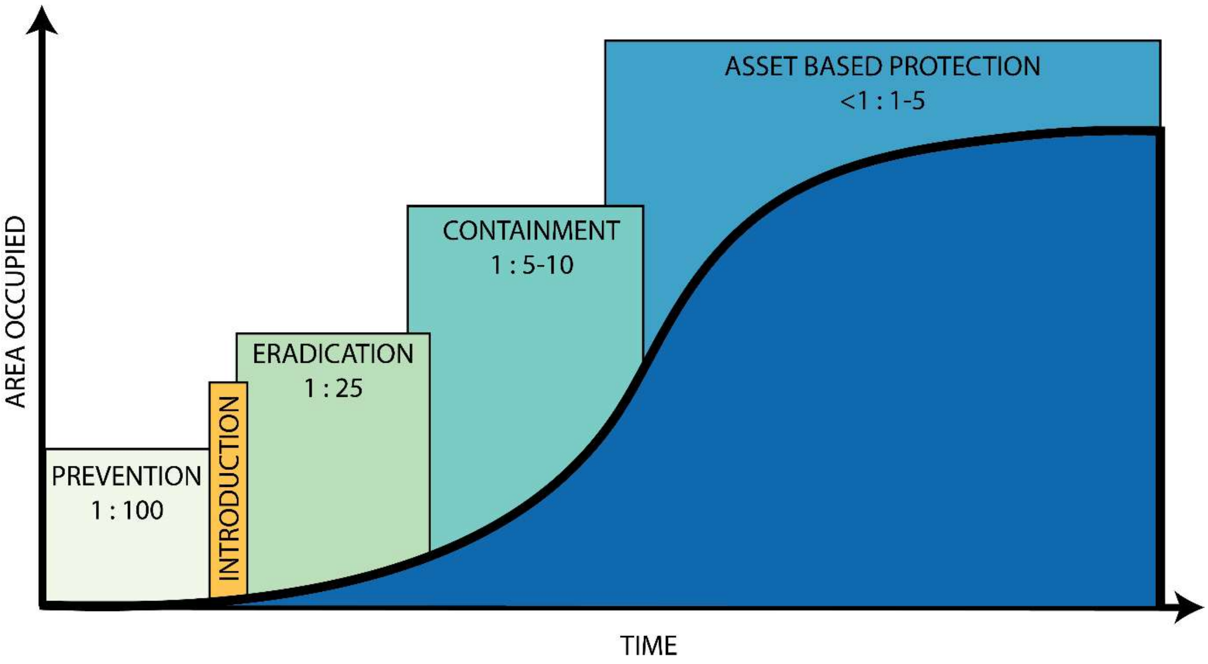


Figure 1: Generalised invasion curve, with the economic returns associated with each stage highlighted (indicative only). Adapted from Invasive Plants and Animals Policy Framework, State of Victoria, Department of Primary Industries, 2010

through a mixture of targeted (active) surveillance activities to detect or demonstrate absence of a high priority pest (Low-Choy, 2015), and more general (passive) surveillance that leverages public awareness, reports of pest symptoms, and biodiversity surveys conducted by researchers and natural resource managers (Bishop & Hutchings, 2011; F. C. Jarrad et al., 2011; Thomas et al., 2017). In Australia, post-border biosecurity surveillance is coordinated and conducted by various national, state, and industry organisations, depending on the geographic scale and relevant jurisdictions covered (Anderson et al., 2017). Surveillance programmes targeting insect pests generally employ traps containing targeted pheromone lures (Witzgall et al., 2010), host semiochemicals (e.g. Cha et al., 2014; Cunningham, Carlsson, Villa, Dekker, & Clarke, 2016), or simply relying upon wind and insect flight (Hardulak et al., 2020). Depending on the selectivity of the lure and the environment they are deployed in, surveillance traps can collect just a few specimens of a single species, or more commonly, hundreds of mixed specimens from a broad range of species (Batovska et al., 2018, 2020; Spears & Ramirez, 2015). In the latter case, detection of a newly introduced species requires it to first be located and identified within the mixed trap (Boykin, Armstrong, Kubatko, & De Barro, 2012), which in itself presents a major bottleneck to the design and implementation of cost-effective surveillance.

Biosecurity diagnostics

While visual morphological examination has long been the routine for insect identification, the microscope has recently been supplemented with a varied toolbox of molecular assays that allow standardised identification of diverse taxa without requiring specialist taxonomic expertise (Roe et al., 2019). As quarantine and regulatory decisions are often based on species names (Boykin, Armstrong, Kubatko, & de Barro, 2012), these assays provide species-level identification either through targeting distinct mutational signatures (Kim et al., 2016), or by comparing the molecular variation contained within a conserved gene against a reference database (Armstrong & Ball, 2005). Sometimes, however, a species name is insufficient: for example, when testing for virulent biotypes (Herbert et al., 2010), pesticide resistant populations (Van Leeuwen et al., 2020), or when tracing the geographic source of an intercepted specimen (Barr et al., 2014). In these cases, molecular methods are often the only approach for obtaining the required intra-specific data, historically involving 'sanger' sequencing of short strands of DNA (Sanger et al., 1977), or by comparing physical size differences in simple sequence repeats or 'microsatellites' (Goldstein & Pollock, 1997). While these traditional techniques have provided valuable insights into the identity and structure of invasive populations (Darling & Blum, 2007; Kirk et al., 2013), they only yield polymorphism information for a small fraction of the genome and are now largely being replaced by genomic datasets provided by High-Throughput Sequencing (HTS) technologies (McCartney et al., 2019; North et al., 2021; Tay & Gordon, 2019).

1.2 Genomic biosurveillance

Genomic biosurveillance involves the use of genomic datasets to investigate the identity, spread, and evolutionary dynamics of invasive pests and pathogens at a fine spatial and temporal scale (Bilodeau et al., 2019; Hamelin & Roe, 2020; Roe et al., 2019). While the term is relatively new, the concept has its roots in genomic epidemiology of human pathogens (Achidi et al., 2008; Gardy & Loman, 2018), a field which has recently entered the broader public consciousness as a result of the global COVID19 pandemic (Lu et al., 2020). Similar to an infectious disease, the process of invasion by an insect pest consists of four major steps; transport, introduction, establishment and spread, each of which can be mitigated by an associated management response (Figure 1). Genomic techniques can contribute to monitoring and management actions at each of these stages through accelerating species identification (Batovska et al., 2018, 2020; Dupuis, Bremer, et al., 2018), tracing introduction pathways (Lee et al., 2019; Schmidt et al., 2021; Tay et al., 2020), and providing information on the demographic processes (Bergey et al., 2020; Wu et al., 2019; You et al., 2020), and genetic architecture underlying successful establishment in a new environment (Calfee et al., 2020; Dupuis, Sim, et al., 2018). While many of these applications represent finer scale investigations of classic questions in invasion biology (Bock et al., 2015), genomic techniques also permit entirely new questions to be addressed. For instance, the ability of HTS platforms to sequence diverse template molecules enables the simultaneous identification of entire communities of native and introduced species (Comtet et al., 2015; Tedersoo et al., 2019), a significant step towards the universal invasive species identification chip envisioned by Darling & Blum, (2007). Genomic biosurveillance may therefore contribute a range of new diagnostic methods to the insect biosecurity toolbox, and provide valuable insights into invasion biology that can be leveraged to better anticipate and respond to future invasions (Poland & Rassati, 2019; Roe et al., 2019).

1.3 Research overview

This thesis applies the concept of genomic biosurveillance to improving the detection and control of invasive insect pests within Australian horticulture. A central aim of the research is to develop and evaluate practical tools for uptake by laboratories conducting insect diagnostics as part of biosecurity surveillance, focussing on two main approaches: (i) the ability for broad-scope HTS assays to simultaneously identify multiple invasive species within mixed trap catches, thereby overcoming the diagnostic bottleneck of morphological specimen sorting; and (ii) the use of genome-wide information to locate the geographic origin of invasive populations and explore patterns of genetic diversity during colonisation and establishment. As diagnostic laboratories commonly operate across a broad scope of invasive insects, the tools developed here are designed to be species-independent, and thus readily applicable to new targets with minimal change in protocol. Each tool is developed and validated on a series of case studies representing high-priority exotic or established pest threats for Australia, serving to demonstrate a flexible genomic biosurveillance pipeline that could be readily expanded to the next emerging threat.

Chapter 2 combines a comprehensive literature review with novel analyses of public sequence data to evaluate the prospects for DNA metabarcoding to act as a universal diagnostic assay for invasive insect pests. This chapter synthesises current knowledge from the metabarcoding literature into a set of procedural best practices, then identifies technical and regulatory challenges which must be overcome before application within the highly regulated field of invasive insect diagnostics.

Chapter 3 aims to determine the taxonomic breadth across which short subregions of COI can achieve species-level resolution and summarise the many published metabarcoding primers into a recommended list for diagnostic use. To achieve this, a large database of public COI reference sequences is curated, then computational methods are used to evaluate the optimal placement, diagnostic sensitivity, and taxonomic bias for 68 published and novel metabarcoding primers.

Chapter 4 then applies these optimal primers within a non-destructive metabarcoding assay aiming to detect *Drosophila suzukii*, a high-priority exotic pest for Australian horticulture, within mixed trap catches. Laboratory and bioinformatic methods appropriate for detecting low abundance specimens are developed; sensitivity, specificity, and overall accuracy of the assay is established; and the required number of technical replicates to ensure robust results is determined.

Chapter 5 assesses whether metabarcoding can provide quantitative measurements to support decision making during pest eradication or suppression efforts. Iterating upon the protocol developed in the previous chapter, six statistical models are evaluated for their ability to correct for taxonomic bias and transform the sequence read data provided by metabarcoding into counts of individual specimens. This approach is then validated on pheromone trapped *Carpophilus* beetles collected within an integrated pest management programme, with the quantitative measurements provided by the bias-corrected metabarcoding assay compared to traditional morphological sorting.

Chapter 6 evaluates the use of genome wide SNP data for tracing the geographic origin of new outbreaks and exploring the patterns of genetic diversity occurring during colonisation and establishment. A low-coverage whole genome sequencing assay is developed and validated on the Queensland fruit fly (*Bactrocera tryoni*), a highly polyphagous pest endemic to Australia but only recently established in the temperate fruit growing regions of Victoria.

Chapter 7 comprises a general discussion that considers the results and findings of each experimental chapter with regards to the broader implications for applied biosecurity and fundamental invasion biology. Practical recommendations are made for integrating genomic approaches into active biosurveillance, and promising avenues for future research are identified.

1.4 Bibliography

Achidi, E. A., Agbenyega, T., Allen, S., Amodu, O., Bojang, K., Conway, D., Corran, P., Deloukas, P., Djimde, A., Dolo, A., Doumbo, O., Drakeley, C., Duffy, P., Dunstan, S., Evans, J., Farrar, J., Fernando, D., Hien, T. T., Horstmann, R., … Gottlieb, M. (2008). A global network for investigating the genomic epidemiology of malaria. *Nature*, 456(7223), 732–737. https://doi.org/10.1038/nature07632

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

Anderson, C., Low-Choy, S., Whittle, P., Taylor, S., Gambley, C., Smith, L., Gillespie, P., Löcker, H., Davis, R., & Dominiak, B. (2017). Australian plant biosecurity surveillance systems. *Crop Protection*, 100, 8–20. https://doi.org/10.1016/j.cropro.2017.05.023

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

Barr, N., Ruiz-Arce, R., & Armstrong, K. (2014). Using molecules to identify the source of fruit fly invasions. In *Trapping And The Detection, Control, and Regulation of Tephritid Fruit Flies: Lures, Area-Wide Programs, and Trade Implications*. https://doi.org/10.1007/978-94-017-9193-9_10

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

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

Beale, R., Fairbrother, J., Inglis, A., & Trebeck, D. (2008). One Biosecurity: A Working Partnership The Independent Review of Australia’s Quarantine and Biosecurity Arrangements Report to the Australian Government. *Commonwealth of Australia, Canberra.*

Bergey, C. M., Lukindu, M., Wiltshire, R. M., Fontaine, M. C., Kayondo, J. K., & Besansky, N. J. (2020). Assessing connectivity despite high diversity in island populations of a malaria mosquito. *Evolutionary Applications*, 13(2), 417–431. https://doi.org/10.1111/eva.12878

Bilodeau, P., Roe, A. D., Bilodeau, G., Blackburn, G. S., Cui, M., Cusson, M., Doucet, D., Griess, V. C., Lafond, V. M. A., Nilausen, C., Paradis, G., Porth, I., Prunier, J., Srivastava, V., Stewart, D., Torson, A. S., Tremblay, E., Uzunovic, A., Yemshanov, D., & Hamelin, R. C. (2019). Biosurveillance of forest insects: part II—adoption of genomic tools by end user communities and barriers to integration. *Journal of Pest Science*, 92(1), 71–82. https://doi.org/10.1007/s10340-018-1001-1

Binimelis, R., Born, W., Monterroso, I., & Rodríguez-Labajos, B. (2007). Socio-Economic Impact and Assessment of Biological Invasions. *Biological Invasions*, 193, 331–347. https://doi.org/10.1007/978-3-540-36920-2_19

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

Bock, D. G., Caseys, C., Cousens, R. D., Hahn, M. A., Heredia, S. M., Hübner, S., Turner, K. G., Whitney, K. D., & Rieseberg, L. H. (2015). What we still don't know about invasion genetics. *Molecular Ecology*, 24(9), 2277–2297. https://doi.org/10.1111/mec.13032

Boykin, L. M., Armstrong, K. F., Kubatko, L., & de Barro, P. (2012). Species delimitation and global biosecurity. *Evolutionary Bioinformatics*, 8, 1–37. https://doi.org/10.4137/EBO.S8532

Boykin, L. M., Armstrong, K., Kubatko, L., & De Barro, P. (2012). DNA barcoding invasive insects: Database roadblocks. *Invertebrate Systematics*, 26, 507–514. https://doi.org/10.1071/IS12025

Bradshaw, C. J. A., Leroy, B., Bellard, C., Roiz, D., Albert, C., Fournier, A., Barbet-Massin, M., Salles, J. M., Simard, F., & Courchamp, F. (2016). Massive yet grossly underestimated global costs of invasive insects. *Nature Communications*, 7, 12986. https://doi.org/10.1038/ncomms12986

Calfee, E., Agra, M. N., Palacio, M. A., Ramírez, S. R., & Coop, G. (2020). Selection and hybridization shaped the rapid spread of African honey bee ancestry in the americas. *PLoS Genetics*, 16(10), e1009038. https://doi.org/10.1371/journal.pgen.1009038

Cha, D. H., Adams, T., Werle, C. T., Sampson, B. J., Adamczyk, J. J., Rogg, H., & Landolt, P. J. (2014). A four-component synthetic attractant for Drosophila suzukii (Diptera: Drosophilidae) isolated from fermented bait headspace. *Pest Management Science*, 70(2), 324–331. https://doi.org/10.1002/ps.3568

Chown, S. L., Hodgins, K. A., Griffin, P. C., Oakeshott, J. G., Byrne, M., & Hoffmann, A. A. (2015). Biological invasions, climate change and genomics. *Evolutionary Applications*, 8(1), 23–46. https://doi.org/10.1111/eva.12234

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

Cooney, R. (2004). The Precautionary Principle in Biodiversity Conservation and Natural Resource Management: An issues paper for policy-makers, researchers and practitioners. In IUCN (Issue 2). http://data.iucn.org/dbtw-wpd/edocs/pgc-002.pdf

Cunningham, J. P., Carlsson, M. A., Villa, T. F., Dekker, T., & Clarke, A. R. (2016). Do Fruit Ripening Volatiles Enable Resource Specialism in Polyphagous Fruit Flies? *Journal of Chemical Ecology*, 42(9), 931–940. https://doi.org/10.1007/s10886-016-0752-5

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

Department of Primary Industries Victoria. (2010). *Invasive plants and animals policy framework.*

Dupuis, J. R., Bremer, F. T., Kauwe, A., San Jose, M., Leblanc, L., Rubinoff, D., & Geib, S. M. (2018). HiMAP: robust phylogenomics from highly multiplexed amplicon sequencing. *Molecular Ecology Resources*, 18(5), 1000–1019. https://doi.org/10.1111/1755-0998.12783

Dupuis, J. R., Sim, S. B., San Jose, M., Leblanc, L., Hoassain, M. A., Rubinoff, D., & Geib, S. M. (2018). Population genomics and comparisons of selective signatures in two invasions of melon fly, Bactrocera cucurbitae (Diptera: Tephritidae). *Biological Invasions*, 20(5), 1211–1228. https://doi.org/10.1007/s10530-017-1621-z

Ehrenfeld, J. G. (2010). Ecosystem consequences of biological invasions. *Annual Review of Ecology, Evolution, and Systematics*, 41, 59–80. https://doi.org/10.1146/annurev-ecolsys-102209-144650

Elton, C. S. (1958). *The Ecology of Invasions by Animals and Plants.* https://doi.org/10.1007/978-1-4899-7214-9

Epanchin-Niell, R. S., & Liebhold, A. M. (2015). Benefits of invasion prevention: Effect of time lags, spread rates, and damage persistence. *Ecological Economics*, 116, 146–153. https://doi.org/10.1016/j.ecolecon.2015.04.014

FAO. (2003). *Biosecurity in food and agriculture. Report on the 17th Session of the committee on agriculture, Rome 31 March–4 April 2003.* http://www.fao.org/3/Y8453E/Y8453E.htm

Finnoff, D., Shogren, J. F., Leung, B., & Lodge, D. (2007). Take a risk: Preferring prevention over control of biological invaders. *Ecological Economics*, 62(2), 216–222. https://doi.org/10.1016/j.ecolecon.2006.03.025

Gardy, J. L., & Loman, N. J. (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics*, 19(1), 9. https://doi.org/10.1038/nrg.2017.88

Goldstein, D. B., & Pollock, D. D. (1997). Launching Microsatellites: A Review of Mutation processes and methods of phylogenetic inference. *Journal of Heredity*, 88(5), 335–342.

Hamelin, R. C., & Roe, A. D. (2020). Genomic biosurveillance of forest invasive alien enemies: A story written in code. *Evolutionary Applications*, 13(1), 95–115. https://doi.org/10.1111/eva.12853

Hardulak, L. A., Morinière, J., Hausmann, A., Hendrich, L., Schmidt, S., Doczkal, D., Müller, J., Hebert, P. D. N., & Haszprunar, G. (2020). DNA metabarcoding for biodiversity monitoring in a national park: screening for invasive and pest species. *Molecular Ecology Resources*, 20(6), 1542–1557. https://doi.org/10.1111/1755-0998.13212

Herbert, K. S., Umina, P. A., Mitrovski, P. J., Powell, K. S., Viduka, K., & Hoffmann, A. A. (2010). Clone lineages of grape phylloxera differ in their performance on Vitis vinifera. *Bulletin of Entomological Research*, 100(6), 671–678. https://doi.org/10.1017/S0007485310000027

Hort Innovation. (2020). *Australian Horticulture Statistics Handbook*, 2019/20.

Hulme, P. E. (2011). Biosecurity: the changing face of invasion biology. In *Fifty years of invasion ecology: the legacy of Charles Elton*. Blackwell Publishing Oxford.

Jarrad, F. C., Barrett, S., Murray, J., Stoklosa, R., Whittle, P., & Mengersen, K. (2011). Ecological aspects of biosecurity surveillance design for the detection of multiple invasive animal species. *Biological Invasions*, *13*(4), 803–818. https://doi.org/10.1007/s10530-010-9870-0

Kalaris, T., Fieselmann, D., Magarey, R., Colunga-Garcia, M., Roda, A., Hardie, D., Cogger, N., Hammond, N., Martin, P. A. T., & Whittle, P. (2014). The Role of Surveillance Methods and Technologies in Plant Biosecurity. In G. Gordh & S. McKirdy (Eds.), *The Handbook of Plant Biosecurity: Principles and Practices for the Identification, Containment and Control of Organisms that Threaten Agriculture and the Environment Globally* (pp. 309–337). Springer Netherlands. https://doi.org/10.1007/978-94-007-7365-3_11

Kenis, M., Auger-Rozenberg, M. A., Roques, A., Timms, L., Péré, C., Cock, M. J. W., Settele, J., Augustin, S., & Lopez-Vaamonde, C. (2009). Ecological effects of invasive alien insects. *Biological Invasions*, *11*, 21–45. https://doi.org/10.1007/978-1-4020-9680-8_3

Kim, Y. H., Hur, J. H., Lee, G. S., Choi, M. Y., & Koh, Y. H. (2016). Rapid and highly accurate detection of Drosophila suzukii, spotted wing Drosophila (Diptera: Drosophilidae) by loop-mediated isothermal amplification assays. *Journal of Asia-Pacific Entomology*, 19(4), 1211–1216. https://doi.org/10.1016/j.aspen.2016.10.015

Kirk, H., Dorn, S., & Mazzi, D. (2013). Molecular genetics and genomics generate new insights into invertebrate pest invasions. *Evolutionary Applications*, 6(5), 842–856. https://doi.org/10.1111/eva.12071

Kogan, M. (1988). Integrated pest management theory and practice. *Entomologia Experimentalis et Applicata*, 49, 59–70. https://doi.org/10.1111/j.1570-7458.1988.tb02477.x

Lee, Y., Schmidt, H., Collier, T. C., Conner, W. R., Hanemaaijer, M. J., Slatkin, M., Marshall, J. M., Chiu, J. C., Smartt, C. T., Lanzaro, G. C., Mulligan, F. S., & Cornel, A. J. (2019). Genome-wide divergence among invasive populations of Aedes aegypti in California. *BMC Genomics*, 20, 204. https://doi.org/10.1186/s12864-019-5586-4

Leung, B., Lodge, D. M., Finnoff, D., Shogren, J. F., Lewis, M. A., & Lamberti, G. (2002). An ounce of prevention or a pound of cure: Bioeconomic risk analysis of invasive species. *Proceedings of the Royal Society B: Biological Sciences*, 269(1508), 2407–2413. https://doi.org/10.1098/rspb.2002.2179

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

Low-Choy, S. (2015). Getting the Story Straight: Laying the Foundations for Statistical Evaluation of the Performance of Surveillance. In F. Jarrad, S. Low-Choy, & K. Mengersen (Eds.), *Biosecurity Surveillance. Quantitative approaches* (6th ed., pp. 43–73). CABI.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., … Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, 395(10224), 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8

Mack, R. N., Simberloff, D., Lonsdale, W. M., Evans, H., Clout, M., & Bazzaz, F. A. (2000). Biotic invasions: causes, epidemiology, global consequences, and control. *Bulletin of the Ecological Society of America*, 10(3), 689–710. https://doi.org/10.1890/0012-9623(2008)89[341:iie]2.0.co;2

MacLeod, A. (2015). The relationship between biosecurity surveillance and risk analysis. In F. Jarrad, S. Low-Choy, & K. Mengersen (Eds.), *Biosecurity Surveillance. Quantitative approaches* (pp. 109–120). CABI.

Martin, R. R., Constable, F., & Tzanetakis, I. E. (2016). Quarantine Regulations and the Impact of Modern Detection Methods. *Annual Review of Phytopathology*, 54(1), 189–205. https://doi.org/10.1146/annurev-phyto-080615-100105

McCartney, M. A., Mallez, S., & Gohl, D. M. (2019). Genome projects in invasion biology. *Conservation Genetics*, 20(6), 1201–1222. https://doi.org/10.1007/s10592-019-01224-x

North, H., Mcgaughran, A., & Jiggins, C. (2021). The population genomics of invasive species. *Authorea Preprints*. https://doi.org/10.22541/au.160968166.65928724/v1

Pejchar, L., & Mooney, H. A. (2009). Invasive species, ecosystem services and human well-being. *Trends in Ecology and Evolution*, 24(9), 497–504. https://doi.org/10.1016/j.tree.2009.03.016

Pluess, T., Jarošík, V., Pyšek, P., Cannon, R., Pergl, J., Breukers, A., & Bacher, S. (2012). Which Factors Affect the Success or Failure of Eradication Campaigns against Alien Species? PLoS ONE, 7(10). https://doi.org/10.1371/journal.pone.0048157

Poland, T. M., & Rassati, D. (2019). Improved biosecurity surveillance of non-native forest insects: a review of current methods. *J. Pest Sci.*, 92(1), 37–49. https://doi.org/10.1007/s10340-018-1004-y

Quinlan, M., Stanaway, M., Mengersen, K., & others. (2015). Biosecurity surveillance in agriculture and environment: a Review. In F. Jarrad, S. Low-Choy, & K. Mengersen (Eds.), *Biosecurity Surveillance. Quantitative approaches* (pp. 9–42).

Reaser, J. K., Burgiel, S. W., Kirkey, J., Brantley, K. A., Veatch, S. D., & Burgos-Rodríguez, J. (2020). The early detection of and rapid response (EDRR) to invasive species: a conceptual framework and federal capacities assessment. *Biological Invasions*, 22, 1–19. https://doi.org/10.1007/s10530-019-02156-w

Roe, A. D., Torson, A. S., Bilodeau, G., Bilodeau, P., Blackburn, G. S., Cui, M., Cusson, M., Doucet, D., Griess, V. C., Lafond, V., Paradis, G., Porth, I., Prunier, J., Srivastava, V., Tremblay, E., Uzunovic, A., Yemshanov, D., & Hamelin, R. C. (2019). Biosurveillance of forest insects: part I—integration and application of genomic tools to the surveillance of non-native forest insects. *Journal of Pest Science*, 92(1), 51–70. https://doi.org/10.1007/s10340-018-1027-4

Rout, T. M., Moore, J. L., Possingham, H. P., & McCarthy, M. A. (2011). Allocating biosecurity resources between preventing, detecting, and eradicating island invasions. *Ecological Economics*, 71(1), 54–62. https://doi.org/10.1016/j.ecolecon.2011.09.009

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Schmidt, T. L., Swan, T., Chung, J., Karl, S., Demok, S., Yang, Q., Field, M. A., Odwell Muzari, M., Ehlers, G., Brugh, M., Bellwood, R., Horne, P., Burkot, T. R., Ritchie, S., & Hoffmann, A. A. (2021). Spatial population genomics of a recent mosquito invasion. *Molecular Ecology*, 30, 1174–1189. https://doi.org/10.1111/mec.15792

Seebens, H., Blackburn, T. M., Dyer, E. E., Genovesi, P., Hulme, P. E., Jeschke, J. M., Pagad, S., Pyšek, P., van Kleunen, M., Winter, M., Ansong, M., Arianoutsou, M., Bacher, S., Blasius, B., Brockerhoff, E. G., Brundu, G., Capinha, C., Causton, C. E., Celesti-Grapow, L., … Essl, F. (2018). Global rise in emerging alien species results from increased accessibility of new source pools. *Proceedings of the National Academy of Sciences*, 115(10), E2264–E2273. https://doi.org/10.1073/pnas.1719429115

Sharma, S., McKirdy, S., & Macbeth, F. (2014). The biosecurity continuum and trade: Tools for post-border biosecurity. In G. Gordh & S. McKirdy (Eds.), *The Handbook of Plant Biosecurity* (pp. 189–206). Springer.

Spears, L. R., & Ramirez, R. A. (2015). Learning to love Leftovers: Using By-Catch to Expand Our Knowledge in Entomology. *American Entomologist*, 61(3), 168–173. https://doi.org/https://doi.org/10.1093/ae/tmv046

Stenberg, J. A. (2017). A Conceptual Framework for Integrated Pest Management. *Trends in Plant Science*, 22(9), 759–769. https://doi.org/10.1016/j.tplants.2017.06.010

Tay, W. T., & Gordon, K. H. J. (2019). Going global – genomic insights into insect invasions. *Current Opinion in Insect Science*, 31, 123–130. https://doi.org/10.1016/j.cois.2018.12.002

Tay, W. T., Rane, R., Padovan, A., Walsh, T., Elfekih, S., Downes, S., Nam, K., D'Alençon, E., Zhang, J. P., Wu, Y., Nègre, N., Kunz, D., Kriticos, D. J., Czepak, C., Otim, M., & Gordon, K. H. J. (2020). Global FAW population genomic signature supports complex introduction events across the Old World. *BioRxiv*. https://doi.org/10.1101/2020.06.12.147660

Tedersoo, L., Drenkhan, R., Anslan, S., Morales-Rodriguez, C., & Cleary, M. (2019). High-throughput identification and diagnostics of pathogens and pests: overview and practical recommendations. *Molecular Ecology Resources*, 19(1), 47–76. https://doi.org/10.1111/1755-0998.12959

Thomas, M. L., Gunawardene, N., Horton, K., Williams, A., O'Connor, S., McKirdy, S., & van der Merwe, J. (2017). Many eyes on the ground: citizen science is an effective early detection tool for biosecurity. *Biological Invasions*, 19(9), 2751–2765. https://doi.org/10.1007/s10530-017-1481-6

Van Leeuwen, T., Dermauw, W., Mavridis, K., & Vontas, J. (2020). Significance and interpretation of molecular diagnostics for insecticide resistance management of agricultural pests. *Current Opinion in Insect Science*, 39, 69–76. https://doi.org/10.1016/j.cois.2020.03.006

Waage, J. K., & Mumford, J. D. (2008). Agricultural biosecurity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1492), 863–876. https://doi.org/10.1098/rstb.2007.2188

Whattam, M., Clover, G., Firko, M., & Kalaris, T. (2014). The biosecurity continuum and trade: border operations. In G. Gordh & S. McKirdy (Eds.), *The handbook of plant biosecurity* (pp. 149–188). Springer.

Witzgall, P., Kirsch, P., & Cork, A. (2010). Sex pheromones and their impact on pest management. *Journal of Chemical Ecology*, 36(1), 80–100. https://doi.org/10.1007/s10886-009-9737-y

Wu, N., Zhang, S., Li, X., Cao, Y., Liu, X., Wang, Q., Liu, Q., Liu, H., Hu, X., Zhou, X. J., James, A. A., Zhang, Z., Huang, Y., & Zhan, S. (2019). Fall webworm genomes yield insights into rapid adaptation of invasive species. *Nature Ecology and Evolution*, 3(1), 105–115. https://doi.org/10.1038/s41559-018-0746-5

You, M., Ke, F., You, S., Wu, Z., Liu, Q., He, W., Baxter, S. W., Yuchi, Z., Vasseur, L., Gurr, G. M., Ward, C. M., Cerda, H., Yang, G., Peng, L., Jin, Y., Xie, M., Cai, L., Douglas, C. J., Isman, M. B., … Zhuang, M. (2020). Variation among 532 genomes unveils the origin and evolutionary history of a global insect herbivore. *Nature Communications*, 11, 2321. https://doi.org/10.1038/s41467-020-16178-9

2

Prospects and Challenges of implementing DNA Metabarcoding for High-Throughput Insect Surveillance

2.1 Chapter preface:

This chapter combines an in-depth literature review with novel analyses of publicly available sequence data to evaluate the prospects for implementing universal metabarcoding assays within insect diagnostic laboratories. This chapter consolidates current best practices from the largely ecology focussed metabarcoding literature into a set of recommendations for laboratory processing, bioinformatic analysis, quality control, and data reporting. In the process, a series of technical challenges are identified that may prove barriers to adoption within the highly regulated field of invasive insect diagnostics. Several of these challenges are then addressed in later chapters, while many of the regulatory and policy considerations highlighted here are referred to throughout the thesis. This chapter is presented in published format.

2.2 Publication details:

Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance

Stage of publication: Published

Journal details: GigaScience, Volume 8, Issue 8, August 2019, giz092, https://doi.org/10.1093/gigascience/giz092

Authors: Alexander M. Piper, Jana Batovska, Noel O. I. Cogan, John Weiss, John Paul Cunningham, Brendan C. Rodoni, Mark J. Blacket

2.3 Statement of joint authorship:

A.M.P. and M.J.B. conceptualized the study. A.M.P. wrote the first draft of the manuscript with contributions from J.B., M.J.B, and J.P.C. J.W. contributed to the sections on detection probability and sampling considerations. B.C.R. contributed to the sections on reporting detections and regulatory considerations. N.O.I.C. contributed to the discussion of sequencing platforms and costs involved. J.P.C., M.J.B. and N.O.I.C. provided supervision. All authors contributed to the editing of the final manuscript and approved the version submitted for publication

Statement from co-author confirming the contribution of the PhD candidate:

"As co-author of the manuscript 'Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. GigaScience, 8(8), giz092.', I confirm that Alexander M. Piper has made the contributions listed above."

Associate Professor John Paul Cunningham

30/03/2021

GigaScience, 8, 2019, 1–22

doi: 10.1093/gigascience/giz092
Review

REVIEW

Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance

Alexander M. Piper[1,2,*], Jana Batovska[1,2], Noel O.I. Cogan[1,2], John Weiss[1], John Paul Cunningham[1], Brendan C. Rodoni[1,2] and Mark J. Blacket[1]

[1]Agriculture Victoria Research, AgriBio Centre, 5 Ring Road, Bundoora 3083, VIC, Australia; and [2]School of Applied Systems Biology, La Trobe University, Bundoora 3083, VIC, Australia

*Correspondence address. Alexander M. Piper. AgriBio Centre, 5 Ring Road, Bundoora 3083, VIC, Australia ; E-mail: alexander.piper@ecodev.vic.gov.au http://orcid.org/0000-0002-0664-7564

Abstract

Trap-based surveillance strategies are widely used for monitoring of invasive insect species, aiming to detect newly arrived exotic taxa as well as track the population levels of established or endemic pests. Where these surveillance traps have low specificity and capture non-target endemic species in excess of the target pests, the need for extensive specimen sorting and identification creates a major diagnostic bottleneck. While the recent development of standardized molecular diagnostics has partly alleviated this requirement, the single specimen per reaction nature of these methods does not readily scale to the sheer number of insects trapped in surveillance programmes. Consequently, target lists are often restricted to a few high-priority pests, allowing unanticipated species to avoid detection and potentially establish populations.

DNA metabarcoding has recently emerged as a method for conducting simultaneous, multi-species identification of complex mixed communities and may lend itself ideally to rapid diagnostics of bulk insect trap samples. Moreover, the high-throughput nature of recent sequencing platforms could enable the multiplexing of hundreds of diverse trap samples on a single flow cell, thereby providing the means to dramatically scale up insect surveillance in terms of both the quantity of traps that can be processed concurrently and number of pest species that can be targeted. In this review of the metabarcoding literature, we explore how DNA metabarcoding could be tailored to the detection of invasive insects in a surveillance context and highlight the unique technical and regulatory challenges that must be considered when implementing high-throughput sequencing technologies into sensitive diagnostic applications.

Keywords: biosecurity; alien species; biosurveillance; early detection; bioinformatics; reference database; quality assurance; controls; validation; non-destructive

Background

Increasing globalization of trade and tourism, along with changing climates, is expected to further increase the rate of biological invasions over coming decades [1–3]. Insects form a dominant component of this global spread of invasive species [4], posing a major threat to agroecosystems [5], the environment [6], and human health [7] through disruption of ecological networks, plant herbivory, and the transmission of pathogens and disease [8]. Once established in a new environment, ongoing containment and control of invasive insect pests imposes substantial costs to industry, government, and private landowners [8], and conse-

Received: 11 January 2019; **Revised:** 25 June 2019; **Accepted:** 9 July 2019

quently major efforts are made to forecast incursion risk [9–11] and implement quarantine of entry pathways [12–14]. Despite these measures, the exponential increase in global movement of food, commerce, and humans complicates traceability and makes quarantine inspection of more than a fraction of arriving cargo an impossible task [15, 16]. Therefore, proactive post-border surveillance within agricultural and natural landscapes is becoming an increasingly important component of effective biosecurity programmes, aiming to detect invasive species early before populations escalate or spread and eradication becomes unfeasible [17–19].

Insect invasions can initiate and disperse across vast and highly heterogenous landscapes [20], and therefore surveillance programmes often involve extensive trapping conducted across a range of spatial scales, from large geographic areas to precise crop-monitoring activities within agricultural properties [21]. Because it is generally unclear whether a new introduction has occurred or what species it may be, surveillance programmes can extend over many years and target diverse taxonomic groups [22, 23]. In many cases surveillance traps will capture non-target endemic species in vast excess of the target pests and the sheer number of specimens that need to be sorted through and identified by highly trained entomologists forms a major diagnostic bottleneck. While insect diagnostics still largely relies on traditional morphological examination [24], in recent years this has been supplemented by a range of molecular techniques that allow standardized identification of a wide range of taxa without specialist taxonomic expertise (Table 1). DNA barcoding in particular has become a central component of the modern diagnostic toolbox, owing to the ability to compare a single unknown specimen against many potential species in a single assay, and standardized protocols that allow transparent and objective comparison of specimen identifications between laboratories, regulatory agencies, and trading partners [24–26]. Despite these advantages, the time-consuming process of conducting single PCR and sequencing reactions on individual specimens has restricted the use of DNA barcoding to confirming the identity of specimens already deemed suspect by prior morphological sorting, or for identification of taxa or life stages where a taxonomic key may not be available or key diagnostic structures are degraded or missing [24, 27]. Without access to a scalable and cost-effective diagnostic method for large trap catches, current surveillance programmes generally do not identify all specimens to species level [23, 28]. Instead, target lists are confined to relatively few priority pest species identified by previous risk assessment [9] or statistical methods are used to select only a subset of specimens for species-level identification [29]. These restrictions can result in the non-detection of unanticipated or cryptic invasive species that are not being actively monitored for [30].

In order to overcome the limitations of current identification methods for processing large numbers of specimens, recent studies have looked to high-throughput sequencing (HTS) technologies to allow DNA barcode-based identification to be conducted in a massively parallel manner. This process, termed “metabarcoding” [31] or “marker gene sequencing” [32], generates a large number of individual barcode sequences in a single reaction, enabling the simultaneous identification of individuals in large mixed communities [33, 34], such as a trap sample containing many different insect species. The ability to rapidly and cost-effectively survey biodiversity has led to metabarcoding being taken up across numerous fields of applied ecology [34–37], including the identification of invasive species (Fig. 1A) [33, 38–40]. By identifying both endemic and potential exotic species in a bulk DNA analysis approach, metabarcoding can obviate the time-consuming specimen sorting required by previous molecular and morphological diagnostic methods, and allow detection of not just key pests but also other unanticipated species that are not being actively searched for [38, 41, 42]. This aspect is particularly advantageous for the detection of environmental threats because when one considers impacts beyond just agriculture and the time lag that can occur between introduction of a new species and perceptible damage to the environment [43], it becomes clear that there are far more invasive species of threat than can be identified by risk assessment and incorporated into target lists [23, 44]. A further advantage arises from the ability of HTS to count occurrences of specific sequences in a mixed sample [45], potentially allowing simultaneous pest identification and population size estimation. Finally, the rapidly increasing output of HTS technologies enables multiplexing of hundreds of trap samples in a single sequencing run, providing an avenue to dramatically scale up insect surveillance to the level required for effective, affordable, and proactive management response.

Despite the advantages that metabarcoding may offer to insect surveillance programs, uptake of new diagnostic tools into operational use depends on more than just the cost-effectiveness of the tool, but also on factors such as ease of use, accuracy, reproducibility, and perceived usefulness to the end users, as well as compatibility with existing policy frameworks [46, 47]. With the introduction of the World Trade Organisation Agreement on the Application of Sanitary and Phytosanitary measures (SPS) came new obligations for exporting nations to demonstrate freedom of a geographic area from particular pests using scientifically rigorous surveillance practices [48]. This agreement has in turn led to harmonization of routine diagnostic procedures into internationally standardized protocols to ensure that all end users are aware of the particulars involved and therefore committed to accepting any risk management actions that arise through its use [46, 49]. The SPS agreement recognizes the International Plant Protection Convention (IPPC) and the World Organisation of Animal Health (OIE) as the international standard-setting bodies for plant and animal health, respectively [48], and adoption of new standards stems from exhaustive workgroup efforts by these agencies [13, 50]. While the opportunities that HTS approaches could offer have been widely recognized by the diagnostics community [51, 52], because of the relative infancy of the technology, standards and guidelines around their use is a rapidly evolving space and validated protocols do not yet exist. Despite this, there is flexibility within the SPS framework for trading partners to introduce novel sanitary or surveillance procedures if it can be demonstrated that they are equivalent to or better than previous methods [49] and both the IPPC and OIE have now released guidelines for those laboratories preparing to implement HTS approaches in routine diagnostics applications. These guidelines highlight the need for robust experimental designs, assay validation, and quality assurance [51, 53, 54], reflecting recent discussions in the wider metabarcoding community [55]. In this review we explore the application of metabarcoding for high-throughput species-level identification of insects, providing an overview of common metabarcoding workflows (Fig. 2) and considerations required at each step to ensure reliable detection and quantification of taxa within complex mixed communities. We further discuss the unique technical and regulatory challenges of integrating broad-spectrum HTS assays into diagnostic laboratories and offer a perspective on the future adoption of high-throughput insect surveillance within international biosecurity frameworks.

Table 1: Methods used for insect identification, with suitability assessed according to accuracy, expertise, general applicability, time, and throughput criteria

Identification method	Taxonomic expertise	Identify specific taxa	Identify broad range of taxa	Throughput level	Time per identification
Morphological					
Microscopic examination	High	High*	High*	Low	Moderate
Molecular					
PCR–restriction fragment length polymorphism	Low	Moderate	Low	Moderate	Moderate
DNA barcoding	Low	High	High	Low	Moderate
Quantitative PCR/droplet digital PCR	Low	High	Low	High	Low
Loop-mediated isothermal amplification	Low	High	Low	Low	Low
Metabarcoding	Low	High	High	Very high	Low

*This morphological identification score assumes a high level of taxonomic knowledge and a low human error rate.

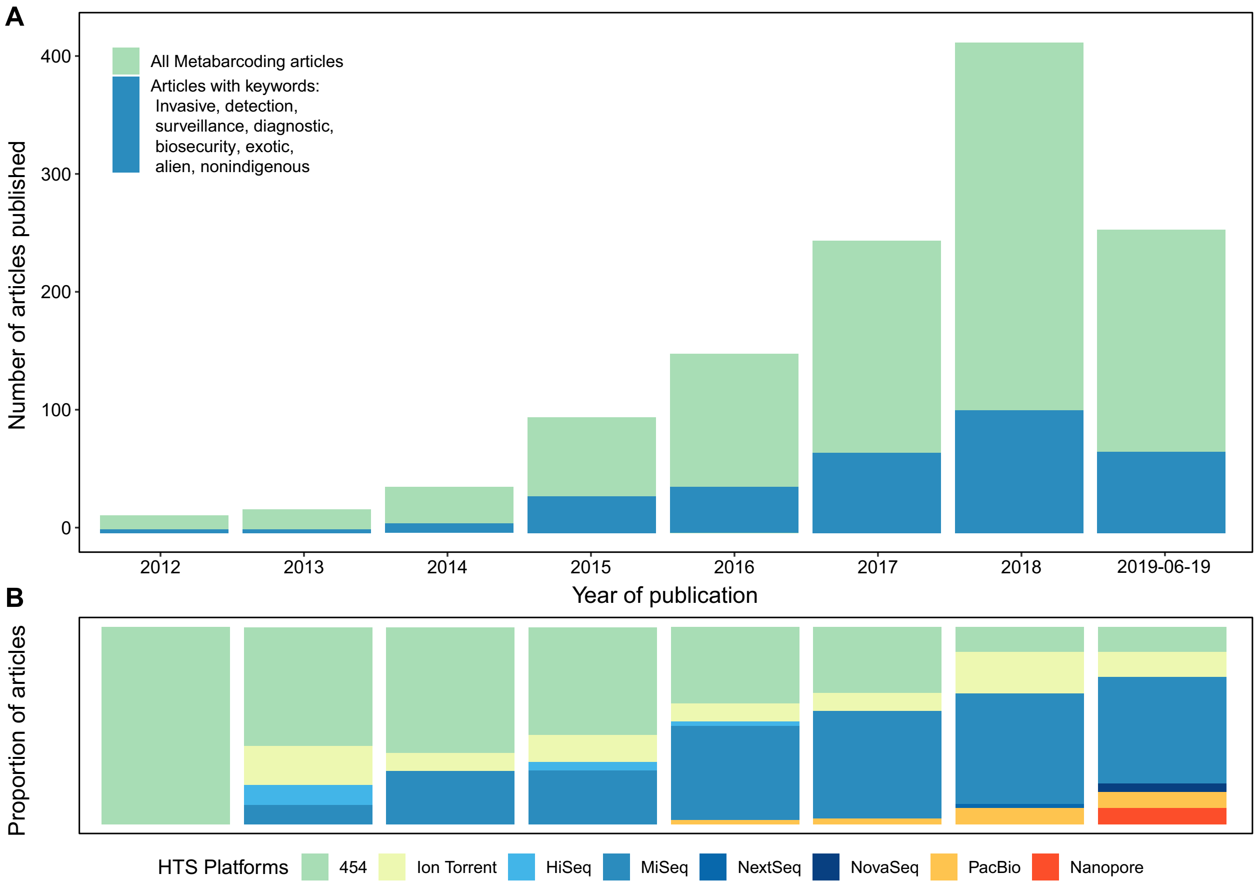


Figure 1: Metabarcoding in the literature. (A) Published articles obtained from Scopus, Crossref, and PubMed searches on 6 June 2019 for all metabarcoding studies, and those containing keywords in title or abstract relevant to invasive insect surveillance. (B) Sequencing platforms used in the above metabarcoding studies displayed as a proportion for each year.

Review

Selecting a taxonomic marker

Appropriate selection of a taxonomic marker or barcode locus is a critical first step in design of a metabarcoding assay because all downstream species detection and identification will rely on how conserved this marker is across taxa, and the discriminatory power of the nucleotide variation contained within it [56].

The markers most commonly used in metabarcoding studies are those already widely adopted for conventional DNA barcoding, and therefore the mitochondrial cytochrome oxidase I (COI) locus has been the most widely used marker for metabarcoding of insects to date. The 658-bp region of COI [57] used for conventional DNA barcoding has a strong track record of delivering species-level identification of insect pests [58]; however, many HTS platforms impose strict limitations in molecule length that can be sequenced (Table 2) and therefore smaller stretches of the

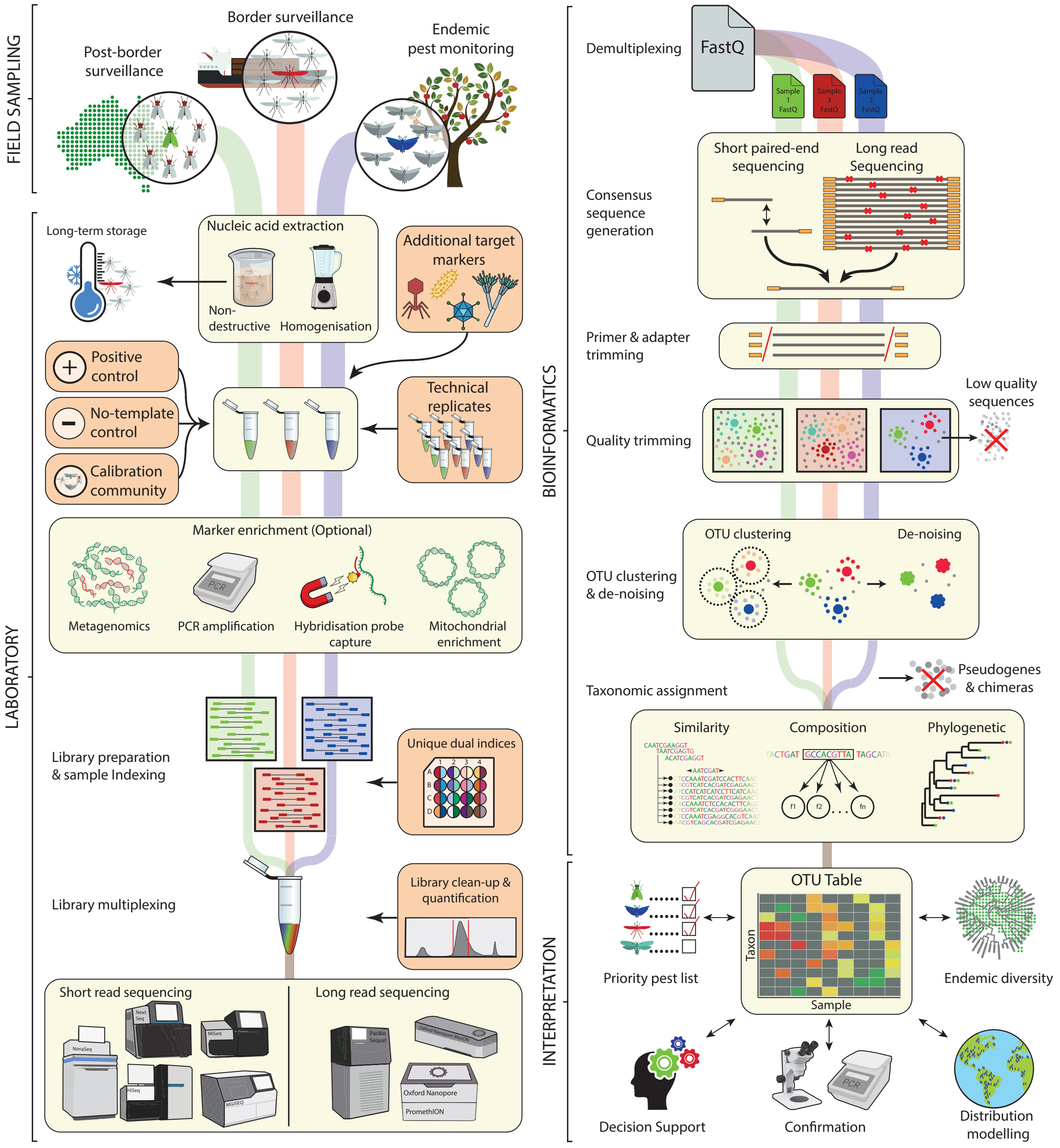


Figure 2: Overview of common metabarcoding workflows for identification of trapped insect species

conventional barcode loci or “mini-barcodes” must be used [59]. Nevertheless, research into degraded DNA samples has shown that singular COI barcode of sizes between 135 [60] and 250 bp [61] can reliably distinguish most animal species; however, appropriate placement within the larger barcode region is essential [62]. Despite the excellent taxonomic resolution provided by COI, since its application to metabarcoding a number of further limitations have become particularly apparent. Because 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 [63]. This mismatch inevitably leads to primers having variable affinity for different template molecules, biasing the amplification towards well-matched taxa and failing to amplify others [64]. Unlike conventional DNA barcoding where a failed amplification will result in a noticeably absent PCR product, in a bulk sample failed amplification of a particular taxon will be masked by the recovery of sequences from other taxa and therefore will go unnoticed [63]. A further issue inherent to mitochondrial loci such as COI is the proliferation of nuclear mitochondrial pseudogenes (numts) in many insect

Table 2: Comparison of sequence throughputs, error rate, and associated costs among high-throughput sequencing platforms

	Short-read platforms							Long-read platforms			
	Illumina MiSeq	Illumina NextSeq	Illumina HiSeq 3000/4000	Illumina NovaSeq	MGISeq-200	MGISeq-2000	MGISeq-T7	PacBio Sequel	PacBio Sequel II	ONT MinION	ONT PromethION
Maximum throughput (Gb)	15	120	750/1,500 (8/16 lanes)	6,000 (8 lanes)	60	1,080	6,000	20	160	20	150 per flow cell (up to 48)
Maximum read length	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 100 bp	2 × 150 bp	2 × 150 bp	~100 kb	~100 kb	~2 Mb	~2 Mb
Error rate	Low	Low	Low	Low	Low	Low	Low	Low (consensus error)	Low (consensus error)	High	High
Instrument cost	Low	Medium	High	High	Low	Medium	High	High	High	Extremely low	Low
Set-up time (labour)	Medium	Medium	Medium	Medium	Medium	Medium	Medium	High	High	Low	Low
Run time (hours)	56	30	84	40	<48	<48	24	15	15	1–72	1–72
Sequencing cost per sample*,†	<$50	<$15	<$10	<$5	<$50	<$10	<$5	<$25	<$15	<$25	<$5

*Costs are presented in Australian Dollars (AUD) and consider chemistry cost, depreciation, servicing, and computational cost over the lifespan of the instrument; however, total costs and read lengths will further depend on target enrichment and library preparation methods used.

†Assuming pooled sequencing of many traps with 250-Mb sequencing effort per sample.

orders [65–67], the result of historical recombination between the mitochondrial and nuclear genomes [68]. Co-amplification or preferential amplification of these pseudogenes instead of the true mitochondrial locus can complicate species identification [67] and result in overestimation of taxonomic diversity in the sample [69].

As a result of the aforementioned issues, as well as the inability for COI to differentiate certain pest groups [70], a range of alternative universal barcode markers have been proposed (reviewed by Freeland [56]). Ribosomal RNA (rRNA) genes are particularly appealing owing to their high copy number and stem-loop structure that consists of highly conserved core sequences for primer binding, interspaced with variable regions providing taxonomic resolution [71, 72]. Despite this, rRNA regions are on average more conserved than COI and therefore while appropriate for reconstructing higher level relationships they require longer spans of nucleotides to be informative at the species level. For single-specimen barcoding this can be overcome by concatenating several markers to increase phylogenetic resolution [73]; however, this presents a challenge for metabarcoding of mixed communities because there is no way of knowing whether 2 non-overlapping markers are from the same individual [74]. Therefore, while multi-locus approaches can be useful for expanding the taxonomic diversity an assay can recover [75–77], in particular cross-kingdom diversity (Box 2), they do not necessarily provide greater resolution [45]. Consequently, closely related and difficult-to-diagnose pest taxa may require further studies to identify appropriate diagnostic loci [78], or the development of novel analytical methods to integrate taxonomic assignments from multiple independent barcode loci. Finally, the application of alternative markers to insect diagnostics will suffer from a lack of reference sequence data because many taxa, including those of economic importance, currently only have COI sequence data publicly available (Fig. 3B, 3C). Therefore, because species-level resolution is a requirement of many diagnostic standards [24, 49, 79], for the taxa in which it has sufficient resolution, the high mutation rate and extensive reference information obtainable for COI will maximize the utility of metabarcoding within a broad-spectrum surveillance programme [80].

Box 1:

Reference sequence databases

As with conventional DNA barcoding, accurate taxonomic assignment in metabarcoding studies relies on a well-curated reference database of DNA marker sequences tied to vouchered morphological specimens to compare query sequences against [81]. The primary public nucleotide databases of relevance to insect metabarcoding are the Barcode of Life Data System (BOLD) [82] and the NCBI GenBank database [83]. While GenBank hosts greater overall sequence data, BOLD represents a curated DNA barcoding database that aims to maintain consistent links between sequences, validated morphological specimens, and associated specimen collection metadata [84]. Concerted efforts to generate mitochondrial COI barcodes for major insect orders have led to broad coverage of insects of biosecurity concern in both major public databases [58]; however, many geographic regions are still under-sampled (Fig. 3A) and reference sequences for alternative loci are mostly unavailable (Fig. 3B and C). While continued public submission and high-throughput reference sequence generation [85] will increase the representation of missing taxa and loci over time, ensuring the quality of submitted sequences from correctly identified specimens is crucial [24]. There are numerous examples of barcode sequences being either insufficiently annotated [34], annotated with the incorrect species in public databases [81, 86–89], or multiple morpho-species assigned to the same DNA barcode, which may reflect misidentifications or the existence of species complexes [58]. These issues highlight the importance of engaging taxonomic experts to ensure a priori identification of a specimen before submitting a reference barcode to a public database [90, 91]. Furthermore, the use of non-destructive DNA extraction methods when generating barcode sequences would allow the retention of voucher specimens to ensure traceability between the molecular and morphological features, especially in the case of taxonomic reassignments [92].

While some metabarcoding studies have responded to the aforementioned issues by exclusively using in-house reference databases for taxonomic assignment [90, 93–95], because many insect surveillance programmes aim to detect species that are not locally present, the reliance on public data to supplement in-house sequences may be unavoidable. Some taxonomic classifiers used in metabarcoding studies provide the option to weight classifications towards certain reference sequences [96, 97], which could be beneficial when combining high-confidence in-house sequences with public sequences of more variable quality, or when the endemic diversity for the target region is well characterized [74, 98]. Regardless of source, barcode sequences will be compiled together and formatted appropriately for use with automatic taxonomic classification software [99–101], and this presents an ideal point where automated or semi-automated curation methods can be used to identify and remove any taxonomically mislabelled sequences or non-homologous regions such as pseudogenes [74, 102]. Finally, curated databases used in an active surveillance program should only be updated after rigorous testing with standardized datasets to ensure that assay results remain accurate and reproducible following addition of new sequences [103].

Marker enrichment

Similar to conventional DNA barcoding, most metabarcoding studies use a set of universal oligonucleotide primers to exponentially amplify a target barcode marker until it reaches a concentration appropriate for sequencing. This “amplicon sequencing” methodology has proven reliable and sensitive for detection of low-abundance taxa in bulk samples [40]. However, differential PCR amplification efficiencies between taxa generally result in a biased depiction of relative abundances of community members [104]. This bias is thought to mainly arise from primer-template mismatches, particularly at the 3′ end of the primer where extension takes place [64, 105] and therefore comprehensive *in silico* evaluation should be conducted at the beginning of a project to ensure that primer sequences are appropriate for the underlying target community [106–108]. Where mismatches with certain taxa are predicted to occur, inclusion of degenerate bases can overcome taxonomic bias inherent to a specific primer sequence [109, 110]; however, high levels of degeneracy can also lead to undesirable off-target amplification or formation of dimers [87, 111], which will require further laboratory validation to detect [71, 109, 112]. In addition to the effects of PCR primers, a range of template-specific factors including copy number of the loci [113], nucleotide composition and secondary structure [114], variable amplicon lengths [115], specimen biomass [116], and complexity of the species mixture [105, 117] can further contribute bias. While the cumulative bias from all these factors may suggest that amplicon sequencing can only be used for presence-absence data, importantly, sequencing reads are still correlated with DNA input in a predictable way, and biases should only affect the slope of that correlation [113]. Therefore the calculation of taxon-specific correction factors shows great promise for improving abundance estimates from metabarcoding data [113, 118–120], particularly for simpler communities such as those trapped using targeted attractant lures [17]. Nevertheless, if accurate quantification is essential for the surveillance programme, removing the PCR amplification process altogether should also be considered for improving taxon abundance estimates from metabarcoding data.

PCR-free approaches

The major alternative to amplicon sequencing–based metabarcoding involves simply fragmenting the genomic DNA extract to lengths appropriate for the sequencing platform and directly sequencing it without any prior bias-inducing enrichment step. This methodology, termed “shotgun metagenomics,” generates sequence reads comprising a random subsample of the mixed community DNA and relies on the higher representation of taxonomically informative multi-copy mitochondria and nuclear rRNA in this subsample to identify community members [121–123]. In addition, these high-copy regions can be assembled into long contigs and even full-length mitochondrial genomes for further phylogenetic inference and systematics applications [124, 125]. Despite this, restricting taxonomic analysis to just mitochondrial and nuclear rRNA regions still leaves the vast majority of reads corresponding to DNA that is not taxonomically informative or easily assembled from a bulk sample to be discarded [121] and deep sequencing will be required to reliably detect rare specimens in the community [125, 126]. While the rapid growth in sequencing capabilities is making this brute force approach to community identification increasingly possible, for routine surveillance a cost-effective method for enriching taxonomically informative loci should be used prior to sequencing. A range of potential methods for PCR-free sequence enrichment have been reviewed elsewhere (see Mamanova et al. [127] and Jones and Good [128]), but some examples that have been successfully used for metabarcoding include differential centrifugation to enrich for mitochondria [129] or baiting target barcode markers and whole mitochondria using hybridization probe capture [130–133]. Hybridization capture relies on the use of thousands of synthetic oligonucleotide probes, each with strict complementarity to a target sequence, and therefore should ideally be designed with a priori knowledge of every target sequence [128]. Although this may be a limiting factor for recovery of previously unsequenced diversity, the flexibility to include essentially infinite numbers of probes provides further advantages for building bespoke metabarcoding assays that capture diverse loci for purposes beyond taxonomic inference (Box 2). Nevertheless, while PCR-free approaches have shown improved correlations between sequencing reads and input DNA [123, 134], it is important to remember that HTS counts molecules not individual specimens [45] and therefore biases are likely to still remain due to variation in biomass and copy number between organisms and tissues [131, 134]. Furthermore, the process of PCR amplification is already widely accepted within diagnostics protocols [49], and implementation of alternative PCR-free sequence enrichment methods may require overcoming additional regulatory hurdles.

Box 2:

Modular metabarcoding assays

Many of the insect pests actively monitored by surveillance programs are not targeted because of direct damage they do to animals, plants, or the environment but instead the associated fungi, bacteria, viruses, and viroids for which they can be vectors [52, 135, 136]. Similar to identification of insects, detection of host-associated pathogens has previously required screening of trapped samples on a specimen-by-specimen basis using target-specific assays or culturing and morphological analysis [33]; however, this is rapidly being augmented with metabarcoding and metagenomic approaches [33, 103, 137, 138]. The ability of HTS platforms to sequence a heterogenous mix of loci opens up the opportunity for combining both the identification of insects and the screening of a diverse range of host-associated microbiota within a single multiplexed metabarcoding assay [40, 139]. Nonetheless, developing an integrated assay that allows detection and identification of biologically diverse organisms in a diagnostics context presents a number of challenges. Extraction techniques will need to be optimized to account for the pathogen association with its insect host (i.e., intracellular [140], external [141], gut-borne [142]), and specific microbial life histories may make this incompatible with non-destructive DNA extraction. Furthermore, PCR protocols will need to be optimized to account for the large differences in template quantity between abundant host DNA and low-titre vectored organisms [143].

In contrast with the high resolution that COI provides for identification of insects, the commonly used universal markers for bacterial and fungal barcoding struggle to identify organisms to the species or strain level, which is necessary to separate pathovars from common innocuous environmental organisms [33, 136]. Therefore, diagnostic assays that aim to be universal for identification of both host and vectored organisms will require analysis of a range of group-specific markers in multiplex, or make use of long-read HTS platforms for increased taxonomic resolution [144, 145]. While multiplexing many loci together in single PCR reactions can greatly simplify laboratory protocols and therefore costs involved, for metabarcoding this can be complicated by cross-reactivity between primers and individual primer sensitivities changing depending on community composition [76, 105, 112]. As an alternative, various target loci could be enriched in parallel reactions and then pooled together by sample prior to library preparation in proportions relative to the number of reads desired for each marker [40, 146]. This highly flexible modular approach would then allow group-specific microbial primers or other markers of interest to be added or retracted from the assay depending on the target community and needs of the end user. For example, Swift et al. [147] have demonstrated the ability of modular metabarcoding assays not just to identify cross-kingdom species composition but also to genotype microsatellite loci and sex-specific markers relevant to the community under study. While the field of invasion biology has traditionally been concerned with the transport and movement of species, this doctrine overlooks the intraspecific movement of genetic material such as pesticide resistance alleles [148], transposable elements [149], and genetically modified organisms [150]. The ability to capture essentially any loci in a modular metabarcoding assay may allow integration with a more gene-focused model of biosecurity in the future.

Library preparation and multiplexing

Regardless of whether an enrichment or metagenomics approach was used, platform-specific sequencing adapters need to be attached to the molecules (via ligation [151] or 1-step [152] or 2-step PCR [40, 106]) to form “libraries” that can then bind to the flow cell for sequencing (Fig. 4A). Because current HTS platforms output sequences far in excess of what is required to identify the taxa in a single community, metabarcoding studies commonly multiplex many samples together on a single flow cell and use oligonucleotide index sequences incorporated into the sequencing adapters to link sequencing reads back to origin sample. While a range of indexing strategies exist for HTS [153], for sensitive diagnostics applications it is critical to choose an approach that can adequately cope with the occasional recombination of these indices between molecules. Index-switching has received particular recent attention due to reports of remarkably high levels on current Illumina platforms [154]; however, similar phenomena can affect multiplexed sequencing across all major platforms to various degrees [155–159] (with the possible exception of recent MGI platforms [160]). Suggested causes include contamination from residual adapter/primer oligonucleotides [161], chimera formation during adapter PCR [162], mixed clusters on the flow cell [157], or physical contamination during library preparation or oligo synthesis by the vendor [159, 163, 164]. Regardless of mechanism, when not properly controlled for, index-switching can cause taxa from one sample to “bleed” into others, and while this will only produce false-positive results for a taxon of concern when a true-positive result is present in ≥1 of the samples, the spreading of positive signal across samples can imply that the taxon of interest has a larger geographic distribution than exists in reality. Recent studies have demonstrated that the most effective method for controlling for index-switching is through the use of unique dual indices (Fig. 4C) rather than the commonly used combinatorial indexing (Fig. 4B). When unique dual indices are used, switching events at either end of the molecule will generate an index combination that was not originally applied and, during de-multiplexing, the reads with mismatched indices to the sample sheet will be filtered into an unassigned-reads file and excluded from analysis [159, 162, 165]. Furthermore, sets of indices should be alternated for each sequencing run [51] because carryover of molecules between runs on an HTS machine can be a further cause of false-positive results in high-sensitivity sequencing applications [166]. Finally, it is important that index sequences used are designed with sufficient edit distance between them so that substitution or insertion/deletion errors within the index do not cause further sequence misassignment [131, 167], particularly for higher error rate platforms such as nanopore [115].

High-throughput sequencing platforms

While the rapid growth of HTS over the past decade has produced a variety of techniques and chemistries for discerning the nucleotide sequence of a DNA molecule [168], modern platforms can largely be divided into those producing short-but-accurate sequences or long-but-error-prone sequences (Table 2). To date, the majority of metabarcoding studies have been conducted us-

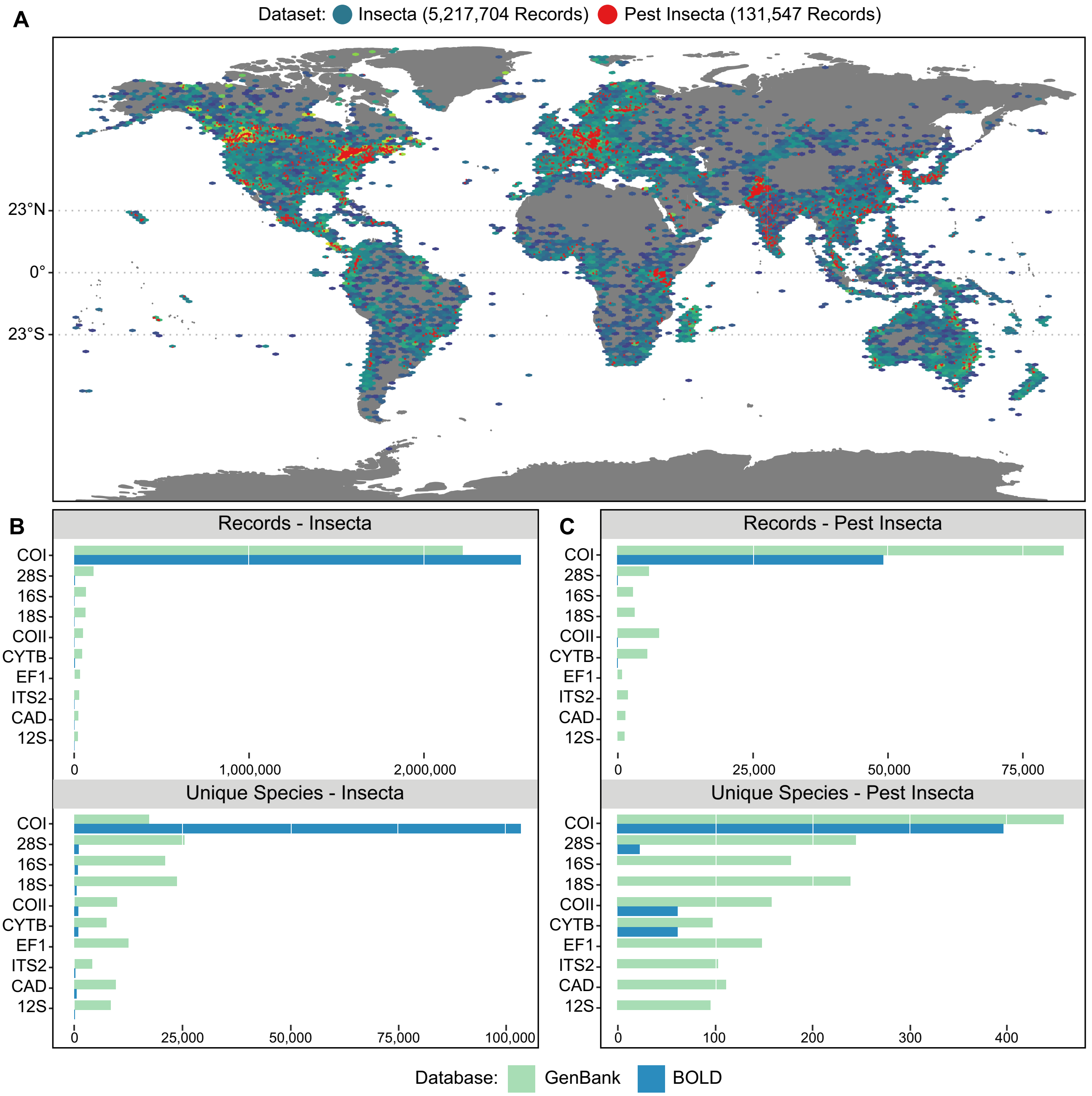


Figure 3: DNA barcodes in public reference databases. (A) Global distribution of all sufficiently annotated DNA barcode records from BOLD and GenBank for all barcode loci; records for all Insecta are displayed as a density map, while those species present on international pest lists are overlaid in red. (B) Distribution of records and unique species within major public databases for the 10 barcode markers with the most reference information for entire Insecta and for (C) Insecta species present on international pest lists.

ing the former, with the Illumina "MiSeq" dominating the recent metabarcoding literature due to its high-quality reads and relatively inexpensive purchase cost (Fig. 1B). Despite the current popularity of the MiSeq for research studies, the cost per sample may be impractical for the number of specimens produced by large-scale surveillance programmes, and instead the production-scale Illumina "NextSeq," "HiSeq," and "NovaSeq" provide progressive increases in throughput and therefore cost reductions (Table 2). Nevertheless this increased sequencing throughput of these platforms must be balanced with diagnostic turnaround times, and effective use of the ultrahigh-capacity HiSeq and NovaSeq flow cells will involve multiplexing of thousands of samples, necessitating substantial logistical efforts in sample collection and processing [103].

Despite the cost-effectiveness of the aforementioned platforms, their restricted read lengths (Table 2) limit the taxonomic resolution achievable with a metabarcoding assay and therefore long-read sequencing platforms such as the Pacific Biosciences (PacBio) "Sequel" and Oxford Nanopore Technologies (ONT) "MinION" and "PromethION" are becoming increasingly attractive alternatives. The ability to sequence barcode regions thousands of bases in length has potential to enable greater

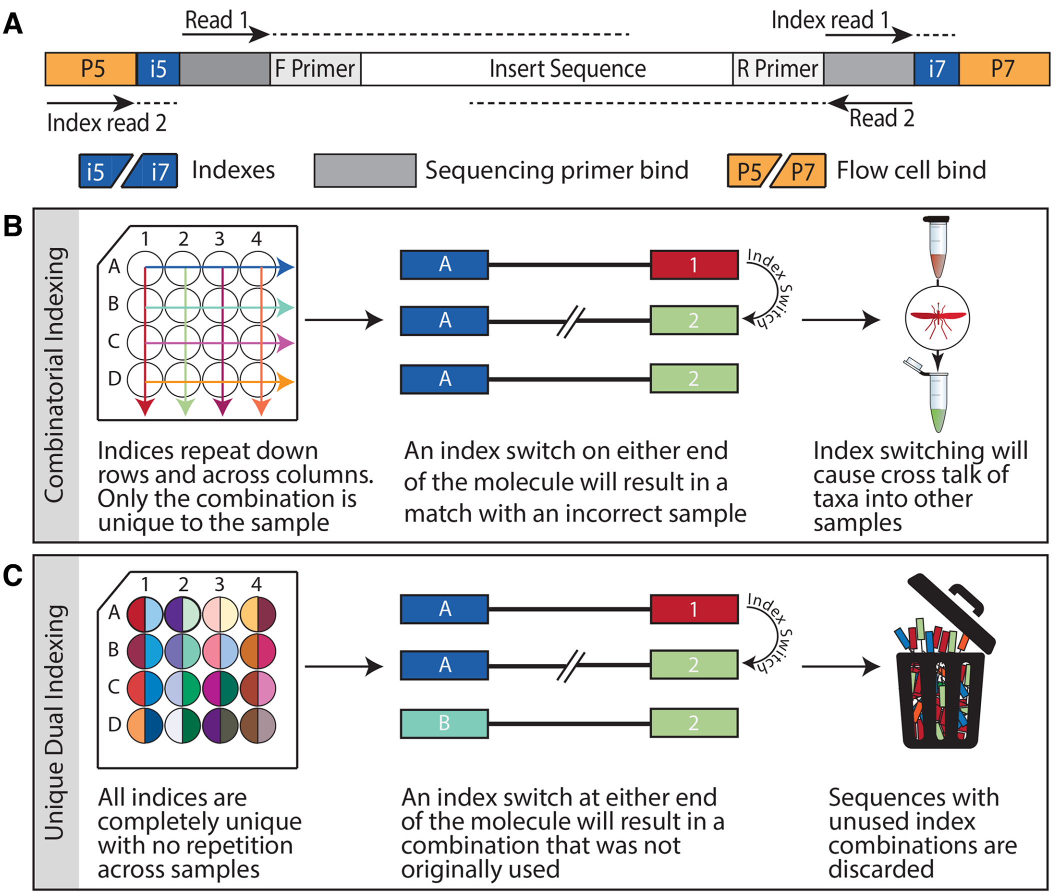

A
Read 1
Index read 1
P5
i5
F Primer
Insert Sequence
R Primer
i7
P7
Index read 2
Read 2
i5 / i7 Indexes
Sequencing primer bind
P5 / P7 Flow cell bind
B
Combinatorial Indexing
1 2 3 4
A
B
C
D
A
1
A
2
A
2
Index Switch
Indices repeat down rows and across columns. Only the combination is unique to the sample
An index switch on either end of the molecule will result in a match with an incorrect sample
Index switching will cause cross talk of taxa into other samples
C
Unique Dual Indexing
1 2 3 4
A
B
C
D
A
1
A
2
B
2
Index Switch
All indices are completely unique with no repetition across samples
An index switch at either end of the molecule will result in a combination that was not originally used
Sequences with unused index combinations are discarded

Figure 4: Unique dual indexing overcomes issues of cross-contamination due to index-switching. (A) An amplified barcode locus with sequencing adapters attached; read locations and orientations are indicated for commonly used Illumina MiSeq platform. Reads 1 and 2 are designed to overlap to facilitate assembly into a consensus sequence. Both sequencing adapters incorporate a unique oligonucleotide index sequence to allow differentiation of multiplexed samples. Strategies for indexing include (B) combinatorial indexing, where indices on either end of the molecule are shared with other samples but the combination of the two is unique, and (C) unique dual indexing, where adapter indices at both ends of the molecule are completely unique to the sample.

recovery of taxonomic diversity with intraspecific resolution [169]; however, in practice the utility of long reads for species identification has been limited by considerably higher per-base error rates that commonly exceed intraspecific distance [115, 170]. Nevertheless, methods for repeatedly sequencing a single molecule to create higher quality consensus sequences [171] are now opening up applications in metabarcoding [144, 158], with natively implemented circular consensus sequencing on the PacBio Sequel producing consensus reads with similar accuracy to traditional Sanger sequencing [172], and third-party protocols mimicking this approach have now been published for the ONT platforms [173, 174]. If similarly robust consensus sequencing can be achieved with nanopore technology, the significantly smaller start-up cost and portability of the handheld MinION platform may in future permit metabarcoding-based diagnostics to be conducted in remote field sites [115], as well as enable lesser resourced laboratories to access these technologies [14].

Bioinformatics

Computational processing of sequence reads represents a series of steps of equal importance to laboratory protocols for ensuring accurate and sensitive detection of invasive species [175, 176]; however, many of the skills and techniques involved in this process have not historically been required within diagnostic laboratories. While there exist a number of popular end-to-end computational pipelines for analysing marker gene data [177–181], many of these have been designed for measuring diversity rather than detection of low-abundance taxa. Each step in the bioinformatic analysis can present trade-offs between sensitivity to rare taxa, amount of erroneous sequences retained, and overall computing time [77, 175, 182–184], and use of metabarcoding in an invasive species surveillance or other sensitive context presents some unique challenges and regulatory requirements that may be best addressed through the creation of a custom analysis pipeline [146, 176].

De-multiplexing and sequence quality trimming

A metabarcoding assay typically involves multiplexing many samples into a single pooled sequencing library in order to make optimal use of the high-capacity flow cells of current sequencing platforms. Therefore, the first step following sequencing (typically automated by the HTS platform's software) is to assign sequences back to their origin sample using unique oligonucleotide sample indices incorporated into the sequencing adapters (Fig. 4). Following de-multiplexing, sequencing adapters and any other non-biological information such as PCR primer sequences are removed, and reads are assembled into consensus sequences using their overlapping bases. While improvements in underlying sequencing chemistries and aforementioned consensus approaches means that the majority of platforms now provide per base accuracies >99.99% (with the notable exception of nanopore platforms) [168, 173, 185], when put in context of the billions of bases sequenced on modern flow cells, tens of thousands of sequences will still contain errors [186]. Raw sequence reads are generated in conjunction with a predicted error profile based on signal intensity and background noise, and these data are generally presented to the user in the form of a FASTQ file [187]. An initial quality-trimming stage uses this profile to truncate or remove sequences that contain excessive ambiguous or low-confidence base calls [186, 188]; this is, however, a coarse filtering process where parameters should be carefully considered, particularly for higher error platforms such as nanopore. While strict quality trimming will more effectively remove sequencing artefacts and erroneous reads that can affect downstream diversity and abundance estimates, overly conservative parameters can result in removal of too many reads and therefore loss of sensitivity to low-abundance taxa [146, 176].

OTU clustering and denoising

While quality trimming can improve accuracy by removing sequencing errors, the PCR amplification process used in the majority of metabarcoding studies can further introduce single-base substitutions [158, 189] and length variation [190] that will not necessarily be associated with low quality scores [191]. Because these noisy sequences can cause spurious results and substantially increase downstream computation, many studies cluster together all sequences within an arbitrary similarity threshold (commonly 97%) into representative bins called "operational taxonomic units" (OTUs). While the 97% similarity threshold is thought to represent a broadly generalizable compromise between interspecific and intraspecific variation and is commonly used to indicate distinct taxa [192, 193], actual coalescent depths between species can differ greatly across taxonomic groups [91]. Therefore when a single global threshold is applied to diverse communities it can result in both the splitting of a single species across multiple OTUs, as well as the lumping of multiple species into the same OTU, resulting in false-negative results [176, 194]. Furthermore, aggregating all similar sequences into a single OTU loses all information on intraspecific diversity, restricting the ability to trace the geographic origin of invasive populations [39, 79]. In addition, the OTUs generated by clustering are dependent on the particular dataset, reference database, and parameters selected [194, 195], and as such they do not lend themselves to ongoing comparison with the constantly evolving data produced by a longitudinal surveillance programme. To overcome the aforementioned limitations, newly developed "denoising" algorithms instead use statistical models to infer true biological sequences from sequencing noise and correct for single-nucleotide differences, without imposing the arbitrary similarity threshold that defines OTUs [196–198]. This single-nucleotide resolution enables binning sequences into "amplicon sequence variants" (ASVs) [196] (also termed "exact sequence variants" [194], sub-OTUs [197], or zero-radius OTUs [zOTUs] [198]) that retain precise haplotype information that can be necessary for diagnostics of closely related taxa or tracking an invasion [199], and act as a consistent label between analyses [194].

OTU quality control

While the above measures account for the majority of low-abundance errors, they are not designed to deal with high-abundance artefacts such as PCR-generated chimeras and non-specific amplification products. Chimeric sequences are the result of incompletely extended PCR products acting as primers for a different closely related sequence [189], and therefore appear as concatenated products of 2 parent sequences. Assuming that parent sequences will be more abundant having undergone more rounds of amplification, chimeras can be algorithmically removed through comparison with other sequences in the sample [196, 200] or with a chimera-free reference database [201]. On the other hand, removing products of non-specific amplification such as intragenomic variants and pseudogenes presents more of a challenge and will generally require manual curation [151, 202]. When targeting protein-coding mitochondrial genes such as COI, the presence of stop codons and frameshifts that disrupt the open reading frame are common indicators of pseudogenes [80], and for rRNA markers secondary structure prediction could be used to ensure that sequences do not contain substantial variation in highly conserved regions [203]. Because it is inefficient to include a manual curation process as part of a high-throughput bioinformatics pipeline, it would be beneficial for future denoising algorithms to incorporate patterns of sequence evolution to allow more precise and automated filtering of barcode loci from erroneous and pseudogenic sequences [80, 204, 205].

Taxonomic assignment

In order to process the large diversity of sequences that a metabarcoding assay typically produces, the assignment of Linnaean taxonomy (e.g., species, genus) is typically conducted in an automated manner. While a large range of software tools exist for this purpose [206], the approaches used can generally be delineated into either sequence similarity searches (i.e., BLAST alignment), sequence composition methods (i.e., hidden Markov models and *k*-mer counts), phylogenetic methods, or a hybrid of the above (see Bazinet and Cummings [207] for an in-depth comparison). To date, the most widely used approach for taxonomic classification in metabarcoding studies has been best-hit classification using alignment based tools such as BLAST [208], which assume that the taxonomy of the query sequence will be identical to the taxonomy of the most similar sequence in a reference database. While this approach is simple to implement and can perform effectively when the reference database contains sequence information from conspecifics, when reference data are absent or when the particular loci cannot distinguish between multiple organisms, best-hit classification is prone to over-classifying the sequence to incorrect species-level taxonomy [209]. In the worst case, this over-classification error could lead to false-positive results by classifying a previously unsequenced but probably innocuous organism as a known pest, owing to the pest being the closest taxon with an existing reference sequence [210].

As the above situation demonstrates, for applications where management decisions are to be based on the results of a taxonomic classification, a central question is the reliability of that classification. A number of taxonomic assignment algorithms aim to address this issue by returning a measure of confidence of inclusion in each taxonomic rank, e.g., by using repeated random sampling [97, 211], lowest common ancestor methods [212], or probabilistic models [96, 213]. In an ideal case, only a single possible taxonomic outcome will obtain a high level of confidence, whereas alternate outcomes will obtain probabilities close to zero. In cases where there may be uncertainty at the species or genus level due to imperfect reference data and multiple taxonomic outcomes obtaining similar probabilities, the sequence may still be robustly assigned to a higher taxonomic rank (e.g., family) [101], providing important information about sample composition and possible presence of novel taxa without producing false-positive results [214]. While using measures of confidence can reduce the incidence of over-classification, many of these approaches are impaired by an inherent bias in that they infer the entire scope of possible taxonomic outcomes exclusively from the reference sequences used for training [215, 216], which in reality only represents taxonomic units that have been previously sequenced. In contrast, the Bayesian framework of PROTAX [96] accepts a reference taxonomy tree alongside the reference sequence database in order to account for taxa that are present in Linnaean taxonomy but not represented by reference sequences. Furthermore, PROTAX explicitly models the probability that a sequence belongs to a taxon that is novel to both the reference sequence database and reference taxonomy, which could be particularly important when conducting surveillance in regions with substantial uncharacterized biodiversity [216, 217]. Nevertheless, even the most complex taxonomic assignment algorithms do not model important aspects of species biology that may limit the possible geographical distribution or habitat in which they could reasonably exist, and therefore the results of taxonomic assignment should be vetted with ecological knowledge of the detected species where possible [35].

Quality assurance

The ability to simultaneously identify many loci from thousands of specimens in a single diagnostic assay underlies the power of the metabarcoding approach to surveillance; however, the resulting increase in sequence diversity and analytical complexity introduces further risk of cross-contamination and technical error [55]. An important challenge for the use of metabarcoding in a diagnostic context is the rate of false-positive errors (incorrect identification of an insect as the pest of concern) and false-negative errors (not identifying a pest of concern). While many ecological studies prioritize minimizing false-positive errors over false-negative errors [37], generally the precautionary principle applies in biosecurity; i.e., it is better to have a false-positive result that can be followed up with an orthologous confirmation method than to miss a serious pest. This is particularly important if the assay is to provide "evidence of absence" to support pest-free status [218], which can be required to access certain international markets [28]. Therefore, a quality assurance system for metabarcoding diagnostics should aim to reduce the frequency of false-positive results as much as possible through the appropriate use of controls, replication, and validation, without in turn increasing the incidence of false-negative results.

Controls and replication

The majority of contamination in next-generation sequencing assays is expected to arise from other samples processed in the same laboratory environment, particularly when PCR is involved [164, 219], and therefore workspaces should be physically or temporally separated for different assay steps, with all surfaces, equipment, and reagents regularly decontaminated [33, 219–221]. Periodic swipe tests of laboratory surfaces can help identify common laboratory contaminants and confirm the absence of environmental DNA from target pests [220, 222]. Despite these precautions, even the cleanest laboratory environment will not account for all possible contaminant sequences and therefore no-template controls should be included throughout the entire laboratory workflow and sequenced alongside the sample libraries to provide a cumulative measure of contamination [162, 223, 224]. When these controls are incorporated sequentially at each step of the laboratory protocol they can further enable partitioning of contamination to the stage in the workflow where it occurred, which can highlight processes that can be improved during assay development [35, 37]. Index-switching is perhaps the most worrisome cause of contaminating sequences in HTS, and while use of unique dual indices (Fig. 4C) can reduce this phenomenon to a level acceptable for most studies, trace levels of index-switching can still persist and cause issues for sensitive diagnostic applications [159]. While index-switching artefacts will be detectable in no-template controls, it can be difficult to discern this phenomenon from sequences arising through physical contamination. Instead, including a positive control library made up of synthetic standard DNA [177, 225, 226] or an "alien" taxon guaranteed to be absent from the sample [88, 227] allows empirical measurement of the index-switch rate. Alternatively, the rate of index-switching can be measured post hoc by comparing read counts between valid and invalid combinations of unique dual indices [131, 228]. Once contaminant sequences have been identified, their presence can be controlled through the application of a minimum abundance filter based on the read counts within negative and/or positive control libraries [35, 229], although choice of an appropriate threshold can be complicated by read depth differences between samples and preferential amplification of contaminants in low-biomass no-template control samples [175, 230]. As an alternative, new statistical methods allow systematic removal of contaminant sequences based on co-occurrence patterns and library quantification data [231–233]; however, if particularly high levels of contamination or abnormally high rates of index-switching are detected in a specific batch of samples, it may be more appropriate to repeat the assay. Finally, including an additional positive control in the form of a well-characterized mock "calibration community" in every sequencing run could further highlight any additional run-specific aberrations or batch effects that may have been introduced during the metabarcoding workflow when taxonomic composition or error rates deviate strongly from expected [205, 234, 235].

In addition to being prone to contamination, library preparation protocols involve a series of molecular bottlenecks where during each subsequent stage of DNA extraction, target enrichment, and binding of molecules onto the flow cell, only a random subsample of molecules are taken forward [37]. Stochasticity in this sampling process is likely to bias the resulting sequences towards more abundant taxa and increase the false-negative rate for rare taxa [236], and this can be further exacerbated by negative primer bias [77]. Potential loss of rare taxa during sample processing can be offset through the use of technical replicates, and these provide a further avenue to identify laboratory cross-contamination in the case that replicates show significant dissimilarities in taxonomic composition [77, 229, 237]. While using higher numbers of replicates can increase the probability of detecting rare taxa [237], this must be weighed against the increased costs of sequencing and library replication as well as the strategy for processing the replicates [37]. Additive processing (i.e., pooling the detections of all replicates) can be most useful for overcoming sampling stochasticity and controlling for false-negative results, while restrictive processing (i.e., only retaining sequences present in several replicates) more effectively controls for cross-contamination. To balance the positives of both approaches, it may be best to include a minimum number of technical replicates to allow a majority-rules approach (e.g., 2/3 replicates count as a detection) [77, 88, 112]. A further aspect to consider is the importance of biological replicates at the sample collection stage [238] because regardless of the effectiveness of the metabarcoding diagnostic assay, if an insect is not caught in a trap, it does not necessarily mean absence in the area. The use of site occupancy models that account for the false-positive– and false-negative–prone nature of metabarcoding surveys could be used to determine the optimal number of both technical and biological replicates to reach the desired statistical power for the survey [239, 240]. Finally, while outside the scope of this review, appropriate trap design [241] and surveillance grid planning [242] must also be adhered to for effective metabarcoding-based surveillance.

Validating metabarcoding assays

Because of the relevance of many invasive insects to international trade and human health, laboratories conducting insect diagnostics generally exist within strict regulatory environments. As part of laboratory accreditations, newly developed assays are required to undergo a validation process in order to provide objective evidence to all end users that an assay is fit for purpose [53, 54, 243, 244]. Traditionally, validation first involves defining the scope of the assay and then establishing performance parameters such as analytical sensitivity, analytical specificity, reproducibility and repeatability for every individual target designated in this scope [26, 244, 245]. However, the universal nature of metabarcoding assays and the taxonomic diversity of potential surveillance catch make this impractical [246]. To overcome this inevitable variation between reference samples and reality, a flexible scope validation process should be used to establish performance parameters on representative samples and identify critical steps in the workflow where variation can be introduced [146, 247]. These critical steps can then be monitored run to run using control samples and appropriate quality control checkpoints (Table 3) to ensure that no sample or sequence data continue without meeting minimum quality requirements [51, 221, 247, 248]. In the case of insect metabarcoding, mock communities made up of the taxonomic groups of interest are generally used for validation, which are then spiked with decreasing concentrations of target species in order to establish assay sensitivity and limits of detection [40, 249]. Because DNA extraction efficiency and primer bias can be affected by overall community complexity [105, 250], mock communities should as closely as possible represent the diversity expected to be recovered in different trapping scenarios. Furthermore, the amount of sequencing effort assigned to an individual sample during multiplexed sequencing can vary across runs [224, 251], and the effect of sequencing depth on detection should also be established using rarefaction curves [107, 117]. On the other hand, analytical specificity will generally depend on choices made during assay design, such as the choice of target marker, availability of appropriately annotated reference sequences for the chosen marker, and taxonomic assignment criteria used [220, 246]. Parameters such as precision and reproducibility of a metabarcoding assay can be established similar to other molecular diagnostics, through replication of samples and controls within and across sequencing runs and inter-laboratory comparisons [146]. Finally, stability of specimens and DNA to environmental factors such as temperature, UV radiation, pH of commonly used drowning or attractant solutions (e.g., vinegar traps [252]), and exposure to environmental microorganisms in the field and during storage [253] should be evaluated and may prompt a need for redesign of insect traps to collect and preserve samples in a manner more suited to DNA-based identification.

Reporting and confirming detections

Even when primers are designed around a specific taxonomic group, metabarcoding can amplify and detect many more taxa outside the scope of the original validated target list [254]. How these incidental detections are reported and eventually acted upon will present a major challenge to diagnostic laboratories and end users, due to the increased number of previously undocumented taxa being discovered for which knowledge of distribution or ecological significance may be missing [51, 53]. Many of these incidental detections will be taxa that simply have not previously been searched for, and when an appropriate management response is considered, it will be important not to conflate "first detection" in an invasion biology sense, where there was prior evidence of absence, with merely the first time a species has been formally identified in a region [255]. Hence a greater emphasis needs to be placed on conducting baseline surveys to establish comprehensive species checklists of endemic diversity and resolve synonymous taxa at the beginning of a surveillance programme to avoid creating sudden market access and trade issues [256]. Furthermore, a decision framework should be developed for evaluating incidental detections that sets out steps for further characterization and risk assessment for the detected organisms in order to establish whether eradication or other management actions are appropriate or achievable [257]. Where necessary, putative detections can be further confirmed using an orthogonal diagnostic method such as quantitative PCR/droplet digital PCR on the original DNA extract [146]; however, these assays require prior development and will therefore not be available for all incidental taxa detected in a metabarcoding assay. Instead, the use of non-destructive DNA extraction methods that use a combination of enzymes, buffers, and heat without mechanical homogenization [227, 258–260], or even amplification of insect DNA from the ethanol used to preserve specimens [261–264], would enable diagnosticians to revisit original samples following metabarcoding to confirm species detections. Development of a non-destructive metabarcoding assay has great potential for bridging the gap between new HTS methods and traditional entomological techniques and may bootstrap the acceptance of metabarcoding into international regulatory frameworks.

Perspectives and conclusions

The ability to accurately, rapidly, and cost-effectively determine the species composition of bulk insect trap contents using metabarcoding has the potential to revolutionize broad-spectrum surveillance for invasive insect pests. Similar to any novel technology, as metabarcoding transitions from purely research to management applications it faces the growing pains that come with integration into established regulatory structures. While rigorous standardization of both laboratory techniques and data analysis has proven essential for the acceptance of conventional DNA barcoding as a validated diagnostic for insects of regulatory concern [26, 79], the sheer pace of development of HTS technologies and platforms may complicate similar standardization of metabarcoding protocols. Historically, the effective lifespan of many HTS platforms has only amounted to a few years before obsolescence [168], and laboratory protocols and bioinformatic methods are therefore constantly evolving to chase this moving target. In response to this constantly shifting state of the art, harmonization efforts by regulatory bodies should avoid the over-prescription of restrictive standards into law because these will quickly become outdated and risk further widening the gap between research and diagnostics capabilities [46]. Instead, development and distribution of certified reference materials in the form of standard and diverse mock communities or DNA standards (similar to the ZymoBIOMICS microbial mock community standards [265]) as well as computational

Table 3: Recommended quality control checkpoints for metabarcoding-based diagnostics

Category	Quality control checkpoint	Consequences
Laboratory preparedness	Are all reagents within expiry date and stored properly?	Poor reagent storage can lead to reduced efficiency and false-negative results
	Is equipment appropriately maintained and calibrated?	Poorly calibrated equipment will generate inconstancies and inaccurate data
	Have laboratory surfaces been decontaminated and swipe testing of laboratory surfaces been conducted?	Dirty laboratories can be a source of DNA contamination, leading to lowered sensitivity or false-positive results
Sample acceptance	Have specimens arrived in a condition appropriate for extracting DNA?	Inappropriately stored specimens can lead to false-negative results and a reduction in sensitivity
	Are specimens traceable to origin location?	Misidentification of sample origin can complicate detection response
Nucleic acid extraction	Is DNA of sufficient quantity and quality?	Insufficient DNA quantity or presence of contaminants can inhibit reactions and result in false-negative results
Marker enrichment	Are the correct fragment sizes present for the target barcode marker?	Incorrect fragment sizes could indicate off-target amplification
	Have the positive control samples successfully amplified?	Absence of product in positive controls indicates amplification failure
	Are negative control samples free of DNA fragments?	Visible DNA fragments in negative controls indicates contamination
Library preparation and multiplexing	Are libraries of the appropriate size and concentration?	Libraries of significantly different sizes or concentrations will complicate multiplexing
	Have sets of unique dual indices been used?	Unique dual indexing is necessary to control for index-switching
	Have index sets been alternated since the previous sequencing run?	Cross-contamination of libraries between sequencing runs can cause false-positive results
High-throughput sequencing	Has the pooled library been appropriately sized and quantified?	Inaccurate sizing and quantification can cause overloading of flow cell and failed runs, or underloading and low data output
	Has the sequencer been appropriately cleaned between runs?	Insufficient cleaning of the sequencer can result in cross-contamination between runs
De-multiplexing and quality trimming	Has minimum sequencing depth been achieved for each sample?	Low sequencing depth can cause false-negative results
	Are an appropriate number of reads passing quality filtering?	Low numbers of reads passing quality filters can indicate issues with sequencing run and result in false-negative results
OTU clustering and denoising	How much of the original data are explained by the final OTUs/ASVs	Lower-than-expected sequences can indicate overly restrictive bioinformatics parameters
	Have chimeras and sequences with disrupted open reading frames been checked for? (for protein coding genes)	Chimeras and pseudogenes can inflate taxonomic diversity, leading to false-positive results
Taxonomic assignment	Has the reference database been curated to remove mislabelled taxonomy and pseudogenic sequences?	Mislabelled reference sequences can lead to both false-positive and false-negative results
	Has the taxonomy been applied with appropriate confidence levels?	Low-confidence assignment indicates incomplete or erroneous reference database
Interpretation of results	Have the taxa received an appropriate number of reads to pass detection threshold?	Taxa under detection threshold could represent laboratory or reagent contamination, or erroneous sequences that have not been sufficiently controlled for
	Has a minimum detection threshold been applied to remove index-switching?	Index-switching can cause spreading of taxa to other samples and result in false-positive results
	Are there any taxa that need to be confirmed with alternative methods?	Any high-risk putative detections should be confirmed with alternative method before reporting, if possible
Reporting and sign-off	Have any exceptions to laboratory standard operating procedure been made?	Non-compliances with standard operating procedure should be highlighted, and diagnostic confidence may be reduced
	Have data been stored appropriately?	Archiving of data allows future re-analysis in case of disputed results
	Have results been signed off by competent individual?	Incorrect reporting or interpretation of significant taxa can lead to incorrect managment response

datasets [266] would enable benchmarking of laboratory and computational methods and begin to characterize the sources of technical variation between laboratories [267, 268]. This could be further developed into an inter-laboratory proficiency testing program where blinded reference samples are periodically distributed for analysis, in order to demonstrate to all stakeholders that an assay is fit for purpose for detecting invasive insect species [248, 269]. The results of these processes would allow further development of best-practice technical guidelines and begin to harmonize approaches across the wider metabarcoding community [270].

Biosecurity and pest management decision making is still largely reliant on the application of a species name to a specimen barcode sequence [81], and issues of mislabelled sequences in public reference databases (Box 1) highlight the importance of maintaining expertise in taxonomy and classical diagnostics to complement high-throughput approaches. Owing to the incomplete nature of reference databases, much of the sequence data currently produced by metabarcoding assays will consist of insufficiently identified sequences [84]. While some of these will no doubt be the result of sequencing errors making it through quality control, many more will represent real taxa and reflect the further work required to more completely describe and acquire reference data for insect biodiversity. Monitoring programs for biological invasions are at their most informative when they are continuous and long term [271, 272], and it would be beneficial for these insufficiently identified sequences to be integrated into reference databases and tracked across analyses and timepoints. Porter and Hajibabaei [84] have highlighted the advantages that ASVs provide over more traditional OTU methods for consistent labelling of insufficiently identified sequences, and embracing non-destructive DNA extraction techniques would further enable taxonomists to verify these sequences using morphological methods and potentially locate previously unbarcoded taxa or novel species, which could then feed back into reference databases [259]. Conventional DNA barcoding and morphological taxonomy currently benefit from a close and reciprocal interaction [273], and we envision a similar relationship for the future of insect metabarcoding. This ability to systematically reanalyse historical datasets with improved reference databases, bioinformatic tools, and biological knowledge presents a major strength of HTS diagnostics [51], and therefore raw datasets should also be archived alongside relevant technical and environmental metadata in a machine-readable format [195]. However the datasets from ongoing longitudinal surveillance quickly amount to terabytes of data [274], the storage, management, and securing of which will require dedicated infrastructure and personnel [53]. Unlike the current drive for open sharing of data in academic research, concerns of misuse harming the international movement of goods means that historically the release of raw diagnostic data to the public has not been common [51]. However, a pathway for declassifying and releasing these data to researchers should be developed because the mass of community-level information generated by metabarcoding bio-surveillance shows great potential for generating new insights into the process and impacts of biological invasion [275].

In an increasingly globalized world, more effective and scalable utilization of surveillance effort will be required to manage the spread and establishment of invasive species. While broad-spectrum approaches to surveillance have historically been limited by the overwhelming amount of diagnostics work generated, metabarcoding-based diagnostics fundamentally change this dynamic by allowing entire communities of diverse organisms containing target pests, endemic species, and unexpected invaders to be simultaneously identified [41]. While present costs of technological investments may currently limit the uptake of HTS tools to only well-funded core diagnostic laboratories, we expect that developments in portable real-time sequencing will further enhance the availability of these tools to a much wider user-base worldwide. Furthermore, it is conceivable that the ongoing miniaturization of sequencers may synergize with advances in microfluidic and lab-on-a-chip technologies [276] to produce a new generation of metabarcoding-based “smart traps” for remote monitoring [277, 278]. Nevertheless, metabarcoding forms just a single component of a larger biosecurity toolbox that contains not only fast, cost-effective, and reliable means of diagnostics but also predictive models, improved risk forecasting, field-tested tools, and an overarching decision support system [46, 52, 135, 137]. The future of biosecurity surveillance and pest management is a distinctly interdisciplinary area, and we encourage future research to involve closer collaboration between academic scientists, diagnosticians, and the end users who rely on effective surveillance data to manage the spread of invasive pests and pathogens.

Methods

All articles containing ”Metabarcoding” in their abstract, title, or keywords were retrieved from the Scopus, PubMed, and Crossref citation databases on 20 June 2019 using the rscopus [279], rentrez [280], and fulltext [281] packages in R 3.5.3 [282]. Duplicated article entries were detected using fuzzy string matching functions from tidystringdist [283], and filtered out using dplyr [284]. All articles containing keywords in their title or abstract indicative of invasive species or sequencing platform used (see supplementary table 1 for full list of keywords) were then represented graphically by year of publication using ggplot2 [285]. A list of global insect pests was then retrieved from Ashfaq et al. [58] and combined with additional pests of concern for Australia [286]. This list was filtered to retain only unique and complete genus species binomials, retaining 558 species, for which all records for these species and the entire Insecta were retrieved from BOLD using the bold package [287]. The list of genes successfully retrieved from BOLD used to query GenBank and all records for species on the pest list and the entire Insecta were retrieved using the Rentrez R package [280]. Records from all databases were combined and specimen collection information was extracted using R and the biofiles package [288]. Of the 5,589,069 records for all loci in the datasets, 4,603,488 were annotated with latitude and longitude information and these were plotted on a world map using ggmap [289]. The number of overall records and unique species within all datasets were then plotted for the top 10 occurring loci.

Availability of supporting data and materials

A snapshot of the datasets and R markdown documents implementing the analyses contained in this manuscript are available in the Zenodo repository [290].

Additional files

Supplementary table 1: Keywords used to filter articles

Supplementary information 1: Reproducible R code used to conduct analyses and produce figure 1

Supplementary information 2: Reproducible R code used to conduct analyses and produce figure 3

Abbreviations

ASV: amplicon sequence variant; BLAST: Basic Local Alignment Search Tool; BOLD: Barcode of Life Data System; bp: base pairs; COI: cytochrome oxidase I; Gb: gigabase pairs; HTS: high-throughput sequencing; IPPC: International Plant Protection Convention; kb: kilobase pairs; Mb: megabase pairs; NCBI: National Center for Biotechnology Information; OIE: World Organisation of Animal Health; ONT: Oxford Nanopore Technologies; OTU: operational taxonomic unit; PacBio: Pacific Biosciences; rRNA: ribosomal RNA; SPS: World Trade Organisation Agreement on the Application of Sanitary and Phytosanitary measures; zOTU: zero-radius operational taxonomic unit.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported 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. Additional funding was provided by the Plant Biosecurity Cooperative Research Centre(PBCRC No. 2153), and Agriculture Victoria's Improved Market Access for Horticulture programme(CMI105584). A.M.P. and J.B. were further supported by an Australian Government Research Training Program Scholarship.

Authors' contributions

A.M.P. and M.J.B. conceptualized the manuscript. A.M.P. drafted the manuscript with contributions from J.B., J.W., J.P.C., N.O.I.C., B.C.R., and M.J.B. All authors read and approved the final manuscript.

Acknowledgements

We thank the reviewers for comments and suggestions that greatly improved this manuscript.

References

1. Hulme PE. Trade, transport and trouble: Managing invasive species pathways in an era of globalization. J Appl Ecol 2009;**46**:10–8.
2. Meyerson LA, Mooney HA. Invasive alien species in an era of globalization. Front Ecol Environ 2007;**5**:199–208.
3. Chown SL, Hodgins KA, Griffin PC, et al. Biological invasions, climate change and genomics. Evol Appl 2015;**8**:23–46.
4. Seebens H, Blackburn TM, Dyer EE, et al. Global rise in emerging alien species results from increased accessibility of new source pools. Proc Natl Acad Sci U S A 2018;**115**:E2264–73.
5. Paini DR, Sheppard AW, Cook DC, et al. Global threat to agriculture from invasive species. Proc Natl Acad Sci U S A 2016;**113**:7575–9.
6. Kenis M, Auger-Rozenberg MA, Roques A, et al. Ecological effects of invasive alien insects. Biol Invasions 2009;**11**:21–45.
7. Mazza G, Tricarico E, Genovesi P, et al. Biological invaders are threats to human health: An overview. Ethol Ecol Evol 2014;**26**(2-3):112–29.
8. Bradshaw CJA, Leroy B, Bellard C, et al. Massive yet grossly underestimated global costs of invasive insects. Nat Commun 2016;**7**:12986.
9. Andersen MC, Adams H, Hope B, et al. Risk assessment for invasive species. Risk Anal 2004;**24**:787–93.
10. Simberloff D, Martin JL, Genovesi P, et al. Impacts of biological invasions: What's what and the way forward. Trends Ecol Evol 2013;**28**:58–66.
11. Lodge DM, Simonin PW, Burgiel SW, et al. Risk analysis and bioeconomics of invasive species to inform policy and management. Annu Rev Environ Resour 2016;**41**:453–88.
12. Martin RR, Constable F, Tzanetakis IE. Quarantine regulations and the impact of modern detection methods. Annu Rev Phytopathol 2016;**54**:189–205.
13. Schrader G, Unger JG. 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. Biol Invasions 2003;**5**:357–64.
14. Early R, Bradley BA, Dukes JS, et al. Global threats from invasive alien species in the twenty-first century and national response capacities. Nat Commun 2016;**7**:12485.
15. Work TT, McCullough DG, Cavey JF, et al. Arrival rate of nonindigenous insect species into the United States through foreign trade. Biol Invasions 2005;**7**:323–32.
16. Joe Moffitt L, Stranlund JK, Osteen CD. Robust detection protocols for uncertain introductions of invasive species. J Environ Manage 2008;**89**:293–9.
17. Liebhold AM, Berec L, Brockerhoff EG, et al. Eradication of invading insect populations: From concepts to applications. Annu Rev Entomol 2016;**61**:335–52.
18. Trebitz AS, Hoffman JC, Darling JA, et al. Early detection monitoring for aquatic non-indigenous species: Optimizing surveillance, incorporating advanced technologies, and identifying research needs. J Environ Manage 2017;**202**:299–310.
19. Yemshanov D, Haight RG, Koch FH, et al. Optimizing surveillance strategies for early detection of invasive alien species. Ecol Econ 2019;**162**:87–99.
20. Epanchin-Niell RS, Haight RG, Berec L, et al. Optimal surveillance and eradication of invasive species in heterogeneous landscapes. Ecol Lett 2012;**15**:803–12.
21. Low-Choy S. Getting the story straight: Laying the foundations for statistical evaluation of the performance of surveillance. In: Jarrad F, Low-Choy S, Mengersen K , eds. Biosecurity Surveillance: Quantitative Approaches. 6th ed. CABI; 2015:43–73.
22. Whittle PJL, Stoklosa R, Barrett S, et al. A method for designing complex biosecurity surveillance systems: Detecting non-indigenous species of invertebrates on Barrow Island. Divers Distrib 2013;**19**:629–39.
23. Davidovitch L, Stoklosa R, Majer J, et al. Info-gap theory and robust design of surveillance for invasive species: The case study of Barrow Island. J Environ Manage 2009;**90**:2785–93.
24. Hodgetts J, Ostojá-Starzewski JC, Prior T, et al. DNA barcoding for biosecurity: Case studies from the UK plant protection program. Genome 2016;**59**:1033–48.
25. Armstrong KF, Ball SL. DNA barcodes for biosecurity: Invasive species identification. Philos Trans Biol Sci 2005;**360**:1813–23.
26. European and Mediterranean Plant Protection Organization. PM 7/129 (1) DNA barcoding as an identification tool for a number of regulated pests. EPPO Bull 2016;**46**:501–37.
27. Armstrong K. DNA barcoding: A new module in New

Zealand's plant biosecurity diagnostic toolbox. EPPO Bull 2010;**40**:91–100.

28. Anderson C, Low-Choy S, Whittle P, et al. Australian plant biosecurity surveillance systems. Crop Prot 2017;**100**:8–20.
29. Raghu S, Hulsman K, Clarke AR, et al. A rapid method of estimating cathes of abundant fruit fly species (Diptera: Tephritidae) in modified Steiner traps. Aust J Entomol 2000;**39**:15–9.
30. Morais P, Reichard M. Cryptic invasions: A review. Sci Total Environ 2018;**613–614**:1438–48.
31. Taberlet P, Coissac E, Pompanon F, et al. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol 2012;**21**:2045–50.
32. Bik HM, Porazinska DL, Creer S, et al. Sequencing our way towards understanding global eukaryotic biodiversity. Trends Ecol Evol 2012;**27**:233–43.
33. Tedersoo L, Drenkhan R, Anslan S, et al. High-throughput identification and diagnostics of pathogens and pests: Overview and practical recommendations. Mol Ecol Resour 2019;**19**:47–76.
34. Porter TM, Hajibabaei M. Scaling up: A guide to high throughput genomic approaches for biodiversity analysis. Mol Ecol 2018;**27**:313–38.
35. Deiner K, Bik HM, Mächler E, et al. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Mol Ecol 2017;**26**:5872–95.
36. Taberlet P, Bonin A, Zinger L, et al. Environmental DNA: For Biodiversity Research and Monitoring. Oxford University Press; 2017, doi:10.1093/oso/9780198767220.001.0001.
37. Alberdi A, Aizpurua O, Bohmann K, et al. Promises and pitfalls of using high-throughput sequencing for diet analysis. Mol Ecol Resour 2019;327–48.
38. Comtet T, Sandionigi A, Viard F, et al. DNA (meta)barcoding of biological invasions: A powerful tool to elucidate invasion processes and help managing aliens. Biol Invasions 2015;**17**:905–22.
39. Darling JA, Blum MJ. DNA-based methods for monitoring invasive species: A review and prospectus. Biol Invasions 2007;**9**:751–65.
40. Batovska J, Lynch SE, Cogan NOI, et al. Effective mosquito and arbovirus surveillance using metabarcoding. Mol Ecol Resour 2018;**18**:32–40.
41. Simmons M, Tucker A, Chadderton WL, et al. Active and passive environmental DNA surveillance of aquatic invasive species. Can J Fish Aquat Sci 2016;**73**:76–83.
42. Lawson Handley L. How will the "molecular revolution" contribute to biological recording? Biol J Linn Soc 2015;**115**:750–66.
43. Epanchin-Niell RS, Liebhold AM. Benefits of invasion prevention: Effect of time lags, spread rates, and damage persistence. Ecol Econ 2015;**116**:146–53.
44. Blackburn TM, Essl F, Evans T, et al. A unified classification of alien species based on the magnitude of their environmental impacts. PLoS Biol 2014;**12**:e1001850.
45. Deagle BE, Thomas AC, McInnes JC, et al. Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? Mol Ecol 2019;**28**:391–406.
46. Bilodeau P, Roe AD, Bilodeau G, et al. Biosurveillance of forest insects: Part II—adoption of genomic tools by end user communities and barriers to integration. J Pest Sci 2019;**92**:71–82.
47. European and Mediterranean Plant Protection Organization. PM 7/76 (4) Use of EPPO diagnostic protocols. EPPO Bull 2017;**47**:7–9.
48. World Trade Organization. Agreement on the Application of Sanitary and Phytosanitary Measures, 59–72. The results of the Uruguay Round of Multilateral Trade Negotiations: The Legal Texts, https://doi.org/10.1017/CB09780511818424 1999, Cambridge University Press.
49. Clover G, Hammons S, Unger JG. International diagnostic protocols for regulated plant pests. EPPO Bull 2010;**40**:24–9.
50. Thiermann AB. Globalization, international trade and animal health: The new roles of OIE. Prev Vet Med 2005:101–8.
51. Olmos A, Boonham N, Candresse T, et al. High-throughput sequencing technologies for plant pest diagnosis: Challenges and opportunities. EPPO Bull 2018;**48**:219–24.
52. Roe AD, Torson AS, Bilodeau G, et al. Biosurveillance of forest insects: Part I—integration and application of genomic tools to the surveillance of non-native forest insects. J Pest Sci 2019:51–70.
53. Food and Agriculture Organization of the UN. Preparing to use high-throughput sequencing (HTS) technologies as a diagnostic tool for phytosanitary purposes. Commission on Phytosanitary Measures Recommendation No 8. Rome; 2019. https://www.ippc.int/en/publications/87199/. Accessed on May 16, 2019.
54. OIE. Standards for high throughput sequencing, bioinformatics and computational genomics. OIE Terrestrial Manual 2019:88–93, www.oie.int/standard-setting/terrestrial-manual/access-online/. Accessed on May 16, 2019.
55. Zinger L, Bonin A, Alsos IG, et al. DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. Mol Ecol 2019;**28**:1857–62.
56. Freeland JR. The importance of molecular markers and primer design when characterizing biodiversity from environmental DNA. Genome 2017;**6**:358–74.
57. Folmer O, Black M, Hoeh W, et al. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 1994;**3**:294–9.
58. Ashfaq M, Hebert PDN, Naaum A. DNA barcodes for bio-surveillance: Regulated and economically important arthropod plant pests. Genome 2016;**59**:933–45.
59. Brandon-Mong G-J, Gan H-M, Sing K-W, et al. DNA metabarcoding of insects and allies: An evaluation of primers and pipelines. Bull Entomol Res 2015;**105**:717–27.
60. Hajibabaei M, Smith MA, Janzen DH, et al. A minimalist barcode can identify a specimen whose DNA is degraded. Mol Ecol Notes 2006;**6**:959–64.
61. Meusnier I, Singer GAC, Landry JF, et al. A universal DNA mini-barcode for biodiversity analysis. BMC Genomics 2008;**9**:4–7.
62. Elbrecht V, Braukmann TWA, Ivanova NV, et al. Validation of COI metabarcoding primers for terrestrial arthropods. PeerJ Preprints 2019;**7**:e27801v1, doi:10.7287/peerj.preprints.27801v2.
63. Deagle BE, Jarman SN, Coissac E, et al. DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. Biol Lett 2014;**10**:20140562.
64. Piñol J, Mir G, Gomez-Polo P, et al. Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. Mol Ecol Resour 2015;**15**:819–30.
65. Song H, Moulton MJ, Whiting MF. Rampant nuclear insertion of mtDNA across diverse lineages within Orthoptera (Insecta). PLoS One 2014;**9**:e110508.
66. Hlaing T, Tun-Lin W, Somboon P, et al. Mitochondrial pseudogenes in the nuclear genome of *Aedes aegypti* mosquitoes:

Implications for past and future population genetic studies. BMC Genet 2009;**10**:1–12.

67. Blacket MJ, Semeraro L, Malipatil MB. Barcoding Queensland fruit flies (*Bactrocera tryoni*): Impediments and improvements. Mol Ecol Resour 2012;**12**:428–36.
68. Bensasson D, Zhang DX, Hartl DL, et al. Mitochondrial pseudogenes: Evolution's misplaced witnesses. Trends Ecol Evol 2001;**16**:314–21.
69. Song H, Buhay JE, Whiting MF, et al. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc Natl Acad Sci U S A 2008;**105**:13486–91.
70. Jiang F, Jin Q, Liang L, et al. Existence of species complex largely reduced barcoding success for invasive species of Tephritidae: A case study in *Bactrocera* spp. Mol Ecol Resour 2014;**14**:1114–28.
71. Clarke LJ, Soubrier J, Weyrich LS, et al. Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. Mol Ecol Resour 2014;**14**:1160–70.
72. Gillespie JJ, Johnston JS, Cannonone JJ, et al. Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of *Apis mellifera* (Insecta: Hymenoptera): structure, organization, and retrotransposable elements. Insect Mol Biol 2006;**15**:657–86.
73. Zaidi F, Wei S, Shi M, et al. Utility of multi-gene loci for forensic species diagnosis of blowflies. J Insect Sci 2011;**11**:59.
74. Axtner J, Crampton-platt A, Lisa AH, et al. An efficient and robust laboratory workflow and tetrapod database for larger scale environmental DNA studies. Gigascience 2019;**8**(4), doi:10.1093/gigascience/giz029.
75. Zhang GK, Chain FJJ, Abbott CL, et al. Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. Evol Appl 2018;**11**:1901–14.
76. De Barba M, Miquel C, Boyer F, et al. DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. Mol Ecol Resour 2014;**14**:306–23.
77. Alberdi A, Aizpurua O, Gilbert MTP, et al. Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods Ecol Evol 2018;**9**:134–47.
78. Krosch MN, Schutze MK, Strutt F, et al. A transcriptome-based analytical workflow for identifying loci for species diagnosis: A case study with Bactrocera fruit flies (Diptera: Tephritidae). Austral Entomol 2017;**58**:395–408.
79. Floyd R, Lima J, de Waard J, et al. Common goals: Policy implications of DNA barcoding as a protocol for identification of arthropod pests. Biol Invasions 2010;**12**:2947–54.
80. Andújar C, Arribas P, Yu DW, et al. Why the COI barcode should be the community DNA metabarcode for the metazoa. Mol Ecol 2018;**27**:3968–75.
81. Boykin LM, Armstrong K, Kubatko L, et al. DNA barcoding invasive insects: Database roadblocks. Invertebr Syst 2012;**26**:506–14.
82. Ratnasingham S, Hebert PDN. BOLD : The Barcode of Life Data System (www.barcodinglife.org). Mol Ecol Notes 2007;**7**:355–64.
83. Benson DA, Cavanaugh M, Clark K, et al. GenBank. Nucleic Acids Res 2018;**46**:D41–7.
84. Porter TM, Hajibabaei M. Over 2.5 million sequences in GenBank and growing. PLoS One 2018;**13**:e0200177.
85. Liu S, Yang C, Zhou C, et al. Filling reference gaps via assembling DNA barcodes using high-throughput sequencing—Moving toward barcoding the world. Gigascience 2017;**6**(12), doi:10.1093/gigascience/gix104.
86. Shen YY, Chen X, Murphy RW. Assessing DNA barcoding as a tool for species identification and data quality control. PLoS One 2013;**8**:e57125.
87. Mioduchowska M, Jan M, Gołdyn B, et al. Instances of erroneous DNA barcoding of metazoan invertebrates: Are universal cox1 gene primers too "universal"? PLoS One 2018;**13**:e0199609.
88. Galan M, Pons JB, Tournayre O, et al. Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. Mol Ecol Resour 2018;**18**:474–89.
89. Bengtsson-Palme J, Boulund F, Edström R, et al. Strategies to improve usability and preserve accuracy in biological sequence databases. Proteomics 2016;**16**:2454–60.
90. Batovska J, Blacket MJ, Brown K, et al. Molecular identification of mosquitoes (Diptera: Culicidae) in southeastern Australia. Ecol Evol 2016;**6**:3001–11.
91. Collins RA, Cruickshank RH. The seven deadly sins of DNA barcoding. Mol Ecol Resour 2013;**13**:969–75.
92. Castalanelli MA, Severtson DL, Brumley CJ, et al. A rapid non-destructive DNA extraction method for insects and other arthropods. J Asia Pac Entomol 2010;**13**:243–8.
93. Carew ME, Nichols SJ, Batovska J, et al. A DNA barcode database of Australia's freshwater macroinvertebrate fauna. Mar Freshw Res 2017;**68**:1788–802.
94. Kocher A, Gantier JC, Gaborit P, et al. Vector soup: High-throughput identification of neotropical phlebotomine sand flies using metabarcoding. Mol Ecol Resour 2017;**17**:172–82.
95. Bergqvist J, Forsman O, Larsson P, et al. Detection and isolation of sindbis virus from mosquitoes captured during an outbreak in Sweden, 2013. Vector Borne Zoonotic Dis 2015;**15**:133–40.
96. Somervuo P, Koskela S, Pennanen J, et al. Unbiased probabilistic taxonomic classification for DNA barcoding. Bioinformatics 2016;**32**:2920–7.
97. Bokulich NA, Kaehler BD, Rideout JR, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 2018;**6**:90.
98. Rodgers TW, Xu CCY, Giacalone J, et al. Carrion fly-derived DNA metabarcoding is an effective tool for mammal surveys: Evidence from a known tropical mammal community. Mol Ecol Resour 2017;**17**:e133–45.
99. Machida RJ, Leray M, Ho SL, et al. Data Descriptor: Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Sci Data 2017;**4**:170027.
100. Richardson R, Bengtsson-Palme J, Gardiner MM, et al. A reference cytochrome c oxidase subunit I database curated for hierarchical classification of arthropod metabarcoding data. PeerJ 2018;**6**:e5126.
101. Porter TM, Hajibabaei M. Automated high throughput animal CO1 metabarcode classification. Sci Rep 2018;**8**:4226.
102. Kozlov AM, Zhang J, Yilmaz P, et al. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. Nucleic Acids Res 2016;**44**:5022–33.
103. Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet 2019;**20**:341–55.
104. Pawluczyk M, Weiss J, Links MG, et al. Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. Anal Bioanal Chem 2015;**407**:1841–8.

105. Piñol J, Senar MA, Symondson WOC. The choice of universal primers and the characteristics of the species mixture determines when DNA metabarcoding can be quantitative. Mol Ecol 2019;**28**:407–19.
106. Rennstam Rubbmark O, Sint D, Horngacher N, et al. A broadly-applicable COI primer pair and an efficient single tube amplicon library preparation protocol for metabarcoding. Ecol Evol 2018;**8**:12335–50.
107. Bylemans J, Gleeson DM, Hardy CM, et al. Toward an ecoregion scale evaluation of eDNA metabarcoding primers: A case study for the freshwater fish biodiversity of the Murray-Darling Basin (Australia). Ecol Evol 2018;**8**:8697–712.
108. Ficetola GF, Coissac E, Zundel S, et al. An in silico approach for the evaluation of DNA barcodes. BMC Genomics 2010;**11**:434.
109. Elbrecht V, Leese F. Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. Front Environ Sci 2017;**5**:11.
110. Elbrecht V, Leese F. PrimerMiner: An R package for development and in silico validation of DNA metabarcoding primers. Methods Ecol Evol 2017;**8**:622–6.
111. Marquina D, Andersson AF, Ronquist F. New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. Mol Ecol Resour 2019;**19**(1):90–104.
112. Corse E, Tougard C, Archambaud-Suard G, et al. One-locus-several-primers: A strategy to improve the taxonomic and haplotypic coverage in diet metabarcoding studies. Ecol Evol 2019;**9**:4603–20.
113. Krehenwinkel H, Wolf M, Lim JY, et al. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Sci Rep 2017;**7**:17668.
114. Nichols R V, Vollmers C, Newsom LA, et al. Minimizing polymerase biases in metabarcoding. Mol Ecol Resour 2018;**18**:927–39.
115. Krehenwinkel H, Pomerantz A, Henderson JB, et al. Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. Gigascience 2019;**8**(5), doi:10.1093/gigascience/giz006.
116. Elbrecht V, Peinert B, Leese F. Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. Ecol Evol 2017;**7**:6918–26.
117. Braukmann TWA, Ivanova N V, Prosser SWJ, et al. Metabarcoding a diverse arthropod mock community. Mol Ecol Resour 2019;**19**:711–27.
118. Thomas AC, Deagle BE, Eveson JP, et al. Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. Mol Ecol Resour 2016;**16**:714–26.
119. Mclaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. bioRxiv 2019, doi:10.1101/559831.
120. Silverman JD, Bloom RJ, Jiang S, et al. Measuring and mitigating PCR bias in microbiome data. bioRxiv 2019, doi:10.1101/604025.
121. Crampton-Platt A, Yu DW, Zhou X, et al. Mitochondrial metagenomics: letting the genes out of the bottle. Gigascience 2016;**5**, doi:10.1186/s13742-016-0120-y.
122. Gómez-Rodríguez C, Crampton-Platt A, Timmermans MJTN, et al. Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. Methods Ecol Evol 2015;**6**:883–94.
123. Tang M, Hardman CJ, Ji Y, et al. High-throughput monitoring of wild bee diversity and abundance via mitogenomics. Methods Ecol Evol 2015;**6**:1034–43.
124. Linard B, Crampton-Platt A, Moriniere J, et al. The contribution of mitochondrial metagenomics to large-scale data mining and phylogenetic analysis of Coleoptera. Mol Phylogenet Evol 2018;**128**:1–11.
125. Papadopoulou A, Taberlet P, Zinger L. Metagenome skimming for phylogenetic community ecology: A new era in biodiversity research. Mol Ecol 2015;**24**:3515–7.
126. Arribas P, Andújar C, Hopkins K, et al. Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. Methods Ecol Evol 2016;**7**:1071–81.
127. Mamanova L, Coffey AJ, Scott CE, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods 2010;**7**:111–8.
128. Jones MR, Good JM. Targeted capture in evolutionary and ecological genomics. Mol Ecol 2016;**25**:185–202.
129. Macher JN, Zizka VMA, Weigand AM, et al. A simple centrifugation protocol for metagenomic studies increases mitochondrial DNA yield by two orders of magnitude. Methods Ecol Evol 2018;**9**:1070–4.
130. Dowle EJ, Pochon X, C. Banks J, et al. Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: A case study using freshwater macroinvertebrates. Mol Ecol Resour 2016;**16**:1240–54.
131. Wilcox TM, Zarn KE, Piggott MP, et al. Capture enrichment of aquatic environmental DNA: A first proof of concept. Mol Ecol Resour 2018;**18**:1392–401.
132. Peñalba J V, Smith LL, Tonione MA, et al. Sequence capture using PCR-generated probes: A cost-effective method of targeted high-throughput sequencing for nonmodel organisms. Mol Ecol Resour 2014;**14**:1000–10.
133. Liu S, Wang X, Xie L, et al. Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. Mol Ecol Resour 2016;**16**:470–9.
134. Wilson JJ, Brandon-Mong GJ, Gan HM, et al. High-throughput terrestrial biodiversity assessments: Mitochondrial metabarcoding, metagenomics or metatranscriptomics? Mitochondrial DNA A DNA Mapp Seq Anal 2019;**30**:490–9.
135. Poland TM, Rassati D. Improved biosecurity surveillance of non-native forest insects: A review of current methods. J Pest Sci 2019;**92**:37–49.
136. Bulman SR, McDougal RL, Hill K, et al. Opportunities and limitations for DNA metabarcoding in Australasian plant-pathogen biosecurity. Australas Plant Pathol 2018;**47**:467–74.
137. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 2018;**19**:9–20.
138. Batovska J, Lynch SE, Rodoni BC, et al. Metagenomic arbovirus detection using MinION nanopore sequencing. J Virol Methods 2017;**249**:79–84.
139. Gibson J, Shokralla S, Porter TM, et al. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc Natl Acad Sci U S A 2014;**111**:8007–12.
140. Whitfield AE, Falk BW, Rotenberg D. Insect vector-mediated transmission of plant viruses. Virology 2015;**479–480**:278–89.
141. Miller KE, Hopkins K, Inward DJG, et al. Metabarcoding of fungal communities associated with bark beetles. Ecol Evol 2016;**6**:1590–600.

142. Orlovskis Z, Canale MC, Thole V, et al. Insect-borne plant pathogenic bacteria: Getting a ride goes beyond physical contact. Curr Opin Insect Sci 2015;**9**:16–23.

143. Sint D, Raso L, Traugott M. Advances in multiplex PCR: Balancing primer efficiencies and improving detection success. Methods Ecol Evol 2012;**3**:898–905.

144. Callahan BJ, Wong J, Heiner C, et al. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res 2019, doi:10.1093/nar/gkz569.

145. Tedersoo L, Anslan S. Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences. Environ Microbiol Rep 2019, doi:10.1111/1758-2229.12776.

146. Arulandhu AJ, Staats M, Hagelaar R, et al. Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. Gigascience 2017;**6**:(10), doi:10.1093/gigascience/gix080).

147. Swift JF, Lance RF, Guan X, et al. Multifaceted DNA metabarcoding: Validation of a noninvasive, next-generation approach to studying bat populations. Evol Appl 2018;**11**:1120–38.

148. Daborn PJ. A single P450 allele associated with insecticide resistance in *Drosophila*. Science 2002;**297**:2253–6.

149. Stapley J, Santure AW, Dennis SR. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Mol Ecol 2015;**24**:2241–52.

150. Ricciardi A, Blackburn TM, Carlton JT, et al. Invasion science: A horizon scan of emerging challenges and opportunities. Trends Ecol Evol 2017;**32**:464–74.

151. Saitoh S, Aoyama H, Fujii S, et al. A quantitative protocol for DNA metabarcoding of springtails (Collembola). Genome 2016;**59**:705–23.

152. Elbrecht V, Leese F. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. PLoS One 2015;**10**:e0130324.

153. Gohl DM, Vangay P, Garbe J, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol 2016;**34**:942–9.

154. Sinha R, Stanley G, Gulati GS, et al. Index switching causes “spreading-of-signal” among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. bioRxiv 2017, doi:10.1101/125724.

155. Wick RR, Judd LM, Holt KE. Deepbinner : Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. PLoS Comput Biol 2018;**14**:e1006583.

156. Carlsen T, Aas AB, Lindner D, et al. Don’t make a mista(g)ke: Is tag switching an overlooked source of error in amplicon pyrosequencing studies? Fungal Ecol 2012;**5**:747–9.

157. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 2012;**40**:1–8.

158. Tedersoo L, Tooming-Klunderud A, Anslan S. PacBio metabarcoding of fungi and other eukaryotes: Errors, biases and perspectives. New Phytol 2018;**217**:1370–85.

159. Costello M, Fleharty M, Abreu J, et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 2018;**19**:1–10.

160. Li Q, Zhao X, Zhang W, et al. Reliable multiplex sequencing with rare index mis-assignment on DNB-based NGS platform. BMC Genomics 2019;**20**:1–13.

161. Illumina. Effects of index misassignment on multiplexing and downstream analysis. 2017. https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf. Accessed on 19 Feb 2018.

162. Schnell IB, Bohmann K, Gilbert MTP. Tag jumps illuminated - Reducing sequence-to-sample misidentifications in metabarcoding studies. Mol Ecol Resour 2015;**15**:1289–303.

163. Hanna RE, Doench JG. A case of mistaken identity. Nat Biotechnol 2018;**36**:802–4.

164. Nguyen NH, Smith D, Peay K, et al. Parsing ecological signal from noise in next generation amplicon sequencing. New Phytol 2015;**205**:1389–93.

165. MacConaill LE, Burns RT, Nag A, et al. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics 2018;**19**:30.

166. Bartram J, Mountjoy E, Brooks T, et al. Accurate sample assignment in a multiplexed, ultrasensitive, high-throughput sequencing assay for minimal residual disease. J Mol Diagnostics 2016;**18**:494–506.

167. Faircloth BC, Glenn TC. Not all sequence tags are created equal: Designing and validating sequence identification tags robust to indels. PLoS One 2012;**7**:e42543.

168. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet 2016;**17**:333–51.

169. Bleidorn C. Third generation sequencing: Technology and its potential impact on evolutionary biodiversity research. Syst Biodivers 2016;**14**:1–8.

170. Benítez-Páez A, Portune KJ, Sanz Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION^{TM} portable nanopore sequencer. Gigascience 2016;**5**:4.

171. van Dijk EL, Jaszczyszyn Y, Naquin D, et al. The third revolution in sequencing technology. Trends Genet 2018;**34**:666–81.

172. Hebert PDN, Braukmann TWA, Prosser SWJ, et al. A sequel to Sanger: Amplicon sequencing that scales. BMC Genomics 2018;**19**:1–14.

173. Calus ST, Ijaz UZ, Pinto AJ. NanoAmpli-Seq: A workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. Gigascience 2018;**7**:(12, doi:10.1093/gigascience/giy140). .

174. Volden R, Palmer T, Byrne A, et al. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. Proc Natl Acad Sci U S A 2018;**115**:9726–31.

175. Murray DC, Coghlan ML, Bunce M. From benchtop to desktop: Important considerations when designing amplicon sequencing workflows. PLoS One 2015;**10**:e0124671.

176. Scott R, Zhan A, Brown EA, et al. Optimization and performance testing of a sequence processing pipeline applied to detection of nonindigenous species. Evol Appl 2018;891–905.

177. Palmer JM, Jusino MA, Banik MT, et al. Non-biological synthetic spike-in controls and the AMPtk software pipeline improve fungal high throughput amplicon sequencing data. PeerJ 2017;213470.

178. Bolyen E, Rideout JR, Dillon MR, et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Preprints 2018;**6**:e27295v2, doi:10.7287/peerj.preprints.27295v2.

179. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: Open-source, platform-independent, community-

supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;**75**:7537–41.

180. Boyer F, Mercier C, Bonin A, et al. obitools: A unix-inspired software package for DNA metabarcoding. Mol Ecol Resour 2016;**16**:176–82.
181. Rognes T, Flouri T, Nichols B, et al. VSEARCH: A versatile open source tool for metagenomics. PeerJ 2016;**4**:e2584.
182. Pauvert C, Buée M, Laval V, et al. Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. Fungal Ecol 2019;**41**:23–33.
183. Flynn JM, Brown EA, Chain FJJ, et al. Toward accurate molecular identification of species in complex environmental samples: Testing the performance of sequence filtering and clustering methods. Ecol Evol 2015;**5**:2252–66.
184. Majaneva M, Hyytiäinen K, Varvio SL, et al. Bioinformatic amplicon read processing strategies strongly affect eukaryotic diversity and the taxonomic composition of communities. PLoS One 2015;**10**:e0130035.
185. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat Rev Genet 2018;**19**:269–85.
186. Edgar RC, Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 2015;**31**:3476–82.
187. Ewing B, Hillier LD, Wendl MC. Base-calling of automated sequencer traces using Phred. Genome Res 1998;**8**:186–94.
188. Bokulich NA, Subramanian S, Faith JJ, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 2013;**10**:57–9.
189. Potapov V, Ong JL. Examining sources of error in PCR by single-molecule sequencing. PLoS One 2017;**12**:e0169774.
190. Elbrecht V, Hebert PDN, Steinke D. Slippage of degenerate primers can cause variation in amplicon length. Sci Rep 2018;**8**:10999.
191. Schirmer M, Ijaz UZ, D’Amore R, et al. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res 2015;**43**:e37.
192. Meyer CP, Paulay G. DNA barcoding: Error rates based on comprehensive sampling. PLoS Biol 2005;**3**:1–10.
193. Hebert PDN, Ratnasingham S, de Waard JR. Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. Proc Biol Sci 2003;**270**, doi:10.1098/rsbl.2003.0025.
194. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 2017;**11**:2639–43.
195. Tedersoo L, Ramirez KS, Nilsson RH, et al. Standardizing metadata and taxonomic identification in metabarcoding studies. Gigascience 2015;**4**:34.
196. Callahan BJ, McMurdie PJ, Rosen MJ, et al. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 2016;**13**:581–3.
197. Amir A, Daniel M, Navas-Molina J, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2017;**2**:e00191–16.
198. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv 2016, doi:10.1101/081257.
199. Marshall NT, Stepien CA. Invasion genetics from eDNA and thousands of larvae: A targeted metabarcoding assay that distinguishes species and population variation of zebra and quagga mussels. Ecol Evol 2019;**9**:3515–38.
200. Edgar RC, Haas BJ, Clemente JC. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 2011;**27**:2194–200.
201. Haas BJ, Gevers D, Earl AM, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 2011;**21**:494–504.
202. Brown EA, Chain FJJ, Crease TJ, et al. Divergence thresholds and divergent biodiversity estimates: Can metabarcoding reliably describe zooplankton communities? Ecol Evol 2015;**5**:2234–51.
203. Decelle J, Romac S, Sasaki E, et al. Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing. PLoS One 2014;**9**:e104297.
204. Turon X, Antich A, Palacín C, et al. From metabarcoding to metaphylogeography: Separating the wheat from the chaff. bioRxiv 2019, doi:10.1101/629535.
205. Olds BP, Jerde CL, Renshaw MA, et al. Estimating species richness using environmental DNA. Ecol Evol 2016;**6**:4214–26.
206. Gardner PP, Watson RJ, Morgan XC, et al. Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. PeerJ 2019;**7**:e6160.
207. Bazinet AL, Cummings MP. A comparative evaluation of sequence classification programs. BMC Bioinformatics 2012;**13**:92.
208. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990;**215**:403–10.
209. Koski LB, Golding GB. The closest BLAST hit is often not the nearest neighbor. J Mol Evol 2001;**52**:540–2.
210. Virgilio M, Backeljau T, Nevado B, et al. Comparative performances of DNA barcoding across insect orders. BMC Bioinformatics 2010;**11**:206.
211. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 2007;**73**:5261–7.
212. Huson D, Auch A, Qi J, et al. MEGAN analysis of metagenome data. Genome Res 2007;**17**:377–86.
213. Wilkinson SP, Davy SK, Bunce M, et al. Taxonomic identification of environmental DNA with informatic sequence classification trees. PeerJ Preprints 2018;**6**:e26812v1, doi:10.7287/peerj.preprints.26812v1.
214. Lan Y, Wang Q, Cole JR, et al. Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. PLoS One 2012;**7**:e32491.
215. Edgar R. SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences. bioRxiv 2016, doi:10.1101/074161.
216. Somervuo P, Yu DW, Xu CCY, et al. Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. Methods Ecol Evol 2017;**8**:398–407.
217. Burgar JM, Murray DC, Craig MD, et al. Who’s for dinner? High-throughput sequencing reveals bat dietary differentiation in a biodiversity hotspot where prey taxonomy is largely undescribed. Mol Ecol 2014;**23**:3605–17.
218. Secretariat of the International Plant Protection Convention (IPPC). ISPM 4 Requirements for the establishment of pest free areas. 2017. https://www.ippc.int/en/publications/requirements-establishment-pest-free-areas/. Accessed on May 7, 2019.
219. Champlot S, Berthelot C, Pruvost M, et al. An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. PLoS One

2010;**5**:e13042.

220. Miller S, Naccache SN, Samayoa E, et al. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res 2019;**29**:831–42.
221. European and Mediterranean Plant Protection Organization Organisation. Basic requirements for quality management in plant pest diagnosis laboratories. EPPO Bull 2007;**37**:580–8.
222. Gu W, Miller S, Chiu CY. Clinical metagenomic sequencing for pathogen detection. Annu Rev Pathol Mech Dis 2019;**14**:319–38.
223. Elbrecht V, Steinke D. Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring. Freshw Biol 2019;**64**:380–7.
224. Ficetola GF, Taberlet P, Coissac E. How to limit false positives in environmental DNA and metabarcoding? Mol Ecol Resour 2016;**16**:604–7.
225. Klymus KE, Marshall NT, Stepien CA. Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. PLoS One 2017;**12**:e0177643.
226. Wilson CC, Wozney KM, Smith CM. Recognizing false positives: Synthetic oligonucleotide controls for environmental DNA surveillance. Methods Ecol Evol 2016;**7**:23–9.
227. Ji Y, Huotari T, Roslin T, et al. SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and abundances using DNA barcodes or mitogenomes. Mol Ecol Resour 2019, doi:10.1111/1755-0998.13057.
228. Wright ES, Vetsigian KH. Quality filtering of Illumina index reads mitigates sample cross-talk. BMC Genomics 2016;**17**:876.
229. Zepeda-Mendoza ML, Bohmann K, Carmona Baez A, et al. DAMe: A toolkit for the initial processing of datasets with PCR replicates of double-tagged amplicons for DNA metabarcoding analyses. BMC Res Notes 2016;**9**:255.
230. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 2014;**12**:87.
231. Davis NM, Proctor DM, Holmes SP, et al. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 2018;**6**:226.
232. McKnight DT, Huerlimann R, Bower DS, et al. microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. Environ DNA 2019;**1**:14–25.
233. Larsson AJM, Stanley G, Sinha R, et al. Computational correction of index switching in multiplexed sequencing libraries. Nat Methods 2018;**15**:305–7.
234. Yeh Y-C, Needham DM, Sieradzki ET, et al. Taxon disappearance from microbiome analysis reinforces the value of mock communities as a standard in every sequencing run. mSystems 2018;**3**:e00023–18.
235. Hardwick SA, Chen WY, Wong T, et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. Nat Commun 2018;**9**:3096.
236. Leray M, Knowlton N. Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. PeerJ 2017;**5**:e3006.
237. Ficetola GF, Pansu J, Bonin A, et al. Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. Mol Ecol Resour 2015;**15**:543–56.
238. Mata VA, Rebelo H, Amorim F, et al. How much is enough? Effects of technical and biological replication on metabarcoding dietary analysis. Mol Ecol 2019;**28**:165–75.
239. Guillera-Arroita G. Modelling of species distributions, range dynamics and communities under imperfect detection: Advances, challenges and opportunities. Ecography 2017;**40**:281–95.
240. Lahoz-Monfort JJ, Guillera-Arroita G, Tingley R. Statistical approaches to account for false-positive errors in environmental DNA samples. Mol Ecol Resour 2016;**16**:673–85.
241. Krehenwinkel H, Fong M, Kennedy S, et al. The effect of DNA degradation bias in passive sampling devices on metabarcoding studies of arthropod communities and their associated microbiota. PLoS One 2018;**13**:e0189188.
242. Berec L, Kean JM, Epanchin-Niell R, et al. Designing efficient surveys: Spatial arrangement of sample points for detection of invasive species. Biol Invasions 2014;**17**:445–59.
243. European and Mediterranean Plant Protection Organization. PM 7/98 (2) Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity. EPPO Bull 2010;**44**:117–47.
244. National Association of Testing Authorities, Technical Note 17 - Guidelines for the validation and verification of quantitative and qualitative test methods . https://www.nata.com.au/phocadownload/gen-accreditation-guidance/Validation-and-Verification-of-Quantitative-and-Qualitative-Test-Methods.pdf . Accessed on December 6, 2018 . 2012.
245. Blaser S, Diem H, von Felten A, et al. From laboratory to point of entry: Development and implementation of a loop-mediated isothermal amplification (LAMP)-based genetic identification system to prevent introduction of quarantine insect species. Pest Manag Sci 2018;**74**:1504–12.
246. Schlaberg R, Chiu CY, Miller S, et al. Validation of metagenomic next-generation sequencing tests for universal pathogen detection. Arch Pathol Lab Med 2017;**141**:776–86.
247. Adams IP, Fox A, Boonham N, et al. The impact of high throughput sequencing on plant health diagnostics. Eur J Plant Pathol 2018;**152**(4):909–19.
248. Gargis AS, Kalman L, Lubin IM. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. J Clin Microbiol 2016;**54**:2857–65.
249. Hatzenbuhler C, Kelly JR, Martinson J, et al. Sensitivity and accuracy of high-throughput metabarcoding methods for early detection of invasive fish species. Sci Rep 2017;**7**:46393.
250. Bell KL, Burgess KS, Botsch JC, et al. Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. Mol Ecol 2018;**28**:431–55.
251. Smith DP, Peay KG. Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. PLoS One 2014;**9**:e90234.
252. Landolt PJ, Adams T, Davis TS, et al. Spotted wing drosophila, *Drosophila suzukii* (Diptera: Drosophilidae), trapped with combinations of wines and vinegars. Florida Entomol 2012;**95**:326–32.
253. Lindahl T. Instability and decay of the primary structure of DNA. Nature 1993;**362**:709–15.
254. Brown EA, Chain FJJ, Zhan A, et al. Early detection of aquatic invaders using metabarcoding reveals a high number of non-indigenous species in Canadian ports. Divers Distrib 2016;**22**:1045–59.
255. Clarke AR, Li Z, Qin Y, et al. *Bactrocera dorsalis* (Hendel)

(Diptera: Tephritidae) is not invasive through Asia: It's been there all along. J Appl Entomol 2019;**00**:1–5.
256. Callan SK, Majer JD, Edwards K, et al. Documenting the terrestrial invertebrate fauna of Barrow Island, Western Australia. Aust J Entomol 2011;**50**:323–43.
257. Massart S, Candresse T, Gil J, et al. A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. Front Microbiol 2017;**8**:45.
258. Carew ME, Coleman RA, Hoffmann AA. Can non-destructive DNA extraction of bulk invertebrate samples be used for metabarcoding? PeerJ 2018;**6**:e4980.
259. Ritter CD, Häggqvist S, Karlsson D, et al. Biodiversity assessments in the 21st century: The potential of insect traps to complement environmental samples for estimating eukaryotic and prokaryotic diversity using high-throughput DNA metabarcoding. Genome 2019;**62**:147–59.
260. Nielsen M, Gilbert MTP, Pape T, et al. A simplified DNA extraction protocol for unsorted bulk arthropod samples that maintains exoskeletal integrity. Environ DNA 2019;**00**:1–11.
261. Martins FMS, Galhardo M, Filipe AF, et al. Have the cake and eat it: Optimising nondestructive DNA metabarcoding of macroinvertebrate samples for freshwater biomonitoring. Mol Ecol Resour 2019;**19**(4):863–76.
262. Zizka VMA, Leese F, Peinert B, et al. DNA metabarcoding from sample fixative as a quick and voucher-preserving biodiversity assessment method. Genome 2018;**62**:122–36.
263. Hajibabaei M, Spall JL, Shokralla S, et al. Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. BMC Ecol 2012;**12**:28.
264. Linard B, Arribas P, Andújar C, et al. Lessons from genome skimming of arthropod-preserving ethanol. Mol Ecol Resour 2016;**16**:1365–77.
265. McIntyre ABR, Ounit R, Afshinnekoo E, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol 2017;**18**:1–19.
266. Duncavage EJ, Abel HJ, Pfeifer JD. In silico proficiency testing for clinical next-generation sequencing. J Mol Diagn 2017;**19**:35–42.
267. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. Nat Rev Genet 2017;**18**:473–84.
268. Sinha R, Abu-Ali G, Vogtmann E, et al. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nat Biotechnol 2017;**35**:1077–86.
269. Schrijver I, Aziz N, Jennings LJ, et al. Methods-based proficiency testing in molecular genetic pathology. J Mol Diagn 2014;**16**:283–7.
270. Knight R, Vrbanac A, Taylor BC, et al. Best practices for analysing microbiomes. Nat Rev Microbiol 2018;**16**:410–22.
271. Latombe G, Pyšek P, Jeschke JM, et al. A vision for global monitoring of biological invasions. Biol Conserv 2017;**213**:295–308.
272. MacLeod A. The relationship between biosecurity surveillance and risk analysis. In: Jarrad F, Low-Choy S, Mengersen K , eds. Biosecurity Surveillance Quantitative Approaches. CABI; 2015:109–20.
273. Schlick-Steiner BC, Steiner FM, Seifert B, et al. Integrative taxonomy: A multisource approach to exploring biodiversity. Annu Rev Entomol 2010;**55**:421–38.
274. Stephens ZD, Lee SY, Faghri F, et al. Big data: Astronomical or genomical? PLoS Biol 2015;**13**:1–11.
275. Evans DM, Kitson JJN, Lunt DH, et al. Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems. Funct Ecol 2016;1904–16.
276. Lafleur JP, Jönsson A, Senkbeil S, et al. Recent advances in lab-on-a-chip for biosensing applications. Biosens Bioelectron 2016;**76**:213–33.
277. Potamitis I, Eliopoulos P, Rigakis I. Automated remote insect surveillance at a global scale and the Internet of Things. Robotics 2017;**6**:19.
278. Bohan DA, Vacher C, Tamaddoni-Nezhad A, et al. Next-generation global biomonitoring: Large-scale, automated reconstruction of ecological networks. Trends Ecol Evol 2017;**32**:477–87.
279. Muschelli J. rscopus: Scopus Database "API" Interface 2018. https://github.com/muschellij2/rscopus.
280. Winter DJ. rentrez: An R package for the NCBI eUtils API. R J 2019;**9**:520.
281. Chamberlain S. fulltext: Full Text of 'Scholarly' Articles Across Many Data Sources. 2019. R package version 1.3.0. https://cran.r-project.org/web/packages/fulltext .
282. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017. http://www.r-project.org/.
283. Fay C. String distance calculation the tidy way. 2019. https://github.com/ColinFay/tidystringdist.
284. Wickham H, Francois R, Henry L, et al.. dplyr: A grammar of data manipulation. 2019, R package version 0.8.3. https://cran.r-project.org/web/packages/dplyr/.
285. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag; 2016. https://ggplot2.tidyverse.org/ .
286. Plant Health Australia, The National Plant Biosecurity Status Report, 130–137. http://www.planthealthaustralia.com.au/national-programs/national-plant-biosecurity-status-report/ . Accessed on Dec 21 2018. 2017.
287. Chamberlain S. bold: Interface to Bold Systems API. 2017. R package version 0.9.0 https://cran.r-project.org/package=bold.
288. Schöfl G. biofiles: An Interface for GenBank/GenPept Flat Files. R package version 1.0.0 https://github.com/gschofl/biofiles.
289. Kahle D, Wickham H. ggmap: Spatial Visualization with ggplot2. R J 2013;**5**:144–61.
290. Piper AM. Supplementary S2: Prospects and challenges of implementing DNA metabarcoding for high throughput insect surveillance (Version 2.0). zenodo 2019, doi:10.5281/zenodo.3252736.

3
Computational Evaluation of DNA Metabarcoding for Universal Diagnostics of Invasive Insect Pests

3.1 Chapter preface

This chapter uses in-silico methods to establish the taxonomic breadth across which COI “mini-barcodes” can achieve species level resolution, and condense the many published metabarcoding primers into a shortlist of those suitable for diagnostic use. To achieve this, a computational pipeline for curating reference sequence data is developed and applied to all insect COI sequences publicly available on the NCBI GenBank and BOLD repositories. Using the resulting curated database, the diagnostic sensitivity and taxonomic bias is evaluated in-silico for 63 published metabarcoding primers, together with 5 novel primers designed in this chapter. Four of the highest performing primer combinations identified in this chapter are then compared on real mixed trap samples in Chapter 4, and the reference sequence database generated here is used for species identification in both Chapters 4 and 5. This chapter is presented as a self-contained manuscript in the final stages of preparation, with intended submission to the journal *Molecular Ecology Resources*, and includes supplementary material at the end.

3.2 Publication details:

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

Stage of publication: In Preparation

Journal details: Molecular Ecology Resources

Authors: Alexander M. Piper, Noel O.I. Cogan, John Paul Cunningham, Mark J. Blacket

3.3 Statement of joint authorship:

A.M.P. conceptualised the study, performed all analyses, and wrote the first draft of the manuscript with input and supervision from J.P.C., N.O.I.C., and M.J.B. All authors contributed to the editing of the final manuscript and approved the version presented here.

Statement from co-author confirming the contribution of the PhD candidate:

"As co-author of the manuscript 'Piper, A. M., Cogan N.O.I, Cunningham, J. P. & Blacket M.J. (In preparation). Computational Evaluation of DNA Metabarcoding for Universal Diagnostics of Invasive Insect Pests, *Molecular Ecology Resources*', I confirm that Alexander M. Piper has made the contributions listed above."

Associate Professor John Paul Cunningham

30/03/2021

3.4 Manuscript

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

Alexander M. Piper[1,2], Noel O.I. Cogan[1,2], John Paul Cunningham[1,2], Mark J. Blacket[1]

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

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

Keywords: DNA barcoding, amplicon sequencing, assay validation, phylogenetics, NGS sequencing, biosecurity

Running title: Computational evaluation of insect pest metabarcoding

Corresponding author:

Alexander M. Piper

Email: alexander.piper@agriculture.vic.gov.au

Abstract

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

Introduction

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

Plant pest and pathogen diagnostics currently rely on a mixture of morphological examination, biochemical techniques, and molecular assays such as diagnostic qPCR, and DNA barcoding (EPPO, 2019a). While these methods provide highly accurate identification for small numbers of specimens (Armstrong & Ball, 2005; Darling & Blum, 2007), their inherent restriction to analysing single specimens per-reaction limits their application to large mixed samples collected in surveillance traps (Batovska, Piper, Valenzuela, Cunningham, & Blacket, 2020; Carnegie & Nahrung, 2019). As an alternative, high-throughput sequencing (HTS) platforms can comprehensively characterise mixed populations of genomic DNA (metagenomics), RNA (metatranscriptomics) or taxonomically informative marker genes (metabarcoding), allowing whole communities to be identified without any prior isolation or specimen sorting (Piper et al., 2019; Tedersoo, Drenkhan, Anslan, Morales-Rodriguez, & Cleary, 2019). While first emerging for exploring biodiversity (Handelsman, 2004; Taberlet, Coissac, Pompanon, Brochmann, &

Willerslev, 2012), broad-scope HTS assays have recently been co-opted by various disciplines of molecular diagnostics, where their potential to act as a universal identification tool was quickly recognised (Adams et al., 2009; Comtet, Sandionigi, Viard, & Casiraghi, 2015). By removing the requirement to separately develop and maintain hundreds of targeted diagnostic assays, universal HTS diagnostics could substantially expand the range of organisms within the scope of a diagnostic laboratory, as well as decrease the costs of implementation (Adams, Fox, Boonham, Massart, & De Jonghe, 2018; Allcock, Jennison, & Warrilow, 2017).

Metabarcoding of the mitochondrial cytochrome c oxidase subunit 1 (COI) gene presents the most readily adoptable HTS approach for diagnostics of insect pests, due to its cost effectiveness, extensive public reference sequence databases, and ability to leverage widespread acceptance of DNA barcoding within regulatory frameworks (Andújar, Arribas, Yu, Vogler, & Emerson, 2018; Comtet et al., 2015; Piper et al., 2019). While conventional single-specimen DNA barcoding targets a 709 bp region of COI (Folmer, Black, Hoeh, Lutz, & Vrijenhoek, 1994), modern HTS platforms impose strict length limitations on sequenced molecules, and therefore "mini-barcodes" must instead be used (Brandon-Mong et al., 2015). The number of diagnostic nucleotides contained within these mini-barcodes largely determines the sensitivity and specificity for single specimens, but mismatch between PCR primers and variable template molecules can bias amplification and cause dropouts of low-abundance taxa within mixed community samples (Elbrecht & Leese, 2015; D. W. Yu et al., 2012). Primer-template mismatch is a particular issue for protein-coding genes such as COI where variability in the third position of each codon leaves no strictly conserved regions for placement of universal PCR primers (Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014). Therefore, degenerate nucleotide bases are commonly incorporated into COI metabarcoding primers to account for this inevitable mismatch (Elbrecht & Leese, 2017b), though overuse can result in undesired amplification of non-target organisms (Collins et al., 2019; Leese et al., 2021).

Historically, molecular diagnostic assays would have undergone stringent laboratory validation in order to resolve the aforementioned issues and establish performance parameters for every target designated in an unambiguously defined scope (EPPO, 2019b).

However, when considering the sheer number of potential targets, hosts and matrices that would need to be evaluated for a universal assay, it is evident that many validation processes cannot be applied in their traditional sense and must be adapted to novel HTS based diagnostics (Maree, Fox, Al Rwahnih, Boonham, & Candresse, 2018; Roenhorst et al., 2018). In-silico methods pose a promising alternative that can leverage public reference sequence data to establish the diagnostic performance of a target barcode region and determine the best placement of degenerate PCR primers, all without requiring physical specimens (Elbrecht & Leese, 2017a; Ficetola et al., 2010). Nevertheless, the use of public reference data comes with some caveats, as issues of mislabelled taxonomic annotations, insufficiently identified specimens, and contamination with non-homologous loci are well documented (Garg, Leipe, & Uetz, 2019; Locatelli, McIntyre, Therkildsen, & Baetscher, 2020; Pentinsaari, Ratnasingham, Miller, & Hebert, 2020; Siddall, Fontanella, Watson, Kvist, & Erséus, 2009). It is therefore essential for public DNA barcode sequences to be appropriately curated before use for in-silico validation procedures or as reference databases for metabarcoding analysis (Piper et al., 2019).

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

Methods

Retrieval and curation of public reference data

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

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

Construction of phylogenetic tree

The curated Insect reference sequences were supplemented with an outgroup of 15 COI sequences from Diplura, the sister taxa to Insecta (Misof et al., 2014), and all positions in the alignment that contained gaps in >95% of sequences were masked. A maximum likelihood phylogenetic tree was generated using *FastTree* (Price, Dehal, & Arkin, 2009) following the General Time-Reversible (GTR) model (Tavaré, 1986) and gamma model of

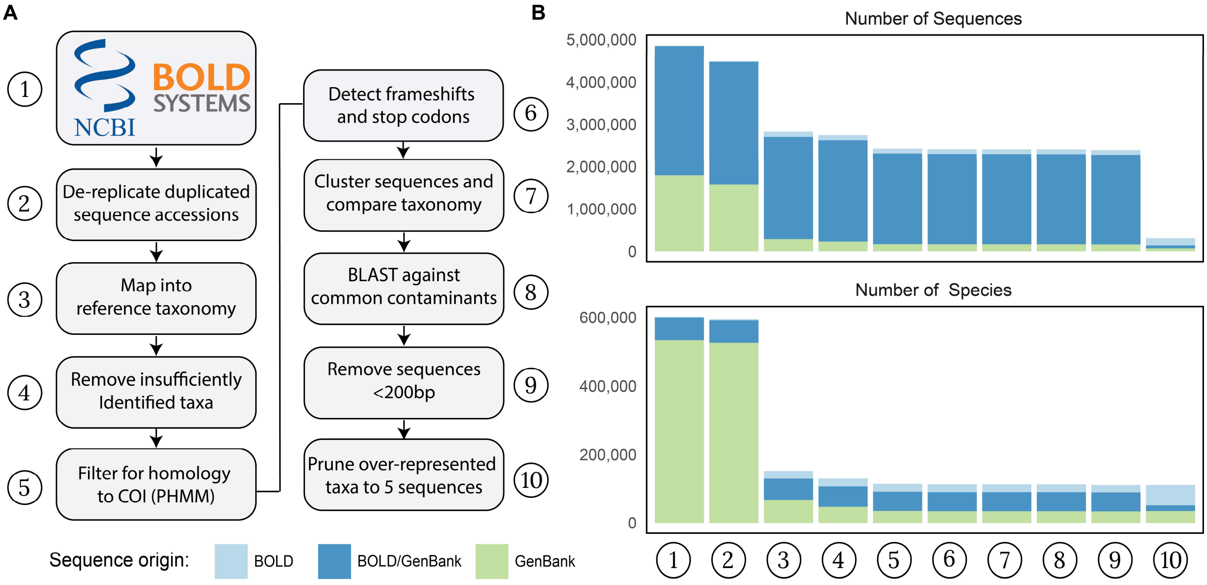


Figure 1 – **A)** Overview of the computational pipeline used to curate public reference sequence data for primer evaluation. **B)** Number of sequences (upper) and species (lower) retained after each curation step, and their origin from BOLD, GenBank or duplicated across both repositories.

rate heterogeneity across sites, with taxonomic identities at the domain, phylum, class and order ranks used to constrain the deeper topology of the tree. The constructed phylogeny was rooted on the edge connecting the Diplura outgroup to the rest of the tree and made ultrametric using the *geiger* R package (Pennell et al., 2014) and PATHd8 (Britton, Anderson, Jacquet, Lundqvist, & Bremer, 2007). All phylogenetic trees were plotted using the *ggtree* (G. Yu, Lam, Zhu, & Guan, 2018; G. Yu, Smith, Zhu, Guan, & Lam, 2017) and *ggplot2* R packages (Wickham, 2016).

Diagnostic information within the COI barcode

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

Characteristics of published and novel primers

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

Diagnostic sensitivity of mini-barcodes

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

Primer-template mismatch

A mismatch score was calculated between each primer and every sequence that contained an appropriate binding site using PrimerMiner (Elbrecht & Leese, 2017a). The default penalties as per Stadhouders et al., (2010) were used to score types of mismatches, with penalty scores doubled for each contiguous mismatch and increased exponentially towards 3' end of primer. To determine the phylogenetic scale at which primer-template mismatch is conserved, $1{\times}10^8$ pairs of tips were randomly selected from the tree and binned into 100 discrete intervals of phylogenetic distance. The phylogenetic autocorrelation function (Zaneveld & Thurber, 2014), or how the value of a trait (primer mismatch score) decays with increasing phylogenetic divergence was then calculated separately for each primer using the *castor* R package (Louca & Doebeli, 2018). As primer-template mismatch was found to be phylogenetically conserved (supplementary Figure 4), phylogenetic independent contrasts (Felsenstein, 1985) was used to impute mismatch scores for species that did not have available sequence data within their respective primer binding sites (Zaneveld & Thurber, 2014). The accuracy of phylogenetic imputation is determined by the depth at which the trait is conserved (measured by the decay of the autocorrelation function), as well as the distance from the tip being imputed to the nearest clade with available data, which was quantified for each primer set using the Nearest Sequence Taxon Index (NSTI) averaged over all tips in the phylogeny (Langille et al., 2013). To determine the importance of primer degeneracy for reducing bias, a linear regression model was fit separately to the imputed and unimputed mismatch data for all forward and reverse primers.

Final primer rankings

To obtain an overall ranking for each evaluated primer, the identification success for both the pest and entire insect datasets, as well as the inverse of the primer mismatch score were z-normalised to be on the same scale. Additionally, each forward and reverse primer was assigned a value of either -1 (poor), 0 (moderate), or 1 (good), depending on how closely their physical characteristics adhered to common primer design recommendations (Supplementary Table 1) (Abd-Elsalam, 2003; Kwok, Chang, Sninsky, & Wang, 1994; Shen et al., 2010), and these were also normalised. The standardised scores from each metric were then weighted by their relative importance for overall primer performance; 1× for mismatch, 0.5× for pest insect identification, 0.5× for all insect identification, and 0.25× for each of the measured physical characteristics; fold-degeneracy, primer length, melting temperature, GC%, presence of GC clamps and longest homopolymer, then summed by primer to obtain a final ranking.

Results

Sequence database assembly

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

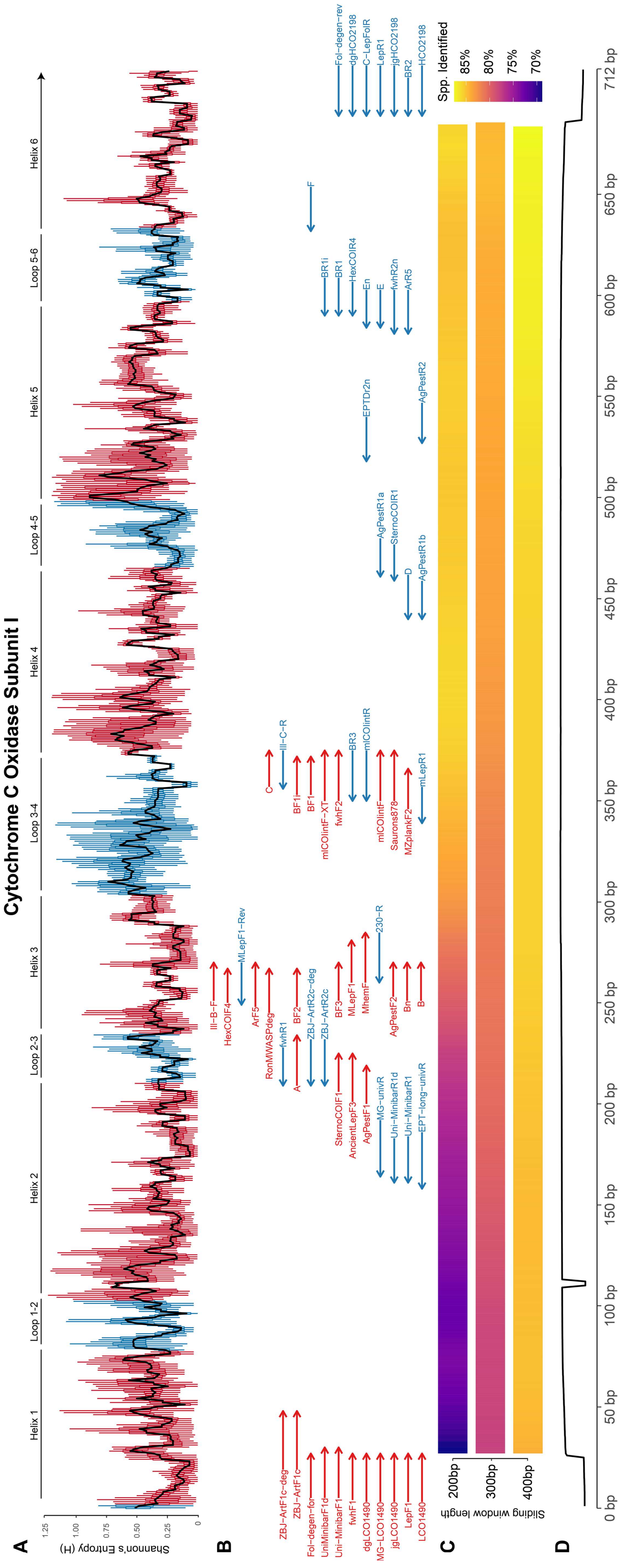


Figure 2 - Summary of the COI barcode region. **A)** Boxplots of Shannon's entropy per nucleotide position calculated separately for each Insect order in the database, with structural motifs annotated as per Pentinsaari et al. (2016). **B)** Binding positions of all published and novel primers evaluated in this study. **C)** Identification success for all insects within sliding windows of length 200 bp, 300 bp and 400 bp. **D)** Number of sequences at each position within the COI barcode locus.

a local database of contaminants removed 30 Wolbachia and 56 human mitochondrial sequences and identified a further 1,667 sequences that were >98% similar to known COI pseudogenes but did not contain any characteristic stop codon or frameshift mutations. All sequences <200 bp in length were then removed, leaving 2,389,404 sequences from 110,676 distinct species remaining. When compared to globally relevant lists of invasive insects, a total of 2,625 species spanning 1,490 genera and 20 taxonomic orders were identified within the curated database (supplementary Figure 2). Each pest species as represented by a significantly higher number of sequences (mean 79.9 ± 5.60) than the average insect taxon (mean 20.2

± 0.428) (Welch t-test: $t_{(2837)} = 11.1$, $p < .001$), however no reference data was available for 1,717 of the listed species. As the number of sequences per species was greatly skewed by a few highly sampled taxa (supplementary Figure 3), the database was pruned to a maximum of 5 sequences per species. This left a total of 315,754 sequences from 110,676 species remaining in the final curated reference database (Figure 1B).

Optimal diagnostic subregions within the COI barcode

The final curated reference database consisted of a 712 bp alignment (Figure 2), which included the 709 bp barcode region and a 3bp insertion at position 110-112 that occurred in the order Thysanoptera as well as some Hymenopteran species. Sequence coverage was relatively even across the COI barcode, with exception of the terminal regions where the standard practice of removing priming sequences before submission to public repositories resulted in low coverage (Figure 2D). The per-site Shannon entropy within regions coding for loop structures was significantly higher than within those coding for transmembrane helices (Welch t-test: $t_{(12306)} = 4.0459$, $p < .001$). However, short segments of low entropy were distributed throughout both (Figure 2A). The large majority of the 58 primers retrieved from the literature were designed around these few low-entropic segments, with many overlapping around 180 bp, 255 bp, 370 bp, and 600 bp into the alignment (Figure 2B). As none of these segments of low entropy extended for the 18-24 bp length of a typical primer, many of the evaluated primers included multiple degenerate bases to account for variable positions (Supplementary Table 1). Sliding window analyses revealed that for a 200 bp or 300 bp amplicon, those positioned towards the 3' end of the barcode region could differentiate substantially more species than those towards the 5' end, however differences were much less pronounced for a 400 bp amplicon (Figure 2C).

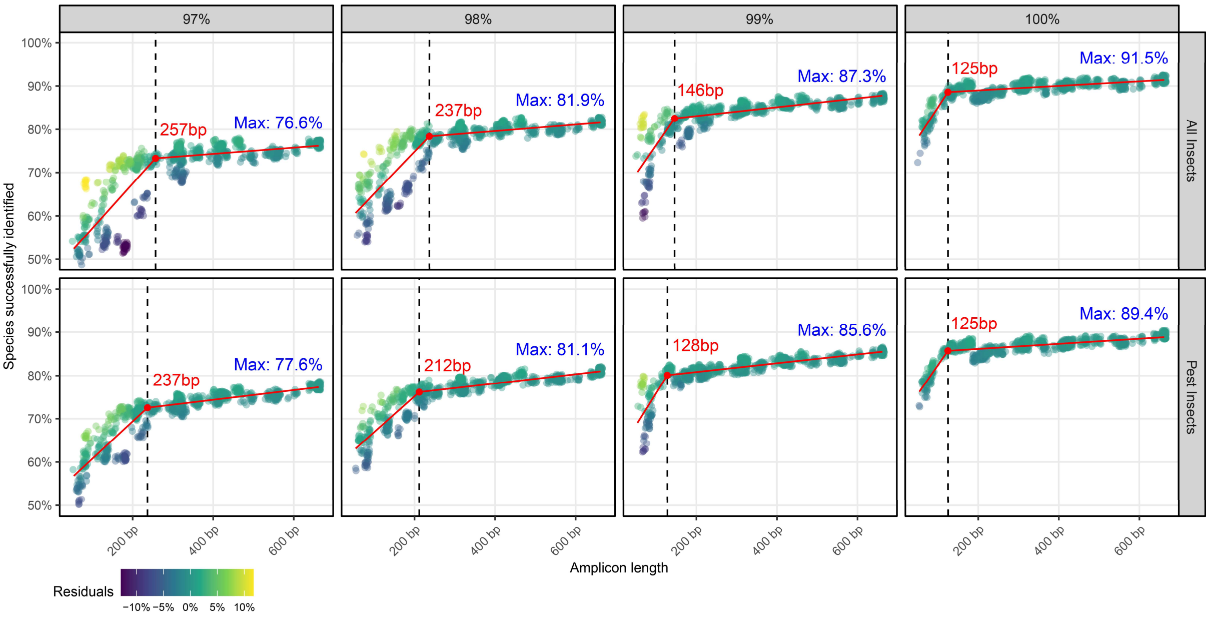


Figure 3) Identification success for all insects (upper) and Insect pests (lower) as a function of amplicon length for all possible combination of mini-barcode primers. Percentage identity threshold used to identify species increases from 97% (leftmost panel) to 100% (rightmost panel), with segmented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

Diagnostic sensitivity of mini-barcodes

PCR amplification was simulated for the 1156 primer combinations that could produce an amplicon >50 bp, as well as for the full-length barcode region. Identification success using mini-barcodes generally increased with amplicon length, but did not follow a simple linear relationship, instead seeing a sharp initial increase up to a certain length, followed by a second more gradual slope (Figure 3). A segmented regression model applied to the whole insect dataset inferred the change point between these trends to be 257 bp at the 97% identity threshold (95% CI: 242 bp – 266 bp), 237 bp at the 98% threshold (95% CI: 218 bp – 248 bp), 146 bp at the 99% threshold (95% CI: 137 bp – 206 bp) and 125 bp when only 100% matches were considered (95% CI: 110 bp – 131 bp). On the other hand, the inferred changepoints for the pest dataset were approximately 20 bp smaller at the 97%, 98% and 99% identity thresholds, but identical to the larger dataset at the 100% threshold. In spite of this general trend, certain amplicons deviated up to 10% above or below the regression line for both datasets (Figure 3), reinforcing that appropriate placement of mini-barcodes within the COI barcode region can be just as important as amplicon length for diagnostic performance.

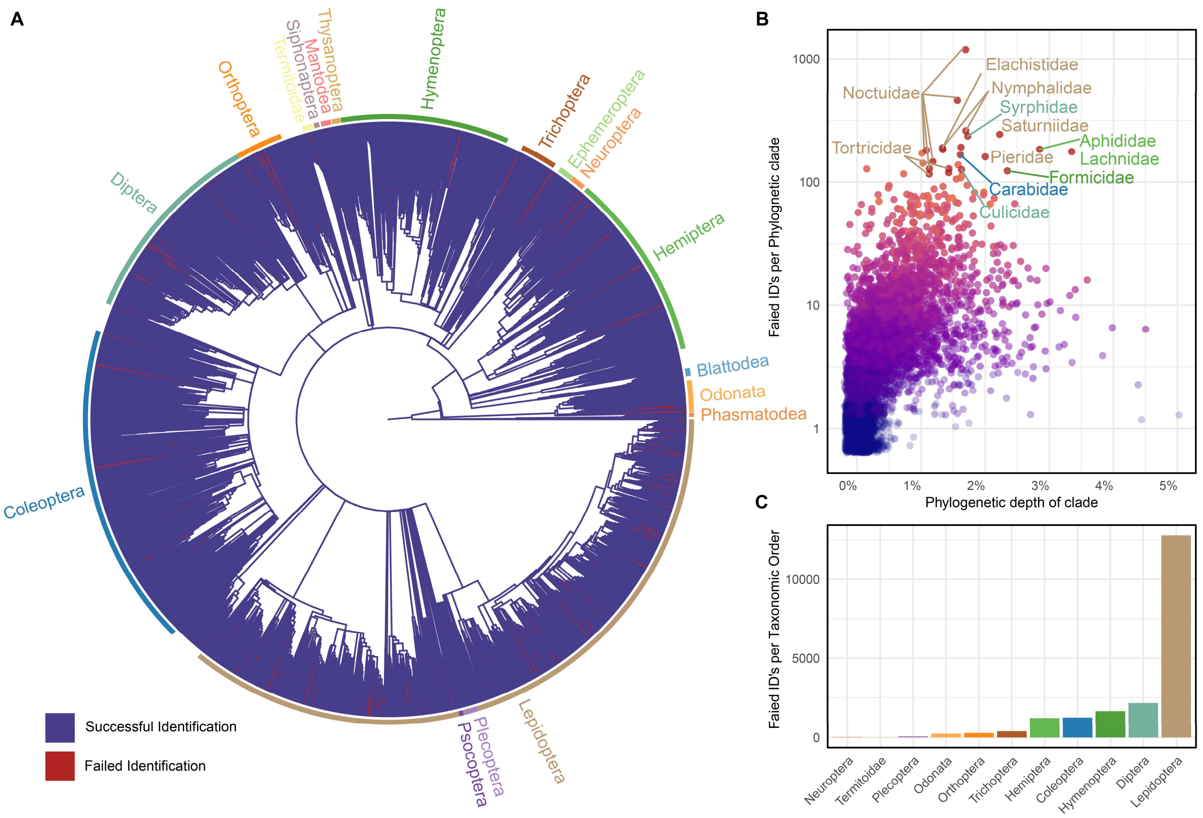


Figure 4: A) Phylogenetic tree of all insect genera contained within the curated reference database with major orders annotated. Clades are highlighted in red where ≥50% of species could not be successfully differentiated from their congenerics by the full-length barcode a 97% identity threshold. **B)**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

The overall proportion of insect species that could be successfully identified by any of the barcodes increased with the stringency of the percentage identity threshold used (Figure 3). For the complete insect dataset, the proportion identified increased from 76.6% when using a 97% identification threshold up to 91.5% when only 100% matches were considered, and similarly increased from 77.6% to 89.4% in the pest dataset. Taxa that could not be identified even by the full-length barcode showed significant phylogenetic clustering (Figure 4A), with the mean phylogenetic depth of problem clades (those in which ≥50% of species couldn't be differentiated from their congenerics) ranging from 0.83% at a 97% identity threshold (consenTRAIT, $p < .001$) to 0.21% at 100% identity threshold ($p < .001$). These patterns indicate that identification failure with DNA barcoding can be considered a phylogenetically conserved trait, but concentrated in smaller clades scattered throughout the tips of the phylogeny rather than being inherent

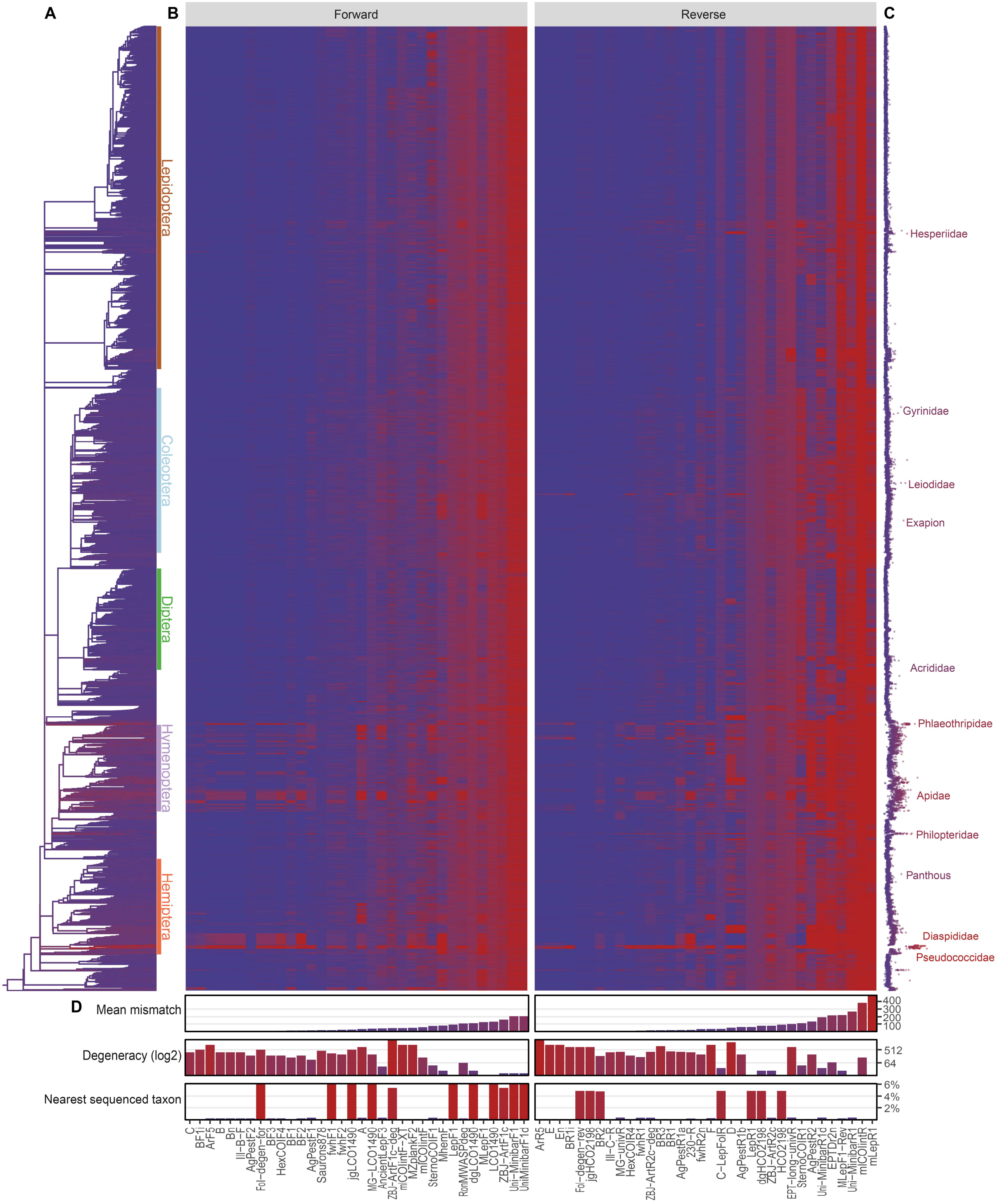


Figure 5: A) Phylogenetic tree of all insect genera contained within the curated sequence database, coloured by their mean primer-template mismatch across all evaluated primers. **B)** Primer-template mismatch for each evaluated sequence and primer, summarised by insect genus. **C)** Dot-plot of mean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

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

Predicted bias for metabarcoding primers

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

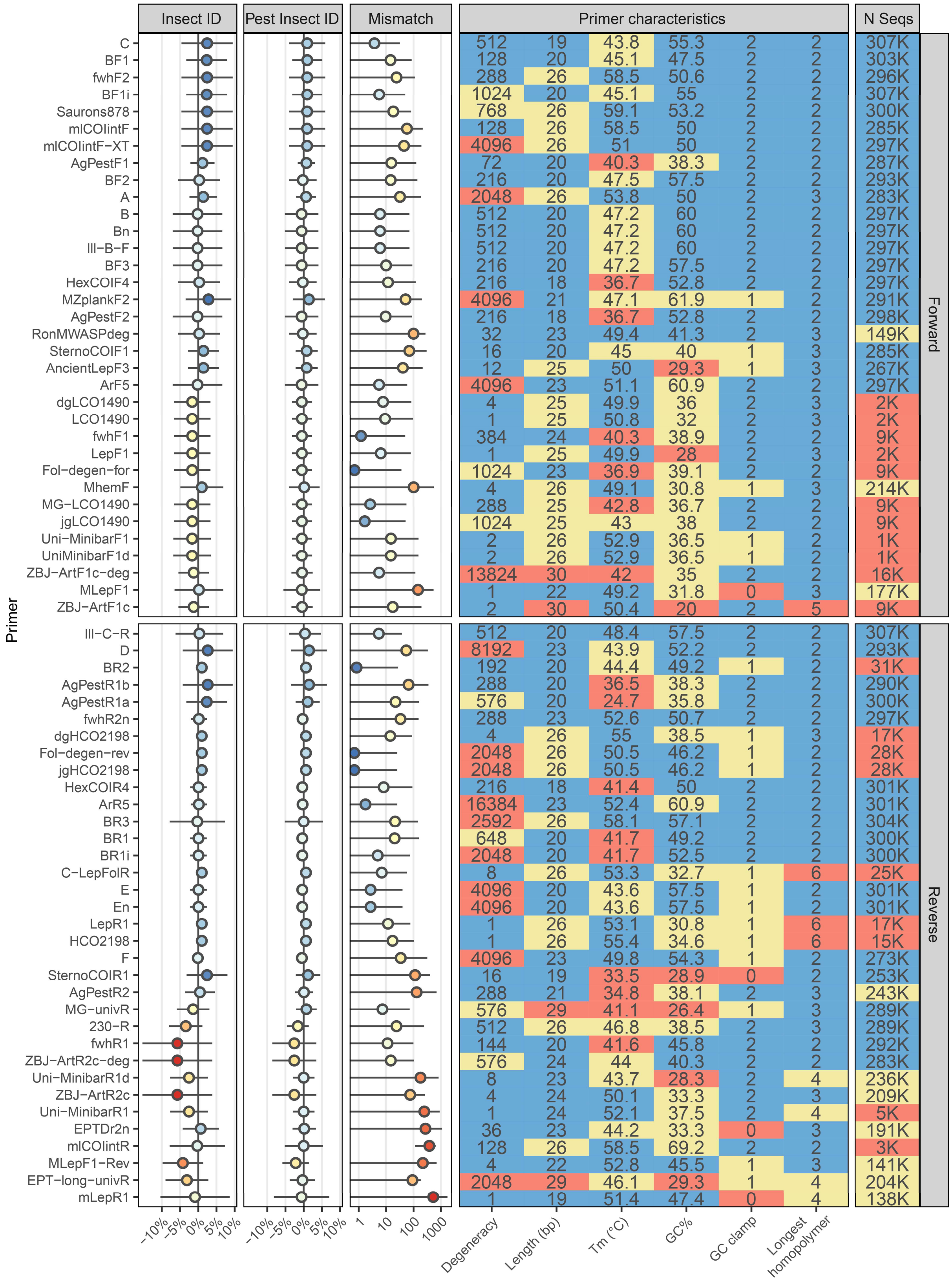


Figure 6) Summary of primer performance across all metrics measured within this study, with primers ranked from best overall performance (top) to worst (bottom). Metrics from left to right: Average performance of primer compared to regression model predictions on all insects and pest insects, mean mismatch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

extremely high degeneracy still showed moderate mismatch across the insect phylogeny

(Figure 5D). The underperformance of highly degenerate primers such as ZBJ-ArtF1c-deg remained when only the unimputed data was viewed (Supplementary Figure 5), suggesting that these results were not confounded by difficulties imputing higher levels of missing data.

Final primer rankings

Many of the evaluated forward primers performed well across all metrics, with C, BF1, fwhF2, BF1i, SauronS878 and MlCOIintF all ranking highly (Figure 6). Nevertheless, there was no perfect forward primer, with many of those showing higher diagnostic sensitivity also having excessive mismatches, or physical characteristics outside of recommended design guidelines (Figure 6). In contrast, there was substantially more variability in the performance of reverse primers, with primers such as D, AgPestR1a and AgPestR1b showing exceptional diagnostic sensitivity, but having either too much degeneracy or a melting temperature well below the recommended guidelines. On the other hand, the reverse primers Ill-C-R, fwhR2n, and BR2 showed slightly less sensitivity, but adhered well to recommended physical characteristics (Figure 6). To comply with restrictions of 2 × 150 bp sequencing, the primer combinations fwhF2-fwhR2n, BF1-BR1, or SauronS878-fwhR2n present the best overall options, amplifying a ~250 bp subregion of COI that contains the most diagnostic nucleotides (Figure 2A, C), while showing little mismatch across all insects and physical characteristics within recommended guidelines. The novel AgPestF1-AgPestR1b primers designed in this study would also provide an appropriate 260 bp amplicon for 2 × 150 bp sequencing, however further laboratory evaluation would be required to ensure their lower than recommended melting temperatures does not introduce non-specific amplification. On the other hand, a much greater range of primer combinations are appropriate for sequencing technologies that can deliver read lengths of 2 × 300 bp. In particular, those that amplify a subregion from 250 bp into the full-length barcode, along to either the low entropy region around 600 bp, or onwards the 3' terminus will capture the most diagnostic nucleotides (Figure 2C). Published primer combinations that amplify these subregions and performed well across all evaluated metrics include HexCOIF4-HexCOIR4, BF2-BR2 and BF3-BR2, but many of the alternative forward or reverse primers which overlap the same positions would also prove suitable (Figure 2B). Nevertheless, once the amplicon has reached approximately 400 bp there is minimal difference in species discrimination across the COI barcode (Figure 2C), and therefore primer combinations that amplify from the 5' end of the barcode onwards would likely also perform adequately for 2 × 300 bp sequencing.

Discussion

High sensitivity across a broad taxonomic scope is a defining feature of DNA metabarcoding assays, facilitating their use as a universal diagnostic assay to rapidly screen mixed samples for a range of target pests or pathogens. Despite several studies applying metabarcoding to certain pest taxa (Batovska et al., 2018, 2020; Bowser et al., 2019; Young, Milián-García, Yu, Bullas-Appleton, & Hanner, 2021), the diagnostic performance of the required mini-barcodes has not until now been systematically evaluated across the broader diversity of invasive insect pests. Using a curated reference database covering 110,676 insect species, including 2,625 species registered on global invasive species lists, we here demonstrate that mini-barcodes can achieve comparable resolution to the full-length COI barcode region already widely accepted within insect diagnostic protocols. Our findings are largely in agreement with previous investigations showing congruence between mini-barcodes and the full-length barcode (Hajibabaei, Smith, et al., 2006; Meusnier et al., 2008), as well as morphospecies (Yeo, Srivathsan, & Meier, 2020). However, our study expands these predictions to a more than five-fold larger sample of insect taxa, with an additional focus on invasive insect pests.

While 97% identity is considered the default threshold for delineating taxonomic units from DNA barcodes (Alberdi et al., 2018; Porter & Hajibabaei, 2020), our analyses demonstrate that a more stringent 98% or 99% identification threshold not only increases the number of insects that can successfully be differentiated, but also reduces the amplicon length required to do so. This is particularly notable for implementing metabarcoding on production scale HTS platforms such as the Illumina NovaSeq, which offer the lowest cost per sample but require much shorter amplicons due to their typical read lengths of only 2 × 150 bp (Piper et al., 2019). Shorter barcodes also improve recovery of taxa when DNA is degraded, which can occur when traps are deployed in the field for extended periods of time without adequate preservative (Krehenwinkel et al., 2018). While use of more stringent identification thresholds has been constrained by the common practice of clustering metabarcoding reads to resolve sequencing errors (Porter & Hajibabaei, 2020), recent denoising algorithms provide single nucleotide resolution that can be leveraged for more accurate and reproducible taxonomic assignment (Callahan, McMurdie, & Holmes, 2017; Porter & Hajibabaei, 2020). Nevertheless, the metric of identification success used within our study (proportion of clusters containing only a single species) does not consider the actual availability of reference sequences to match unknown taxa against, and false negatives may be introduced when using these more stringent thresholds if intraspecific diversity isn't sufficiently represented in the reference database. While this will not be an issue for most invasive insect pests due to their general overrepresentation in public sequence repositories, some taxa were represented by only single sequences, and a further 1,717 had no publicly available COI data whatsoever. Moreover, the 110,676 insect species represented in our curated database barely accounts for 10% of described insect diversity (Stork, 2018), highlighting the considerable efforts still required to increase the taxonomic coverage of reference databases before metabarcoding assays can operate in a truly universal manner.

While many of the evaluated mini-barcodes performed comparably to the full-length barcode region, 10.6% of the insect pests, and 8.5% of all insect species could not be differentiated at all, even when solely considering 100% matches. While this may at first glance seem high, it is largely consistent with previous research that suggests between 12.3% and 26.5% of described Arthropod species are non-monophyletic for the COI barcode region (Funk & Omland, 2003; Mutanen et al., 2016; Ross, 2014). In our study, these misidentified taxa were phylogenetically clustered in "problem clades" towards the tips of the phylogeny, which predominantly occurred within the order Lepidoptera. Issues of DNA barcode failure for Lepidoptera and other speciose taxonomic groups has long been noted (Meier, Shiyang, Vaidya, & Ng, 2006; Wiemers & Fiedler, 2007), but only recently has it been appreciated how much of this can be attributed to underlying misidentifications, databasing errors, or flawed taxonomic delimitation (Locatelli et al., 2020; Mutanen et al., 2016). While our study has followed current best practices in reference database curation, we notably did not go to the extra length of verifying the original source for the taxonomic identities applied to each sequence, a task which would prove insurmountable for a dataset of this size. While computational curation presents an extremely scalable approach for resolving annotation errors and contamination within public datasets, the quality of any identification ultimately depends on the quality of the systematics and taxonomy that originally delimited and described the species (Clarke &

Schutze, 2014). Therefore, additional taxonomic consideration using more comprehensive genomic (Leaché, Fujita, Minin, & Bouckaert, 2014) or integrative methods (Padial, Miralles, De la Riva, & Vences, 2010) may be required to determine if the problem clades identified in our study are actually due to insufficient resolution within the COI barcode, or rather, over-splitting in the underlying taxonomy (Mutanen et al., 2016).

The 11,431 defunct or synonymous species names identified and corrected in our study clearly demonstrates that taxonomic synonyms remain one of the largest and seldom discussed issues within public sequence repositories (Leray, Knowlton, Ho, Nguyen, & Machida, 2019). While taxonomic names must be free to change to reflect revised species concepts, this becomes a problem when historically defunct species names are retained in reference databases, propagating errors through later studies and the management decisions made from them (Clarke et al., 2019). For insect metabarcoding, issues arising from taxonomic synonyms will become most apparent as hierarchical taxonomic classifiers already widely adopted by microbiome researchers become more prevalent (Porter, Gibson, Shokralla, & Baird, 2014; Porter & Hajibabaei, 2018). These methods require a query sequence to reach a certain bootstrap support to be assigned to lower ranks in the taxonomy, but conflicts in taxonomic annotation between genetically similar reference sequences could result in a failure to reach the required confidence, and thus false negative detections. Therefore, we recommend that resolving taxonomic synonyms to their currently accepted name become a default step in all metabarcoding database curation efforts, alongside the more common practices of removing non-homologous sequences, contaminants, and misannotated taxonomy (Kozlov, Zhang, Yilmaz, Glöckner, & Stamatakis, 2016; Richardson, Sponsler, McMinn-Sauder, & Johnson, 2020). While some curation of taxonomic synonyms already occurs within both GenBank and BOLD (Schoch et al., 2020), determining the currently valid taxonomic name from a diverse and constantly evolving primary literature is by no means a trivial task (Schoch et al., 2017). Continued investment into digital infrastructure for distribution of taxonomic information will therefore prove critical for ensuring identifications obtained through metabarcoding remain robust to the inevitable future description and renaming of taxa (Miralles et al., 2020).

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

Despite this study being a purely in-silico evaluation and the many limitations that this entails (Alberdi et al., 2018; Corse et al., 2019; Zhang, Zhao, & Yao, 2020), it represents a comprehensive first step in applying big data principles to inform development of HTS based diagnostics for invasive insect pests. As well as providing a starting point for diagnosticians and researchers selecting mini-barcode primers for insect identification, our results offer a degree of confidence to managers and regulators grappling with the consequences of broad-scope HTS diagnostic assays and how to appropriately respond the incidental detection of regulated species (Darling, Pochon, Abbott, Inglis, & Zaiko, 2020). Whilst we still recommend additional laboratory validation be conducted on high-priority targets before adoption in active surveillance, promisingly, many of the outstanding primers highlighted in our rankings have also shown similarly high performance in a recent in-vitro evaluation on a diverse insect mock community (Elbrecht et al., 2019). Our finding that mini-barcodes of lengths 125-257 bp provide comparable resolution to the full-length barcode opens for use of production-scale sequencing platforms such as the Illumina NovaSeq to cost-effectively process large numbers of bulk samples. Nevertheless, the lower capacity Illumina MiSeq may remain more appropriate when samples are likely to arrive at the laboratory in small batches, due to the increased diagnostic turnaround time from waiting to fill the higher capacity NovaSeq flow cells. In summary, appropriately chosen COI mini-barcode primers perform effectively across the majority of pest and non-pest insects, opening for the adoption of universal metabarcoding assays within diagnostic laboratories. While the computational curation pipeline presented here can resolve many issues inherent to public reference sequence data, regardless of whether the diagnostic tool is a microscope or a HTS assay, the accuracy of results will ultimately depend on the underlying quality and completeness of the taxonomy for the target groups.

Acknowledgments

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

Availability of data and materials:

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

Author contributions

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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*, 10(11), 1985–2001. 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

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

Crooks, J. A. (2005). Lag times and exotic species: The ecology and management of biological invasions in slow-motion. *Ecoscience*, 12(3), 316–329. doi:10.2980/i1195-6860-12-3-316.1

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

Darling, J. A., Pochon, X., Abbott, C. L., Inglis, G. J., & Zaiko, A. (2020). The risks of using molecular biodiversity data for incidental detection of species of concern. *Diversity and Distributions*, 26(9), 1116–1121. doi:10.1111/ddi.13108

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biology Letters*, 10(9), 20140562. doi:10.1098/rsbl.2014.0562

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9), 755–763. 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*, 7, e7745. doi:10.7717/peerj.7745

Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS ONE*, 10(7), e0130324. 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

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

EPPO. (2019b). PM 7/98 (4) Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity. *EPPO Bulletin*, 49(3), 530–563. doi:10.1111/epp.12629

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, *125*(1), 1–15. doi:10.1086/284325

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., … Pompanon, F. (2010). An In silico approach for the evaluation of DNA barcodes. *BMC Genomics*, *11*, 434. doi:10.1186/1471-2164-11-434

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

Fong, Y. (2019). Fast Bootstrap Confidence Intervals for Continuous Threshold Linear Regression. *Journal of Computational and Graphical Statistics*, *28*(2), 466–470. doi:10.1080/10618600.2018.1537927

Fong, Y., Huang, Y., Gilbert, P. B., & Permar, S. R. (2017). chngpt: Threshold regression model estimation and inference. *BMC Bioinformatics*, *18*, 454. doi:10.1186/s12859-017-1863-x

Funk, D. J., & Omland, K. E. (2003). Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 397–423. doi:10.1146/annurev.ecolsys.34.011802.132421

Galan, M., Pons, J. B., Tournayre, O., Pierre, É., Leuchtmann, M., Pontier, D., & Charbonnel, N. (2018). Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. *Molecular Ecology Resources*, *18*(3), 474–489. doi:10.1111/1755-0998.12749

Garg, A., Leipe, D., & Uetz, P. (2019). The disconnect between DNA and species names: Lessons from reptile species in the NCBI taxonomy database. *Zootaxa*, *4706*(3), 401–407. doi:10.11646/zootaxa.4706.3.1

Geller, J., Meyer, C., Parker, M., & Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources*, *13*(5), 851–861. doi:10.1111/1755-0998.12138

Gibson, J. F., Shokralla, S., Curry, C., Baird, D. J., Monk, W. A., King, I., & Hajibabaei, M. (2015). Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLoS ONE*, *10*(10), e0138432. doi:10.1371/journal.pone.0138432

Gibson, J. F., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., … Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences*, *111*(22), 8007–12. doi:10.1073/pnas.1406468111

Hajibabaei, M., Janzen, D. H., Burns, J. M., Hallwachs, W., & Hebert, P. D. N. (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(4), 968–971. doi:10.1073/pnas.0510466103

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., & Baird, D. J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, 6(4), e17497. doi:10.1371/journal.pone.0017497

Hajibabaei, M., Smith, M. A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., & Hebert, P. D. N. (2006). A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, 6(4), 959–964. doi:10.1111/j.1471-8286.2006.01470.x

Hajibabaei, M., Spall, J. L., Shokralla, S., & van Konynenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology*, 12(1), 28. doi:10.1186/1472-6785-12-28

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

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences of the United States of America*, 101(41), 14812–14817. doi:10.1073/pnas.0406166101

Hernández-Triana, L. M., Prosser, S. W., Rodríguez-Perez, M. A., Chaverri, L. G., Hebert, P. D. N., & Ryan Gregory, T. (2014). Recovery of DNA barcodes from blackfly museum specimens (Diptera: Simuliidae) using primer sets that target a variety of sequence lengths. *Molecular Ecology Resources*, 14(3), 508–518. doi:10.1111/1755-0998.12208

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

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

Jordaens, K., Sonet, G., Richet, R., Dupont, E., Braet, Y., & Desmyter, S. (2013). Identification of forensically important Sarcophaga species (Diptera: Sarcophagidae) using the mitochondrial COI gene. *International Journal of Legal Medicine*, 127(2), 491–504. doi:10.1007/s00414-012-0767-6

Kozlov, A. M., Zhang, J., Yilmaz, P., Glöckner, F. O., & Stamatakis, A. (2016). Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research*, 44(11), 5022–5033. doi:10.1093/nar/gkw396

Krehenwinkel, H., Fong, M., Kennedy, S., Huang, E. G., Noriyuki, S., Cayetano, L., & Gillespie, R. (2018). The effect of DNA degradation bias in passive sampling devices on metabarcoding studies of arthropod communities and their associated microbiota. PLoS ONE, 13(1), e0189188. https://doi.org/10.1371/journal.pone.0189188

Kwok, S., Chang, S. Y., Sninsky, J. J., & Wang, A. (1994). A guide to the design and use of mismatched and degenerate primers. *Genome Research*, 3, S39–S47. doi:10.1101/gr.3.4.S39

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., … Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, *31*(9), 814–821. doi:10.1038/nbt.2676

Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species delimitation using genome-wide SNP Data. *Systematic Biology*, 63(4), 534–542. doi:10.1093/sysbio/syu018

Leese, F., Sander, M., Buchner, D., Elbrecht, V., Haase, P., & Zizka, V. M. A. (2021). Improved freshwater macroinvertebrate detection from environmental DNA through minimized nontarget amplification. *Environmental DNA*, 3(1), 261–276. doi:10.1002/edn3.177

Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences of the United States of America*, 116(45), 22651–22656. doi:10.1073/pnas.1911714116

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, 10(1), 34. doi:10.1186/1742-9994-10-34

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

Locatelli, N. S., McIntyre, P. B., Therkildsen, N. O., & Baetscher, D. S. (2020). GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. *Proceedings of the National Academy of Sciences*, 117(51), 32211–32212. doi:10.1073/pnas.2007421117

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, 1082. doi:10.3389/fpls.2018.01082

Marquina, D., Andersson, A. F., & Ronquist, F. (2019). New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. *Molecular Ecology Resources*, 19(1), 90–104. doi:10.1111/1755-0998.12942

Martiny, A. C., Treseder, K., & Pusch, G. (2013). Phylogenetic conservatism of functional traits in microorganisms. *ISME Journal*, 7(4), 830–838. doi:10.1038/ismej.2012.160

Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. L. (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, 55(5), 715–728. doi:10.1080/10635150600969864

Meusnier, I., Singer, G. A. C., Landry, J. F., Hickey, D. A., Hebert, P. D. N., & Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, 9, 214. doi:10.1186/1471-2164-9-214

Meyer, C. P. (2003). Molecular systematics of cowries (Gastropoda: Cypraeidae) and diversification patterns in the tropics. *Biological Journal of the Linnean Society*, 79(3), 401–459. doi:10.1046/j.1095-8312.2003.00197.x

Meyerson, L. A., & Reaser, J. K. (2002). Biosecurity: Moving toward a comprehensive approach. *BioScience*, 52(7), 593–600. doi:10.1641/0006-3568(2002)052[0593:BMTACA]2.0.CO;2

Miralles, A., Bruy, T., Wolcott, K., Scherz, M. D., Begerow, D., Beszteri, B., … Vences, M. (2020). Repositories for taxonomic data: Where we are and what is missing. *Systematic Biology*, 69(6), 1231–1253. doi:10.1093/sysbio/syaa026

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., … Zhou, X. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210), 763–767. doi:10.1126/science.1257570

Mutanen, M., Kivelä, S. M., Vos, R. A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., … Godfray, H. C. J. (2016). Species-level para- and polyphyly in DNA barcode gene trees: Strong operational bias in European Lepidoptera. *Systematic Biology*, 65(6), 1024–1040. doi:10.1093/sysbio/syw044

Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, 7, 16. doi:10.1186/1742-9994-7-16

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

Park, D. S., Foottit, R., Maw, E., & Hebert, P. D. N. (2011). Barcoding bugs: DNA-based identification of the true bugs (insecta: Hemiptera: Heteroptera). *PLoS ONE*, 6(4), e18749. doi:10.1371/journal.pone.0018749

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., Ratnasingham, S., Miller, S. E., & Hebert, P. D. N. (2020). BOLD and GenBank revisited - Do identification errors arise in the lab or in the sequence libraries? *PLoS ONE*, 15(4), e0231814. doi:10.1371/journal.pone.0231814

Pentinsaari, M., Salmela, H., Mutanen, M., & Roslin, T. (2016). Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. *Scientific Reports*, 6, 35275. doi:10.1038/srep35275

Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *GigaScience*, 8(8), 1–22. doi:10.1093/gigascience/giz092

Porter, T. M., Gibson, J. F., Shokralla, S., & Baird, D. J. (2014). Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome c oxidase subunit 1 (COI) DNA barcode sequences using a naive Bayesian classifier. *Molecular Ecology Resources*, 14, 929–942. doi:10.5061/dryad.bc8pc.

Porter, T. M., & Hajibabaei, M. (2018). Automated high throughput animal CO1 metabarcode classification. *Scientific Reports*, 8, 4226. doi:10.1038/s41598-018-22505-4

Porter, T. M., & Hajibabaei, M. (2020). Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis. *Frontiers in Ecology and Evolution*, 8, 248. doi:10.3389/fevo.2020.00248

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26(7), 1641–1650. doi:10.1093/molbev/msp077

Prosser, S. W. J., Dewaard, J. R., Miller, S. E., & Hebert, P. D. N. (2016). DNA barcodes from century-old type specimens using next-generation sequencing. *Molecular Ecology Resources*, 16(2), 487–497. doi:10.1111/1755-0998.12474

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD : The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, 7(3), 355–364. doi:10.1111/j.1471-8286.2006.01678.x

Reaser, J. K., Burgiel, S. W., Kirkey, J., Brantley, K. A., Veatch, S. D., & Burgos-Rodríguez, J. (2020). The early detection of and rapid response (EDRR) to invasive species: a conceptual framework and federal capacities assessment. *Biological Invasions*, 22(1), 1–19. doi:10.1007/s10530-019-02156-w

Reaser, J. K., Frey, M., & Meyers, N. M. (2020). Invasive species watch lists: guidance for development, communication, and application. *Biological Invasions*, 22(1), 47–51. doi:10.1007/s10530-019-02176-6

Reaser, J. K., Meyerson, L. A., & von Holle, B. (2008). Saving camels from straws: How propagule pressure-based prevention policies can reduce the risk of biological invasion. *Biological Invasions*, 10(7), 1085–1098. doi:10.1007/s10530-007-9186-x

Rennstam Rubbmark, O., Sint, D., Horngacher, N., & Traugott, M. (2018). A broadly applicable COI primer pair and an efficient single-tube amplicon library preparation protocol for metabarcoding. *Ecology and Evolution*, 8(24), 12335–12350. doi:10.1002/ece3.4520

Richardson, R. T., Sponsler, D. B., McMinn-Sauder, H., & Johnson, R. M. (2020). MetaCurator: A hidden Markov model-based toolkit for extracting and curating sequences from taxonomically-informative genetic markers. *Methods in Ecology and Evolution*, 11(1), 181–186. doi:10.1111/2041-210X.13314

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

Ross, H. A. (2014). The incidence of species-level paraphyly in animals: A re-assessment. *Molecular Phylogenetics and Evolution*, 76(1), 10–17. doi:10.1016/j.ympev.2014.02.021

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

Scheider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20), 6097–6100.

Schoch, C. L., Aime, M. C., de Beer, W., Crous, P. W., Hyde, K. D., Penev, L., … Miller, A. N. (2017). Using standard keywords in publications to facilitate updates of new fungal taxonomic names. *IMA Fungus*, 8(2), 70–73. doi:10.1007/BF03449466

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., … Karsch-Mizrachi, I. (2020). NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database*, baaa062. doi:10.1093/database/baaa062

Schrader, G., & Unger, J. G. (2003). Plant quarantine as a measure against invasive alien species: The framework of the International Plant Protection Convention and the plant health regulations in the European Union. *Biological Invasions*, 5(4), 357–364. doi:10.1023/B:BINV.0000005567.58234.b9

Shen, Z., Qu, W., Wang, W., Lu, Y., Wu, Y., Li, Z., … Zhang, C. (2010). MPprimer: A program for reliable multiplex PCR primer design. *BMC Bioinformatics*, 11, 143. doi:10.1186/1471-2105-11-143

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., … Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, 5, 9687. doi:10.1038/srep09687

Siddall, M. E., Fontanella, F. M., Watson, S. C., Kvist, S., & Erséus, C. (2009). Barcoding bamboozled by bacteria: Convergence to metazoan mitochondrial primer targets by marine microbes. *Systematic Biology*, 58(4), 445–451. doi:10.1093/sysbio/syp033

Simberloff, D. (2006). Risk assessments, blacklists, and white lists for introduced species: Are predictions good enough to be useful? *Agricultural and Resource Economics Review*, 35(1), 1–10. doi:10.1017/S1068280500010005

Stadhouders, R., Pas, S. D., Anber, J., Voermans, J., Mes, T. H. M., & Schutten, M. (2010). The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5′ nuclease assay. *Journal of Molecular Diagnostics*, 12(1), 109–117. doi:10.2353/jmoldx.2010.090035

Stork, N. E. (2018). How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annual Review of Entomology*, 63, 31–45. doi:10.1146/annurev-ento-020117-043348

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8), 2045–2050. doi:10.1111/j.1365-294X.2012.05470.x

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, 17(2), 57–86.

Tedersoo, L., Drenkhan, R., Anslan, S., Morales-Rodriguez, C., & Cleary, M. (2019). High-throughput identification and diagnostics of pathogens and pests: overview and practical recommendations. *Molecular Ecology Resources*, 19, 47–76. doi:10.1111/1755-0998.12959

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*. doi:10.1093/nar/gks596

Vamos, E. E., Elbrecht, V., & Leese, F. (2017). Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcoding and Metagenomics*, 1, e14625. doi:10.3897/mbmg.1.14625

Wangensteen, O. S., Palacín, C., Guardiola, M., & Turon, X. (2018). DNA metabarcoding of littoral hardbottom communities: High diversity and database gaps revealed by two molecular markers. *PeerJ*, 6, e4705. doi:10.7717/peerj.4705

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wiemers, M., & Fiedler, K. (2007). Does the DNA barcoding gap exist? - A case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology*, 4, 8. doi:10.1186/1742-9994-4-8

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*, 35(19), 3829–3830. doi:10.1093/bioinformatics/btz159

Winter, David, J., & Winter, D. J. (2017). rentrez: An R package for the NCBI eUtils API. *The R Journal*, 9(2), 520–526. 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. doi:10.32614/RJ-2016-025

Yeo, D., Srivathsan, A., & Meier, R. (2020). Longer is Not Always Better: Optimizing Barcode Length for Large-Scale Species Discovery and Identification. *Systematic Biology*, 69(5), 999–1015. doi:10.1093/sysbio/syaa014

Young, R. G., Milián-García, Y., Yu, J., Bullas-Appleton, E., & Hanner, R. H. (2021). Biosurveillance for invasive insect pest species using an environmental DNA metabarcoding approach and a high salt trap collection fluid. *Ecology and Evolution*, 11(4), 1558–1569. doi:10.1002/ece3.7113

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup : metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3, 613–623. doi:10.1111/j.2041-210X.2012.00198.x

Yu, G., Lam, T. T. Y., Zhu, H., & Guan, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular Biology and Evolution*, 35(12), 3041–3043. 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

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

Zeale, M. R. K., Butlin, R. K., Barker, G. L. A., Lees, D. C., & Jones, G. (2011). Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. *Molecular Ecology Resources*, 11(2), 236–244. doi:10.1111/j.1755-0998.2010.02920.x

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

3.5 Supplementary Information

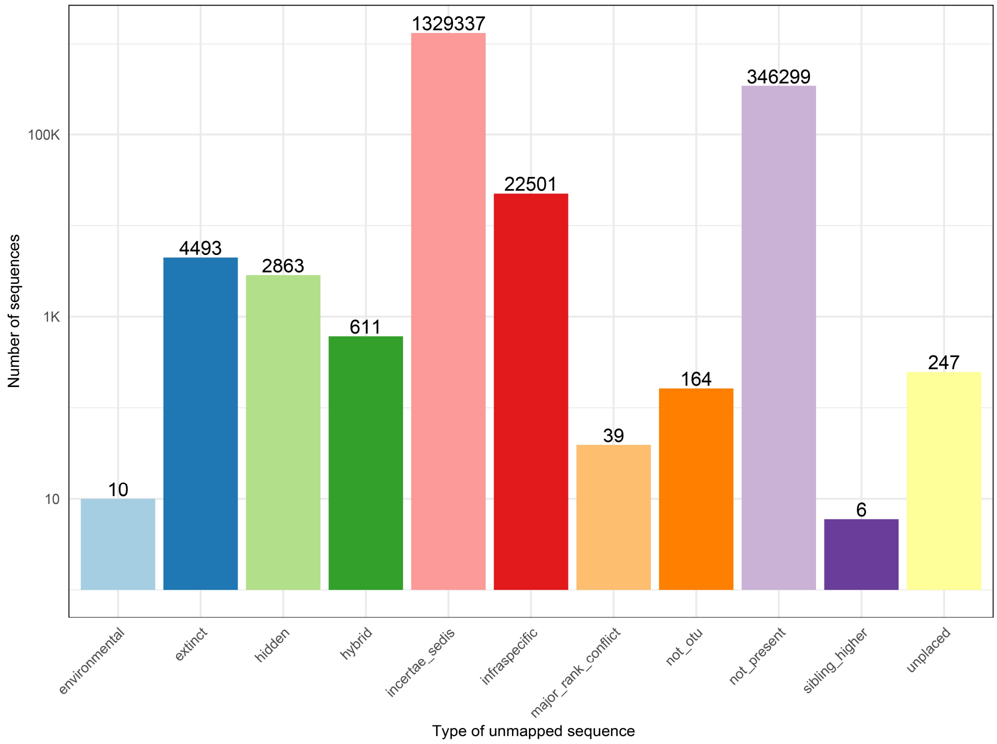
Supplementary Table 1: Published and novel primers evaluated in this study

Primer		Sequence	Citation
SternoCOIF1	F	ATTGGWGGWTTYGGAAAYTG	Batovska et al. (2021)
SternoCOIR1	R	ATRAARTTRATWGCTCCTA	Batovska et al. (2021)
Saurons878	F	GGDRCWGGWTGAACWGTWTAYCCNCC	Rennstam Rubbmark et al. (2018)
AgPestF1	F	ATYATWATTGGDGGDTTYGG	This Study
AgPestF2	F	HGAYATRGCHTTYCCHCG	This Study
HexCOIF4	F	HCCHGAYATRGCHTTYCC	Marquina et al. (2019)
HexCOIR4	R	TATDGTRATDGCHCCNGC	Marquina et al. (2019)
mLepR1	R	CCTGTTCCAGCTCCATTTT	Hebert et al. (2004)
AgPestR1a	R	GTRATRAARTTDAYWGMHCC	This Study
AgPestR1b	R	ARAATWGADGADAYWCCWGC	This Study
AgPestR2	R	RACWGMTCAVAYAAATARDGG	This Study
LCO1490	F	GGTCAACAAATCATAAAGATATTGG	Folmer et al. (1994)
HCO2198	R	TAAACTTCAGGGTGACCAAAAAATCA	Folmer et al. (1994)
Uni-MinibarR1	R	GAAAATCATAATGAAGGCATGAGC	Meusnier et al. (2008)
Uni-MinibarR1d	R	AAAATTATAATAAARGCRTGRGC	Jordaens et al. (2013)
Uni-MinibarF1	F	TCCACTAATCACAARGATATTGGTAC	Meusnier et al. (2008)
UniMinibarF1d	F	TCCACTAATCACAARGATATTGGTAC	Jordaens et al. (2013)
ZBJ-ArtF1c	F	AGATATTGGAACWTTATATTTTATTTTTGG	Zeale et al. (2011)
ZBJ-ArtF1c-deg	F	RGAYATYGGWACHYTWTAYTTYHTHTTYGG	Elbrecht et al. (2019)
ZBJ-ArtR2c	R	WACTAATCAATTWCCAAATCCTCC	Zeale et al. (2011)
ZBJ-ArtR2c-deg	R	WAYTARTCARTTWCCRAAHCCHCC	Elbrecht et al. (2019)
mlCOIintF	F	GGWACWGGWTGAACWGTWTAYCCYCC	Leray et al. (2013)
mlCOIintR	R	GGRGGRTASACSGTTCASCCSGTSCC	Leray et al. (2013)
BR3	R	GGDGGRTANACWGTYCAHCCDGTHCC	Elbrecht et al. (2019)
LepF1	F	ATTCAACCAATCATAAAGATATTGG	Hebert et al. (2004)
EPT-long-univR	R	AARAAAATYATAAYAAAIGCGTGIAIIGT	Hajibabaei et al. (2011)
MLepF1-Rev	R	CGTGGAAAWGCTATATCWGGTG	Brandon-Mong et al. (2015)
Ill-C-R	R	GGIGGRTAIACIGTTCAICC	Shokralla et al. (2015)
Ill-B-F	F	CCIGAYATRGCITTYCCICG	Shokralla et al. (2015)
BF1	F	ACWGGWTGRACWGTNTAYCC	Elbrecht & Leese (2017b)
BF1i	F	ACIGGITGRACIGTITAYCC	Elbrecht et al. (2019)
BF2	F	GCHCCHGAYATRGCHTTYCC	Elbrecht & Leese (2017b)
BF3	F	CCHGAYATRGCHTTYCCHCG	Elbrecht et al. (2019)
BR1	R	ARYATDGTRATDGCHCCDGC	Elbrecht & Leese (2017b)
BR1i	R	ARYATIGTRATIGCICCIGC	Elbrecht et al. (2019)
BR2	R	TCDGGRTGNCCRAARAAYCA	Elbrecht & Leese (2017b)
ArF5	F	GCICCIGAYATRKCITTYCCICG	Gibson et al. (2014)
ArR5	R	GTRATIGCICCIGCIARIACIGG	Gibson et al. (2014)
jgLCO1490	F	TITCIACIAAYCAYAARGAYATTGG	Geller et al. (2013)
jgHCO2198	R	TAIACYTCIGGRTGICCRAARAAYCA	Geller et al. (2013)
MZplankF2	F	RGYNGGNACRGGNTGRACNGT	Elbrecht et al. (2019)
LepR1	R	TAAACTTCTGGATGTCCAAAAAATCA	Hebert et al. (2004)
C-LepFolR	R	TAAACTTCWGGRTGWCCAAAAAATCA	Hernández-Triana et al. (2014)

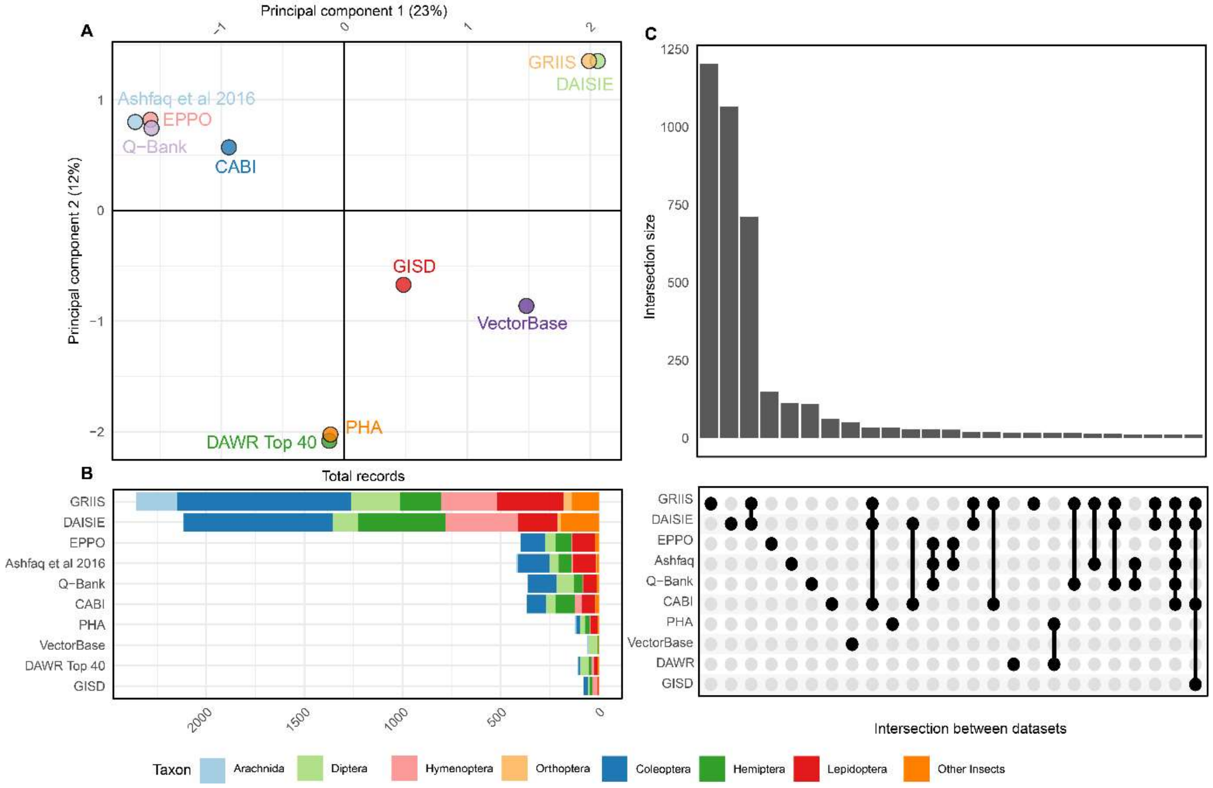
AncientLepF3	F	TTATAATTGGDGGWTTTGGWAATTG	Prosser et al. (2016)
A	F	GGIGGITTTGGIAATTGAYTIGTICC	Hajibabaei et al. (2012)
D	R	CCTARIATIGAIGARAYICCIGC	Hajibabaei et al. (2012)
B	F	CCIGAYATRGCITTYCCICG	Hajibabaei et al. (2012)
Bn	F	CCNGAYATRGCNTTYCCNCG	Elbrecht et al. (2019)
E	R	GTRATIGCICCIGCIARIAC	Hajibabaei et al. (2012)
En	R	GTRATNGCNCCNGCNARNAC	(Elbrecht et al. (2019)
C	F	GITGAACIGTITAYCCICC	Hajibabaei et al. (2012)
F	R	CCIGCIGGRTCIAARAAIGAIGT	Hajibabaei et al. (2012)
fwhF1	F	YTCHACWAAYCAYAARGAYATYGG	Vamos et al. (2017)
fwhR1	R	ARTCARTTWCCRAAHCCHCC	Vamos et al. (2017)
fwhF2	F	GGDACWGGWTGAACWGTWTAYCCHCC	Vamos et al. (2017)
fwhR2n	R	GTRATWGCHCCDGCTARWACWGG	Vamos et al. (2017)
MG-LCO1490	F	ATTCHACDAAYCAYAARGAYATYGG	Galan et al. (2018)
MG-univR	R	ACTATAAARAARATYATDAYRAADGCRTG	Galan et al. (2018)
230-R	R	CTTATRTTRTTTATICGIGGRAAIGC	Gibson et al. (2015)
MhemF	F	GCATTYCCACGAATAAATAAYATAAG	Park et al. (2011)
dgHCO2198	R	TAAACTTCAGGGTGACCAAARAAYCA	Meyer (2003)
dgLCO1490	F	GGTCAACAAATCATAAAGAYATYGG	Meyer (2003)
Fol-degen-for	F	TCNACNAAYCAYAARRAYATYGG	D. W. Yu et al. (2012)
Fol-degen-rev	R	TANACYTCNGGRTGNCCRAARAAYCA	D. W. Yu et al. (2012)
MLepF1	F	GCTTTCCCACGAATAAATAATA	Hajibabaei, Janzen et al. (2006)
RonMWASPdeg	F	GGWTCWCCWGATATAKCWTTTCC	Clare et al. (2019)
mlCOIintF-XT	F	GGWACWRGWTGRACWITITAYCCYCC	Wangensteen et al. (2018)
EPTDr2n	R	CAAACAAATARDGGTATTCGDTY	Leese et al. (2021)

Supplementary Table 2: Criteria used to rank primer characteristics

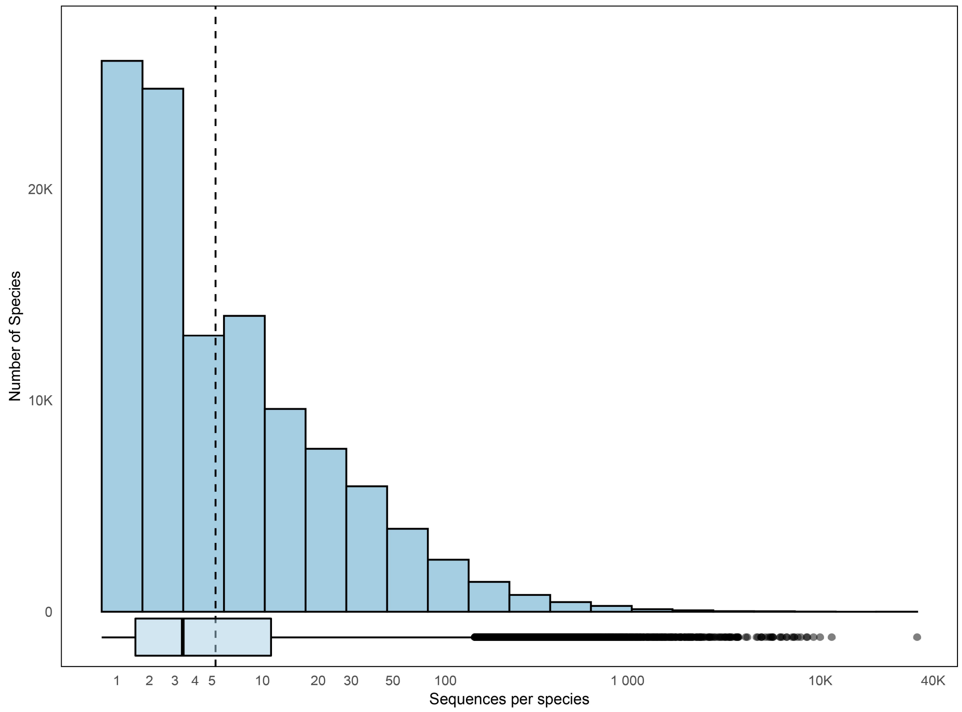
Criterion	Good (1)	Moderate (0)	Bad (-1)
GC%	40%-60%	30%-40% or 60%-70%	<30% or >70%
Degeneracy	0-517 fold	517-1026 fold	>1026 fold
GC clamp (last 2 bases of 5' end)	2 G or C bases	1 G or C base	No G or C bases
Primer length	18-24 bp	16-18 bp or 24-26 bp	<16 bp or >25 bp
Longest homopolymer	≤2 bp	≤4 bp	>4 bp
Melting temperature	48-62 °C	43-48 °C or 63-67 °C	<43 °C or >67 °C



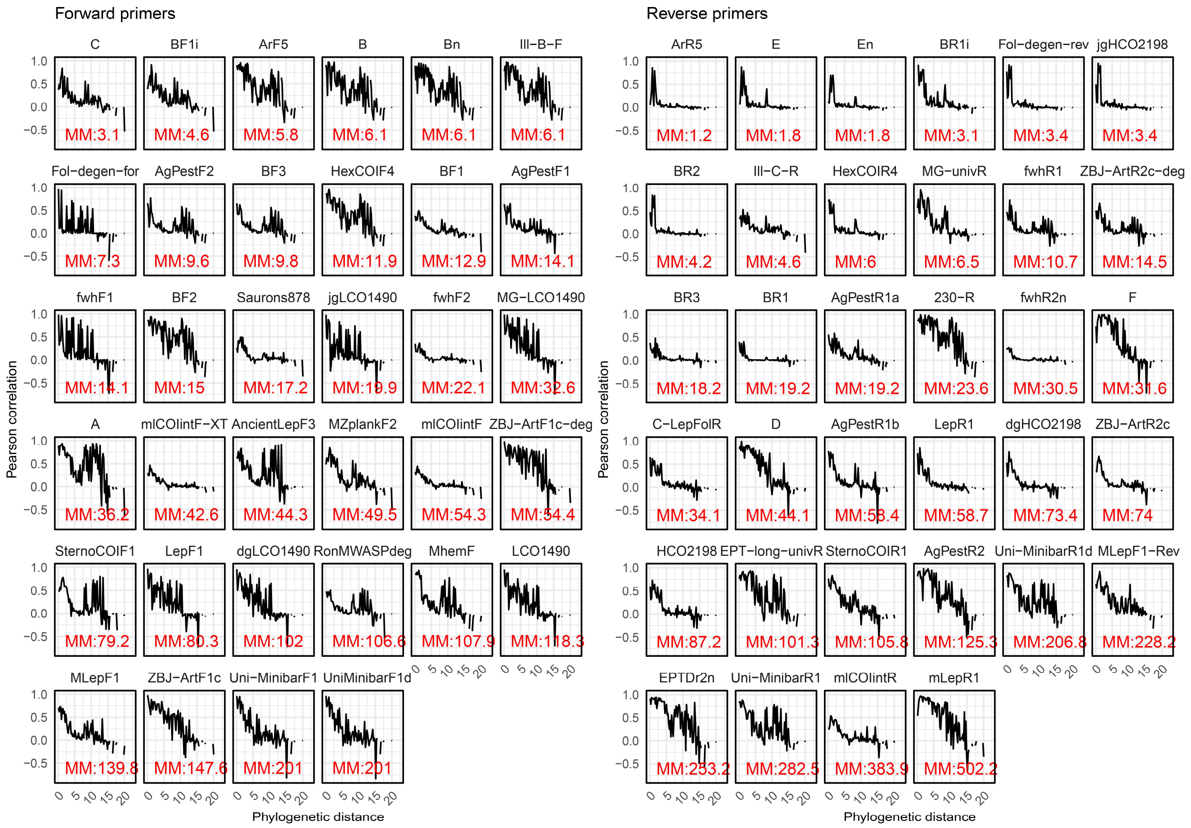
Supplementary Figure 1: Categories of sequences that could not be successfully mapped into the Open Tree of Life taxonomy. Displayed on a log10 scale.



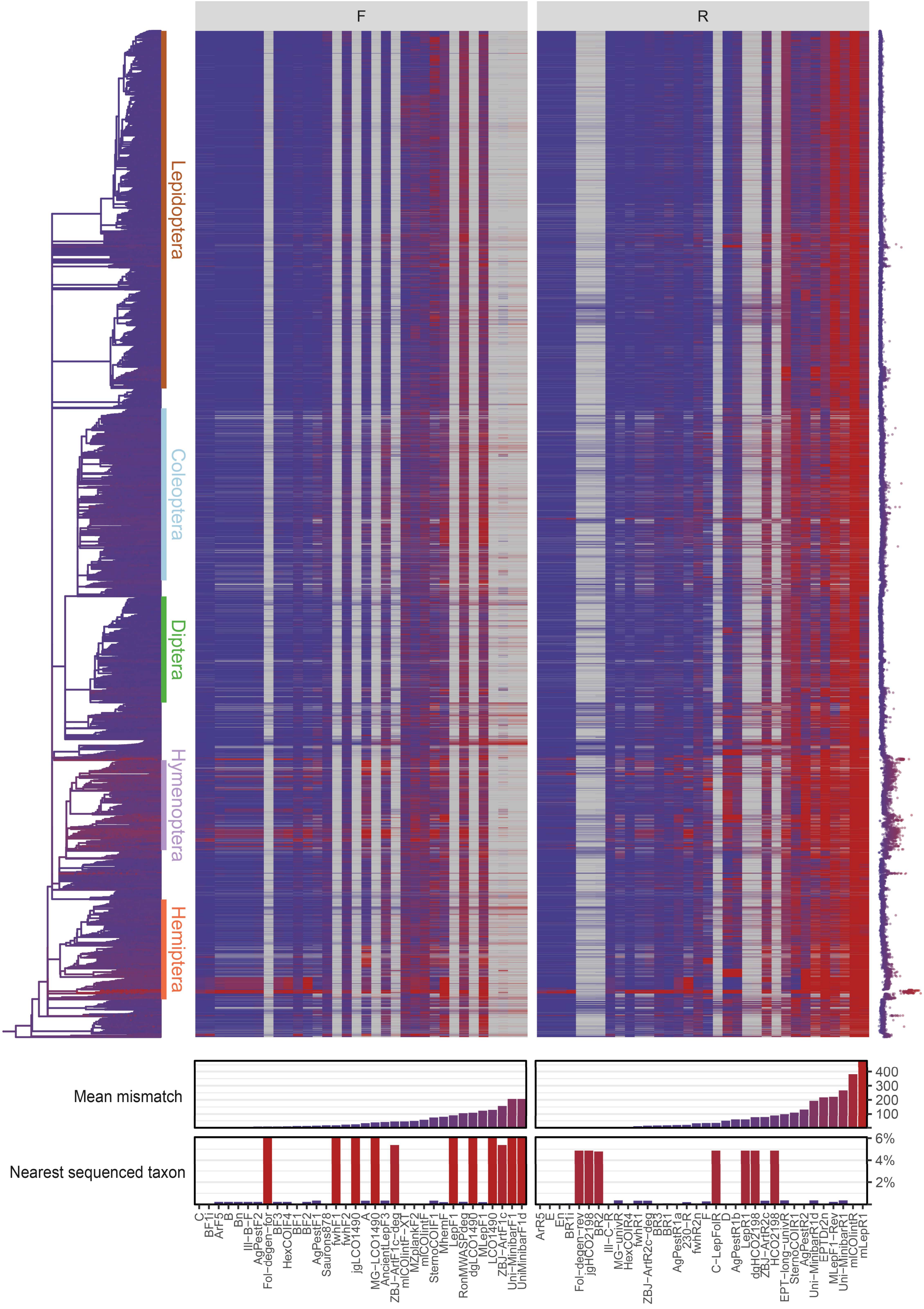
Supplementary Figure 2 – Summary of public pest and invasive insect datasets used to assemble the pest list for primer evaluation. **A)** Principal component analysis of species overlap (Jaccard distance) between pest lists. **B)** Total records for each dataset source, coloured by taxonomic order or class. **C)** 25 largest intersections of species names between invasive or pest species datasets.



Supplementary Figure 3: Number of sequences per species in the reference database before pruning. Vertical line indicates the maximum of 5 sequences per species the final database was pruned to. Displayed on a pseudo-log scale.



Supplementary Figure 4: Phylogenetic autocorrelation function of primer mismatch for each forward and reverse primer. Annotations refer to mean mismatch between primer and all insect species.



Supplementary Figure 5: Alternative to main figure 5 without imputation of missing data

Supplementary Note 1: Sources for invasive or pest species records used to assemble global list of pest Arthropods.

- 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: All sequences that were mapped to nodes within the Open Tree of Life that were annotated with these flags that indicate uncertain placement were removed during sequence filtering.

- incertae_sedis
- major_rank_conflict
- infraspecific
- unplaced
- environmental
- inconsistent
- extinct
- hidden
- hybrid
- not_otu
- viral
- barren

Supplementary Note 3: All sequences with taxonomic annotations containing these words that indicate insufficient identification were 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)

4

DNA Metabarcoding Enables High-Throughput Detection of Spotted Wing Drosophila (*Drosophila suzukii*) Within Unsorted Trap Catches

4.1 Chapter preface:

This chapter extends upon the in-silico results of Chapter 3 by comparing four of the best performing primers for their ability to detect Spotted Wing Drosophila (*Drosophila suzukii*), a high priority pest for Australia, within unsorted trap samples. As part of this laboratory validation, the sensitivity, specificity, and overall accuracy of the assay are established for both the primary target and its close relatives, and the number of biological and technical replicates required for reliable detection is determined. This chapter employs a non-destructive DNA extraction method to retain intact specimens for confirmation of any detected exotic species, and evaluates various methods for deriving a detection threshold to resolve false positives introduced through index-switching, both issues discussed in depth within Chapter 2. The laboratory protocol and bioinformatic pipeline developed here are used again in Chapter 5, where the quantitative performance of the assay is refined. This chapter is presented as a self-contained manuscript in the final stages of preparation, with intended submission to the journal *Environmental DNA*, and includes supplementary material at the end.

4.2 Publication details:

DNA Metabarcoding Enables High-Throughput detection of Spotted Wing Drosophila (Drosophila suzukii) within Unsorted Trap Catches

Stage of publication: In Preparation

Journal details: Environmental DNA

Authors: Alexander M. Piper, John Paul Cunningham, Noel O.I. Cogan, Mark J. Blacket

4.3 Statement of joint authorship:

A.M.P, J.P.C. and M.J.B. conceptualised the study, A.M.P designed and performed all field sampling, laboratory procedures, bioinformatic and statistical analyses. A.M.P. wrote the first draft of the manuscript with input and supervision from J.P.C, N.O.I.C., and M.J.B. All authors contributed to the editing of the final manuscript and approved the version presented here.

Statement from co-author confirming the contribution of the PhD candidate:

"As co-author of the manuscript 'Piper, A. M., Cunningham, J. P, Cogan N.O.I & Blacket M.J. (In preparation). DNA Metabarcoding Enables High-Throughput detection of Spotted Wing Drosophila (Drosophila suzukii) within Unsorted Trap Catches, *Environmental* DNA, I confirm that Alexander M. Piper has made the contributions listed above."

Associate Professor John Paul Cunningham

30/03/2021

4.4 Manuscript

DNA Metabarcoding Enables High-Throughput detection of Spotted Wing Drosophila (*Drosophila suzukii*) within Unsorted Trap Catches

Alexander M. Piper[1,2], John Paul Cunningham[1,2], Noel O.I. Cogan[1,2], Mark J. Blacket[1]

[1] Agriculture Victoria Research, AgriBio Centre, 5 Ring Road Bundoora 3083, Victoria, Australia

[2] School of Applied Systems Biology, La Trobe University, Bundoora 3083, Victoria, Australia

Keywords: Biosurveillance, invasive species, alien species, early detection, amplicon sequencing, COI, DNA barcoding

Running title: Spotted Wing Drosophila metabarcoding

Corresponding author:

Alexander M. Piper

Email: alexander.piper@agriculture.vic.gov.au

Abstract

The spotted wing drosophila (*Drosophila suzukii*, Matsumara) is a rapidly spreading global pest of soft and stone fruit production. Due to lack of selectivity of monitoring traps and the similarity of many of its life stages to other cosmopolitan drosophilids, surveillance for this pest is currently bottlenecked by the required sorting and identification of mixed trap catches. DNA metabarcoding is an untargeted, high-throughput sequencing based assay that allows multi-species identification of mixed communities, and thus may lend itself ideally to rapid and scalable diagnostics of *D. suzukii* within unsorted insect trap samples. In this study we compare the qualitative (identification accuracy) and quantitative (bias towards each species) performance of four recently published metabarcoding primer sets on *D. suzukii* and its close relatives. We then determine the sensitivity of a non-destructive metabarcoding assay (i.e. which retains intact specimens) by spiking target specimens into mock communities of increasing size, as well as diverse field-sampled communities from a cherry and a stone fruit orchard. Metabarcoding successfully detected *D. suzukii* and its close relatives *D. subpulchrella* and *D. biarmipes* in a background of Australian drosophila with a sensitivity of 73.6%, 76% and 81% respectively, and further identified 42 non-target arthropods collected as bycatch by *Drosophila* surveillance methods. Trap designs and surveillance protocols will, however, need to be optimised to adequately preserve specimen DNA for molecular identification. While the non-destructive DNA extraction retained intact voucher specimens, dropouts of low-abundance taxa and entire replicates suggest that these protocols behave more similarly to environmental DNA than tissue homogenisation-based metabarcoding, and thus will require increased replication to ensure reliable detections. Adoption of high-throughput metabarcoding assays for screening mixed trap samples could enable a substantial increase in the geographic scale and intensity of *D. suzukii* surveillance, and thus the likelihood of detecting a new incursion.

Introduction

The combined influences of international trade, tourism, and changing climates are increasing the rate at which new insect pests emerge and spread, creating a global burden on food security (Savary et al., 2019). A particularly striking example is the rapid intercontinental spread of *Drosophila suzukii*, Matsumara (spotted wing drosophila), a significant pest of soft and stone fruits which over the last two decades has expanded from its native range in South East Asia (Kanzawa, 1939; Walsh et al., 2011), to Europe, the Americas, and more recently Africa (Asplen et al., 2015; Cini et al., 2012; Goodhue et al., 2011; Kwadha et al., 2021). The pace of this expansion is attributed to a high fecundity, short generation time, and a broad host range that allows populations to persist throughout the year by alternating between cultivated and wild fruits with different ripening times (Cini et al., 2012). Recent modelling of global climatic suitability predicts further establishment of *D. suzukii* if introduced into regions and continents where it is not yet present, such as Australia and New Zealand (Dos Santos et al., 2017; Maino et al., 2020).

Early detection is critical for containment and eradication of invasive insect populations, with the probability of detecting a new incursion increasing with the intensity of surveillance (Anderson et al., 2017; Liebhold et al., 2016). Surveillance for *D. suzukii* is generally conducted using traps baited with 'food attractant' lures such as apple cider vinegar (Landolt et al., 2012), live yeasts (Hamby et al., 2014), or synthetic formulations mimicking these (Cha et al., 2012, 2014). To complement trapping, infested fruit can be crushed and agitated in a salt solution to float any larvae and eggs to the surface, which can then be collected via filtration (Van Timmeren et al., 2017). However, neither of these surveillance techniques are specifically selective for *D. suzukii*, often capturing hundreds of mixed specimens that must be sorted and identified in order to detect a new incursion (Burrack et al., 2015; Tonina et al., 2018). In addition to the sheer numbers of specimens, rapid morphological identification of *D. suzukii* is hampered by the characteristic "spotted wings" being present only for male flies, unreliable for juvenile adults, and shared by its sister species *D. biarmipes* and *D. subpulchrella* (Cini et al., 2012; Hauser, 2011). Alternative molecular diagnostic assays such as DNA barcoding (Calabria et al., 2012), real-time PCR (Dhami & Kumarasinghe, 2014), PCR-RFLP (S. S. Kim et al., 2014), and loop-mediated isothermal amplification (LAMP) (Y. H. Kim et al., 2016) can provide highly accurate identifications of any life stage, yet the costly and time-consuming process of conducting single reactions on individual specimens has restricted their use to confirming the identity of specimens already suspected to be *D. suzukii* (Boughdad et al., 2021; Calabria et al., 2012). Lack of a cost-effective and high-throughput diagnostic method for bulk trap catches remains a major bottleneck for large-scale *D. suzukii* surveillance, with misidentification or delayed management response incurring considerable costs to individual growers and the wider economy (Hauser, 2011).

DNA metabarcoding is an untargeted molecular assay that couples DNA barcoding with high-throughput sequencing (HTS) in order to simultaneously identify all species within complex mixed communities (Taberlet et al., 2012; Tedersoo et al., 2019). The resulting whole-community data can be compared to both lists of regulated species and baseline knowledge of endemic biodiversity to screen not just for target pests, but also other unanticipated taxa that are not being actively searched for (Batovska et al., 2020; Hardulak et al., 2020). The ability for metabarcoding to be conducted on mixed trap samples without any prior sorting (Nielsen et al., 2019) is particularly appealing for efficiently handling the large number of specimens likely to be produced by an intensive surveillance programme for *D. suzukii*. Nevertheless, ensuring the accuracy of detections must be a priority for use of metabarcoding in an invasive species surveillance context (Piper et al., 2019), due to the risk of false positive and negative detections being introduced by phenomena such as index switching (Schnell et al., 2015a), PCR biases (Deagle et al., 2014), and stochastic sampling of molecules from low abundance specimens (Leray & Knowlton, 2017). Robust metabarcoding assays therefore require both technical replication and use of a detection threshold to resolve true positives from low-abundance contaminants (Zinger et al., 2019), but the number and type of replicates, and appropriate manner for deriving this detection threshold remains unclear for assays which employ non-destructive DNA extractions (Batovska et al., 2020; Carew et al., 2018; Nielsen et al., 2019). These recently developed non-destructive protocols allow high-throughput metabarcoding detections to be confirmed using gold-standard morphological examination and voucher specimens to be retained according to regulatory requirements (Batovska et al., 2020; Martins et al., 2019), yet come at the expense of reduced DNA concentrations compared to more common tissue-homogenisation based protocols (Martoni et al., 2021).

In this study we evaluate the use of a non-destructive DNA metabarcoding assay for detection of *D. suzukii* and its close relatives *D. subpulchrella* and *D. biarmipes* within large unsorted trap samples. Four published primer sets are evaluated for their qualitative and quantitative performance and 6 methods for deriving a detection threshold compared for their ability to resolve false positives caused by index-switching. The sensitivity, specificity, and overall accuracy of the assay, as well as the required number of PCR and DNA extraction replicates is then determined via spiking target species into both mock communities of known composition and field samples collected from a cherry and stone fruit orchard. Analysis of these diverse field samples enabled further assessment of the selectivity of different *D. suzukii* sampling strategies, as well as the effects of commonly used attractant lures on DNA preservation of trapped specimens. Practical implementation of metabarcoding assays into *D. suzukii* surveillance and the wider implications of broad-scope HTS assays for plant pest diagnostics are discussed.

Methods

Assembling mock communities

To assemble mock communities for validating the metabarcoding assay, isofemale lines (David et al., 2005) of *D. melanogaster*, *D. simulans*, *D. hydei* and *Scaptodrosophila lattivitata* were established from individual female drosophila trapped in banana baited traps (Reed, 1938) around Melbourne, Australia. F1 offspring from each isofemale line were identified via DNA barcoding using the LCO1490-HCO2198 primers (Folmer et al., 1994) and those found to be of the same species combined into ongoing colonies. *D. melanogaster*, *D. simulans* and *D. hydei* colonies were maintained at 25 °C on a diet of instant drosophila medium (Carolina Biological Supply, USA) and live brewer’s yeast (Fleischmann's, USA), while *S. lattivitata* was maintained at 25 °C on the diet described by Bock & Parsons (1980). Adult specimens were collected weekly into absolute ethanol, with a random 5 individuals barcoded every 2 months to confirm colony purity. Additional ethanol preserved specimens of *D. suzukii*, *D. subpulchrella*, *D. biarmipes*, and *D. immigrans* were obtained from Cornell *Drosophila* Stock Centre, USA, Ehime University *Drosophila* Species Stock Centre, Japan, and the National Institute of Agricultural Botany

East Malling Research Station, UK. Various numbers of adult or larval specimens from each species were combined to form mock communities with total sizes ranging from 100 to 1000 individuals (Supplementary Table 1), then stored in absolute ethanol at -20 °C until DNA extraction.

Table 1: *Drosophila suzukii* surveillance methods used to collect field samples for evaluation of the non-destructive metabarcoding assay, and the number of samples which had at least one successfully sequenced replicate.

Sampling method	**Method Reference**	**Specimens collected in**	**Samples successfully sequenced**
Apple cider vinegar (ACV)	Landolt et al. (2012)	Apple cider vinegar (pH 2.9)	1/8
Synthetic lure (Syn)	Cha et al. (2014)	Synthetic lure (pH 2.5)	6/8
Synthetic lure + propylene glycol + dichlorvos insecticide cube (SPD)	This study	Propylene glycol	5/5
Fruit crush & floatation **(FF)**	Van Timmeren et al. (2017)	In fruit	7/7

Field sampling

To obtain samples representative of the insect diversity expected to be encountered in a real surveillance programme, 55 red cup traps (Lee et al., 2012) were deployed in a sweet cherry (*Prunus avium* L.) orchard and 44 traps in a mixed stone fruit (*Prunus persica* L.) orchard, each located in Mornington and Tatura, Victoria, Australia. Each trap contained one of either apple cider vinegar as attractant and drowning solution (ACV) (Landolt et al., 2012), the synthetic lure of Cha et al. (2014) as attractant and drowning solution (Syn) or the same synthetic lure with a separate propylene glycol drowning solution and a dichlorvos insecticide cube (SPD). Trap catches were collected every 2 weeks over the course of a 10-week period from January to March 2018, and ~1kg of recently fallen fruits collected at each timepoint. These fruits were crushed and agitated in a 15% w/v salt solution and larvae collected using methods described by Van Timmeren et al. (2017), with exception of the salt solution used here being almost twice the concentration of the original study. To ensure a robust validation of the metabarcoding assay on sufficiently sized communities, all field collected samples were combined by week of collection for each sampling method and orchard. Across the two orchards this yielded a total of 22 trapped samples each containing between 200 and 800 adult insect specimens, as well as 7 fruit crush samples containing between 100 to 800 of predominantly larval specimens. A subset of mock and field collected samples were spiked with either 1 or 5 individuals of *D. suzukii*, *D. subpulchrella* or *D. biarmipes* (Supplementary Table 1) then suspended in absolute ethanol within 15mL falcon tubes and stored at -20 °C until DNA extraction.

Non-destructive DNA *extraction*

The non-destructive Qiagen DNeasy based method of Nielsen et al. (2019) was used to extract DNA from each mixed community, in order to retain voucher specimens for morphological confirmation of any detected exotic species. In brief, ethanol was removed from the samples using a 1000 µL pipette and specimens dried overnight to ensure all residual ethanol was evaporated. The mixed specimens were suspended in a 10:1 mix of Qiagen ATL tissue lysis buffer and Proteinase K (Qiagen, Germany), with the total volume of buffer increased proportionally to the number of specimens to ensure all were fully immersed, then incubated for 24 hours at 56 °C and 220 rpm in a shaking incubator. Following incubation, lysate was removed from the specimens and manually loaded into Qiagen 96 well DNeasy extraction blocks using a multichannel pipette, and the remainder of the Qiagen DNeasy Blood & Tissue protocol followed within the QiaCube automated DNA purification workstation (Qiagen, Germany). Voucher specimens retained after non-destructive DNA extraction were resuspended in absolute ethanol and stored at -20 °C.

COI *amplification and sequencing*

Four candidate primers pairs; BF1-BR1 (Elbrecht & Leese, 2017), fwhF2-fwhR2n (Vamos et al., 2017), fwhF2-HexCOIR4 (Marquina, Andersson, et al., 2019) and fwhF2-SauronS878 (Rennstam Rubbmark et al., 2018) producing 254-258 bp amplicons were selected from those determined as high performing in Piper et al. (2021) and appropriate for 2 × 150 bp sequencing. The qualitative and quantitative performance of each primer pair was compared on a subset of 5 mock and 4 field collected samples, then fwhF2-fwhR2n alone was used for the remaining 20 mock and 18 field samples. Each 25 µL PCR reaction consisted of 5 µL 5X MyFi reaction buffer (Bioline, USA), 1 µL of 10 nM forward and reverse primers, 0.8 µL MyFi DNA polymerase, 11.2 µL BSA and 2 µL of variable concentration template DNA. Cycling conditions were an initial denaturation at 94 °C for 2 min, then 30 cycles of 94 °C for 30 sec, 50°C for 45 sec, and 72 °C for 45 sec, followed by a final 2 min extension at 72 °C. Successful amplification was verified on a 2% w/v agarose gel, then amplicons were diluted 1:10 in ddH20 with no further clean-up step. 1 µL of the diluted COI amplicons were further amplified using 7 cycles of real-time PCR to attach 8 bp unique-dual indices and Illumina sequencing adapters (Costello et al., 2018). Cycling conditions for the second PCR were 98 °C for 10 sec, 65 °C for 30 sec, and 72 °C for 30 sec, with each cycle followed by a SYBR Green fluorescence read.

While only a single DNA extraction and PCR replicate per sample was used for the initial comparison of the four candidate primer sets, after the fwhF2-fwhR2n primers were selected all further samples were replicated twice at the DNA extraction stage and 3 times at the PCR stage. DNA extraction replicates were obtained by splitting the lysate from the 24hr incubation into two aliquots and running each through the QiaCube on separate 96 well DNA extraction blocks, while PCR replicates were obtained by splitting the final DNA extract and amplifying each aliquot in separate thermocyclers (Supplementary Fig. 1). As insufficient unique-dual indices were available for all replicated samples, a 'twin-tagging' approach (Axtner et al., 2019) was used where 3 modified versions of the forward and reverse primers containing an additional 2-4 bp inline tag at the 5'- terminus were used to separately amplify each set of PCR replicates (Supplementary Fig. 2). These inline tags were designed to incorporate length variation in order to improve phasing during the critical first cycles of the sequencing process (Lundberg et al., 2013). Two positive control libraries consisting of 13 equimolarly pooled synthetic gBlock gene fragments (Integrated DNA Technologies, USA), each designed to mimic the COI gene of a certain insect family in base composition and structure (Supplementary Note 1), were included alongside all real communities after DNA extraction but prior to PCR amplification.

Following indexing qPCR, melt curve analysis was used to quantify DNA concentrations, then libraries were pooled in equimolar ratios using a Biomek FX^{P} liquid handling robot (Beckman Coulter, USA). Pooled libraries were purified using a 0.8:1 ratio of AMPure XP beads and then sized and quantified using a 2200 TapeStation (Agilent Technologies, USA) and Qubit 3.0 Fluorometer (Thermo Fisher, USA). Final libraries for the primer comparison were diluted to 7 pM, spiked with 5% PhiX, and sequenced on an Illumina MiSeq V2 flow cell using 2 × 150 bp reads, while the remainder of fwhF2-fwhR2n amplified mock and field collected samples were diluted to 100 pM, spiked with 1% PhiX and sequenced on a portion of an Illumina NovaSeq 6000 S2 flow cell lane, again using 2 × 150 bp reads. To minimise the risk of contamination from the laboratory environment, DNA extraction, preparation of PCR master-mix, PCR amplification, and library preparation were each performed in separate rooms using dedicated equipment and pipettes.

Bioinformatics

Sequence reads were demultiplexed using *bcl2fastq* allowing for zero mismatches to the expected index combinations, followed by a second round of demultiplexing for the inline tags using *Seal* in *BBTools* (Bushnell et al., 2017). Demultiplexed sequencing reads (NCBI SRA acc no: XXXXX, to be assigned later) were trimmed of PCR primer sequences using *BBDuK* and any sequences with >1 expected error (Edgar & Flyvbjerg, 2015), <8 unique 2-mers, or any ambiguous 'N' bases were removed. Remaining sequences were denoised using *DADA2* (Callahan et al., 2016), with the error model determined separately for the MiSeq and NovaSeq data using the "pseudo-pooling" mode for increased sensitivity to rare variants. Due to overfitting of the default Loess error model to the binned quality scores provided by the NovaSeq, the estimated error matrix of nucleotide transitions was modified to enforce monotonicity as suggested by the DADA2 developers (Callahan, 2019). Following denoising, the Amplicon Sequence Variants (ASVs) inferred separately from each sequencing run were merged into a single table and any chimeric sequences removed de-novo using the *removeBimeraDenovo* function in *DADA2*. To further remove any non-specific amplification products and pseudogenes, the ASVs were aligned to a Profile Hidden Markov Model (PHMM) of the COI barcode region (Piper et al., 2021) using the *aphid* R package (Wilkinson, 2019), and checked for frame shifts or stop codons that commonly indicate pseudogenes (Roe & Sperling, 2007).

Hierarchical taxonomy was assigned to the filtered ASVs with a minimum bootstrap support of 60% using the *IDTAXA* algorithm (Murali et al., 2018) trained on the curated insect reference database of Piper et al. (2021), followed by additional species level assignment using a nucleotide *BLAST* search against the same reference database (Altschul et al., 1990). Where taxonomic clashes occurred at the species level due to ties between *BLAST* top hits, species occurrence records from the Atlas of Living Australia (https://www.ala.org.au/) and the Australian Faunal Directory (https://biodiversity.org.au/afd/) were used to resolve the most likely species name for the geographic location. Following taxonomic assignment, all samples which received <1000 reads, and all ASVs that were not classified to Arthropoda were removed. A maximum likelihood phylogenetic tree was then constructed from the remaining ASVs using *FastTree* (Price et al., 2009) following the General Time-Reversible (GTR) model (Tavaré, 1986) and gamma distribution of rate variation among sites. Taxonomic identities at the phylum, class and order were used to constrain the deeper topology of the tree, and the constructed phylogeny was rooted on the edge connecting the synthetic positive controls to the rest of the tree. All phylogenetic trees were plotted using the *ggtree* R package (Yu et al., 2017, 2018).

Determining a detection threshold

A baseline detection threshold of 0.01% relative abundance was used to resolve false positive observations within the initial primer comparison, which approximates the expected rate of index-switching of both i5 and i7 indices (Costello et al., 2018; MacConaill et al., 2018). For the later fwhF2-fwhR2n amplified samples, this baseline threshold was compared to 5 additional methods for empirically deriving a detection threshold: (i) the 'unassigned indices' used the abundance ratio of valid (applied during library preparation) to invalid (pairs that could only arise due to switching) index combinations as per Wilcox et al. (2018). (ii) the 'positive control' method used the abundance ratio of synthetic COI sequences that were correctly assigned to the positive control libraries to those that were found in other samples. (iii) the 'mock community' method used the abundance ratio of expected to unexpected taxon observations across all mock communities. (iv) the 'logistic regression' method fit a logistic model of the per-sample relative abundance of each detection, trained on the expected and unexpected taxon observations within the mock communities, with the sequencing run included as an additional covariate to account for run-specific variation in contamination rates (Batovska et al., 2020). With this method the predictive equation from the logistic model describes the probability of each observation being a true positive (Coughlin et al., 1992), and all observations with probability ≥50% were considered detections. (v) the final method used the same logistic regression model but included both the number of DNA extraction and PCR replicates that each observation was detected in as additional covariates. To evaluate the predictive performance of each approach, all taxon observations within the mock communities were randomly split into 80% training and 20% test sets and the logistic regression classifiers and all detection thresholds compared for their ability to remove cross contamination within the test dataset (Quinn et al., 2021). To ensure the comparisons were robust to whichever observations were assigned to the training and test sets, the random splitting, training, and evaluation was repeated 1,000 times and the results averaged.

Statistical analyses

Overlap in detected species between replicates was quantified using Jaccard's index (Jaccard, 1908), and the influence of collection method and sequencing depth on replicate dissimilarity tested for significance using Analysis of Variance (ANOVA) and linear regression respectively. Using the known presence or absence of target specimens spiked into both mock and field collected communities, the diagnostic sensitivity (proportion of known positives that were correctly identified as positives), diagnostic specificity (proportion of known negatives that were correctly identified as negatives), and the overall accuracy of the assay (average of the sensitivity and specificity) were calculated separately for each taxon. Species-specific quantitative bias was estimated for each primer set using a linear regression of the compositional error (ratio between expected and observed abundances) with taxon as the predictor, and the results geometrically centred to be relative to the 'average taxon' as per McLaren et al. (2019).

Community diversity

To ensure comparisons of species diversity (α-diversity) between field collected communities were not confounded by differing sequencing depths between samples (Willis, 2019), the *breakaway* R package was used to estimate the number of unobserved species for each sample using the frequency ratios of detected species (Willis & Bunge, 2015). Once it was confirmed that there were no unobserved species in samples with lower sequencing depths, both the observed species richness and Shannon index

(Shannon, 1948) were calculated using the *phyloseq* R package (McMurdie & Holmes, 2013). ANOVA was then used to test whether differences in α-diversity could be explained by the collection method or the orchard the sampling was conducted in, with post-hoc pairwise comparisons made using Tukey tests. Differences in species composition (β-diversity) between communities was quantified using the weighted-UniFrac distance, which considers both the phylogenetic relatedness and relative abundance of taxa within each sample (Lozupone & Knight, 2005). The effect of sampling method and orchard type on β-diversity was tested for significance using multivariate generalised linear models as implemented in the *manyglm* function from the *mvabund* R package (Wang et al, 2012). Principal coordinate analysis of the weighted-UniFrac distances was used to visualise the clustering of samples, with 95% confidence ellipses drawn using a multivariate t-distribution. All statistical analyses were conducted within the R4.1 statistical programming environment (R Core Team, 2019) using *tidyverse* (Wickham et al., 2019) and *tidymodels* (Kuhn & Wickham, 2020) packages, and figures plotted with *ggplot2* (Wickham, 2016).

Results

Comparison of 4 mini-primer sets

A MiSeq paired-end sequencing run (2 × 150 bp) was conducted for a subset of 5 mock and 4 field collected communities in order to compare the 4 candidate primer sets, yielding 4,743,487 total reads (mean 121,628 ± 6,193 per sample). All taxa within the mock communities were recovered by the 4 primer combinations, apart from *D. biarmipes* which was absent from BF1-BR1 (Fig. 1A). The absence of *D. biarmipes* for this primer set where it should have been present in two samples at 1% abundance was not related to low total sequencing depth, as these libraries received 66,670 and 146,188 reads respectively. Instead, low amplification efficiency for BF1-BR1 on this taxon left its relative abundance below the 0.01% detection threshold used to control index-switching within this first sequencing run. In addition to the dropout of *D. biarmipes*, between 6 and 9 false positive detections per primer set were recorded within the mock communities (Fig. 1A). Of these, only the *D. immigrans* and *D. hydei* false positives were recorded across all primers with a relative abundance >1%, indicating they may be due to physical cross contamination of a specimen when the mock communities were assembled. In contrast,

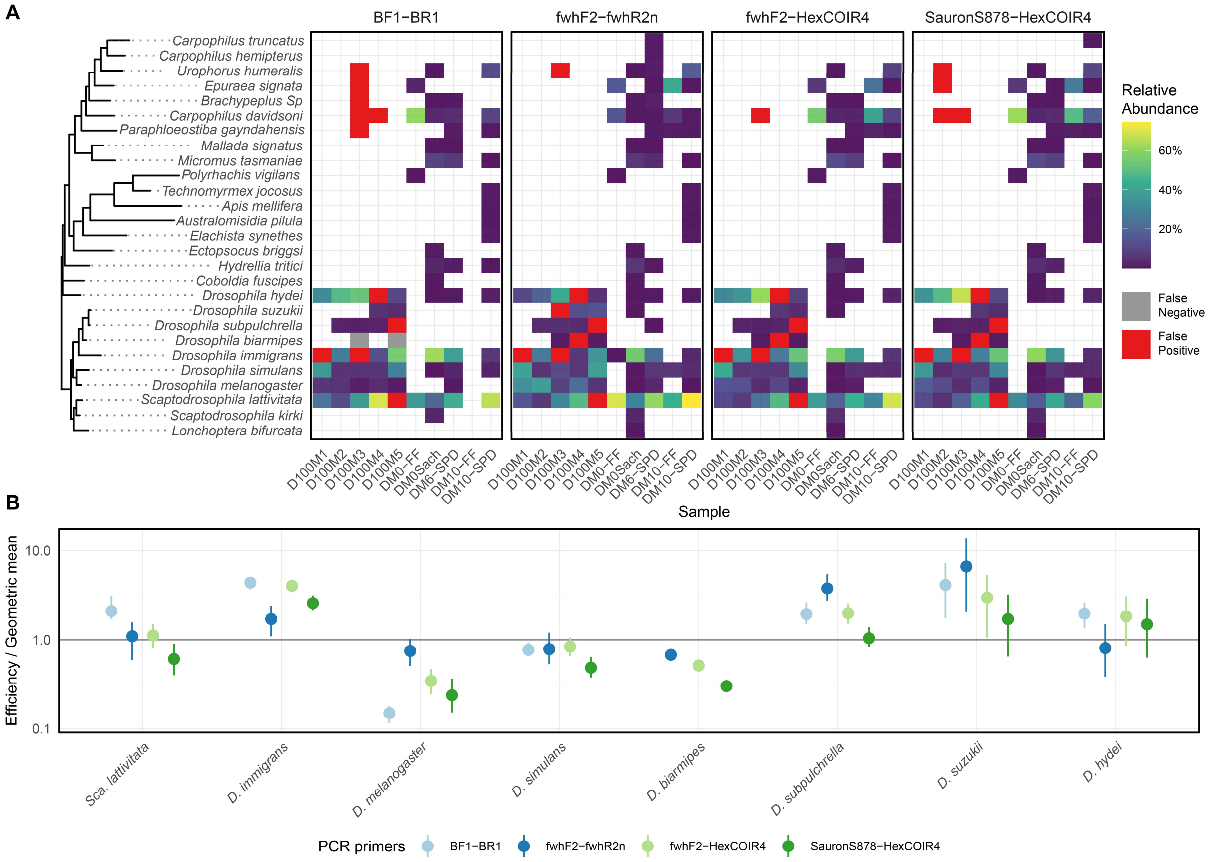


Figure 1 A) Per-sample relative abundance of detected species across the four evaluated primer sets, with false positive or false negative results within the mock community samples highlighted. **B)** Relative efficiency of each Drosophila species in mock communities compared to the geometric mean efficiency, with 95% confidence intervals obtained from 1000 bootstrap resamples.

the remaining false positives were detected with relative abundances between 0.01% and 0.08%, which alongside their presence at high abundance in other sequenced communities suggests they arose through index-switching. For the 4 field samples used for primer evaluation, the fwhF2-fwhR2n primers detected 38 taxa, while fwhF2-HexCOIR4 and SauronS878-HexCOIR1 detected 32 and 33 taxa respectively (Fig. 1A). Despite an entire sample amplified with BF1-BR1 receiving insufficient sequence reads to pass quality control steps, BF1-BR1 still detected 34 distinct taxa. Primer-specific differences were also seen in the identities of detected species (Fig. 1A), with *Carpophilus hemipterus* only being detected by fwhF2-fwhR2n, *Carpophilus truncatus* only by fwhF2-fwhR2n and SauronS878-HexCOIR1, and *Lonchoptera bifurcata* detected with all primer combinations except BF1-BR1. However, in all cases where a taxon was not detected by every primer combination it was <1% relative abundance within the respective sample.

In addition to qualitative differences in taxa detected, primer specific quantitative biases were also seen across the mock community taxa (Fig. 1B). BF1-BR1 showed a high efficiency for S. *lattivitata*, *D. immigrans*, *D. subpulchrella*, *D. suzukii*, and *D. hydei*, a below average efficiency for *D. simulans*, and a very low efficiency for *D. melanogaster*, while the drop-out of *D. biarmipes* meant the efficiency for this taxon was unable to be measured. fwhF2-HexCOIR4 showed similar quantitative performance to BF1-BR1 across most taxa, except for S. *lattivitata* which showed an average efficiency, and *D. melanogaster* where efficiency was slightly higher. In contrast, fwhF2-fwhR2n showed close to average efficiency for S. *lattivitata*, *D. melanogaster*, *D. simulans*, and *D. hydei*, while preferentially amplifying *D. immigrans*, *D. subpulchrella, and D. suzukii*, leaving *D. biarmipes* with slightly below average efficiency. Finally, SauronS878-HexCOIR4 preferentially amplified *D. immigrans*, but showed average efficiency for *D. suzukii*, *D. subpulchrella*, and *D. hydei*, and below average efficiency for *D. biarmipes*, S. *lattivitata*, *D. melanogaster*, and *D. simulans*. Ultimately, the fwhF2-fwhR2n primer combination was chosen to proceed for the remainder of the study as it identified the most species in the field samples (Fig. 1A) and showed the highest efficiency for the targets *D. suzukii*, *D. subpulchrella* and *D. biarmipes* (Fig. 1B), which should increase the probability of detecting them even at low abundance.

Replicate similarity

The 22 field collected samples, remaining 20 mock communities, and 2 synthetic positive control samples were each replicated twice at the DNA extraction and 3 times at the PCR stage, and the resulting 264 libraries sequenced on a portion of a NovaSeq S2 flow cell lane. This yielded a total of 314.5 million reads following bioinformatic quality control (mean 1,502,245 ± 121,999 per replicate), however, a large number of replicate dropouts occurred across both the mock and field collected communities (Fig. 2A). For the mock communities, 79% of the replicates from the adult samples and 67% from the larval samples were successfully sequenced. While for the field samples, 80% of replicates from the SPD treatment, 67% from the fruit crush, 58% from synthetic lure, and only 10% of replicates from the apple cider vinegar samples were successful. Most of these replicate dropouts occurred within the second set of PCR replicates from extraction replicate 1, where 39 of 50 were unsuccessful, including one of the positive controls (Fig. 2A). As each set of replicates was processed in a separate microtiter plate and thermocycler (Supplementary Fig. 1), this likely indicates a systematic failure during PCR amplification

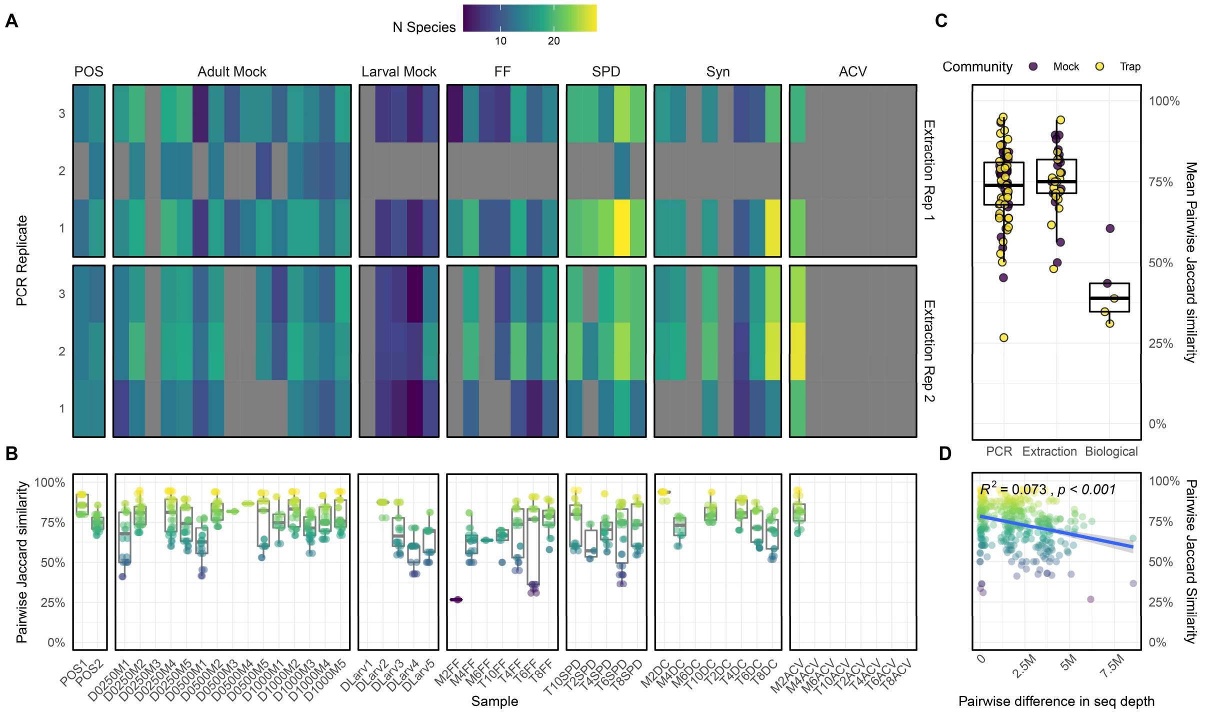


Figure 2 A) Total number of species observed within each replicated sample prior to application of detection threshold, with complete replicate dropouts indicated in grey. **B)** Pairwise Jaccard similarity coefficients (presence/absence of taxa) between all replicates of each sample **C)** Mean Jaccard similarity coefficient between PCR replicates of the same DNA extract, DNA extraction replicates of the same sample, and separate samples obtained using the same collection method (biological replicates). **D)** Relationship between pairwise Jaccard similarity and sequencing depth difference for all replicates.

or when these replicates were pooled into the final libraries. With the exception of this whole set of replicates, dropouts of singular replicates seemed to randomly occur across the samples (Fig. 2A). On the other hand, when considering samples where no replicates were successfully sequenced, there were apparent DNA preservation effects relating to collection method used (Table 1). For the field collected samples, all the fruit crush and SPD samples had at least one successfully sequenced replicate which could be analysed further, while 75% of the synthetic lure samples and only 12.5% of apple cider vinegar samples produced any usable data. For the mock communities, 93% of the adult communities and 80% of the larval communities as well as both positive control samples had at least one successfully sequenced replicate.

While most successful replicates reached saturation in species accumulation (Supplementary Fig. 3), the pairwise Jaccard similarity between each ranged from 25% to 98% (Fig. 2B), and showed a weak but statistically significant relationship with pairwise differences in sequencing depth ($R^2 = .073$, $p < .001$; Fig. 2D). A significant relationship was also found between replicate dissimilarity and the community type (field or mock) or collection method used ($F_{(4, 683)} = 15.43$, $p < .001$), which post-hoc comparisons revealed to be driven by replicates of the larval mock communities, fruit crush, and SPD being less similar to each other than those from the synthetic lure or the adult mock communities (p < .001). However, in all cases where a detection occurred in ≤50% of the sequenced replicates, the taxon was ≤1% relative abundance within the physical community. Overall, extraction replicates of the same samples were as similar to each other as PCR replicates of the same DNA extraction, but separate samples collected from the same orchard using the same collection method (biological replicates) showed much less overlap in species detected (Fig. 2C). Finally, there were no significant differences seen in the quantitative performance between the 3 tagged primer sets on any of the mock community taxa, confirming that replicate dissimilarity was not due to the twin-tagging approach to multiplexing (Supplementary Fig. 4).

Determining a detection threshold

All methods for deriving a detection threshold increased the proportion of true positive detections over the uncorrected data, however the degree of improvement varied substantially (Fig. 3A). While the baseline 0.01% filtering threshold more than halved the number of false positives for the MiSeq run, this threshold was overly strict for the NovaSeq and introduced a significant number of false negatives. Surprisingly, the positive control method (included in the NovaSeq run only) performed worse than the baseline threshold, only marginally reducing the number of false positives compared to the uncorrected data. As these two positive control samples were included after DNA extraction, this limited success could indicate that false positives arose through physical cross contamination during or prior to DNA extraction, rather than index switching. Yet the unassigned indices method, which should only account for index switching, removed substantially more false positives than the positive control approach. The mock community method on the other hand did not improve the proportion of true positives above the baseline threshold for the MiSeq run but performed the best for the NovaSeq data (Fig. 3A). While both logistic regression classifiers fit to this same mock community abundance information performed slightly worse than using just using the abundance

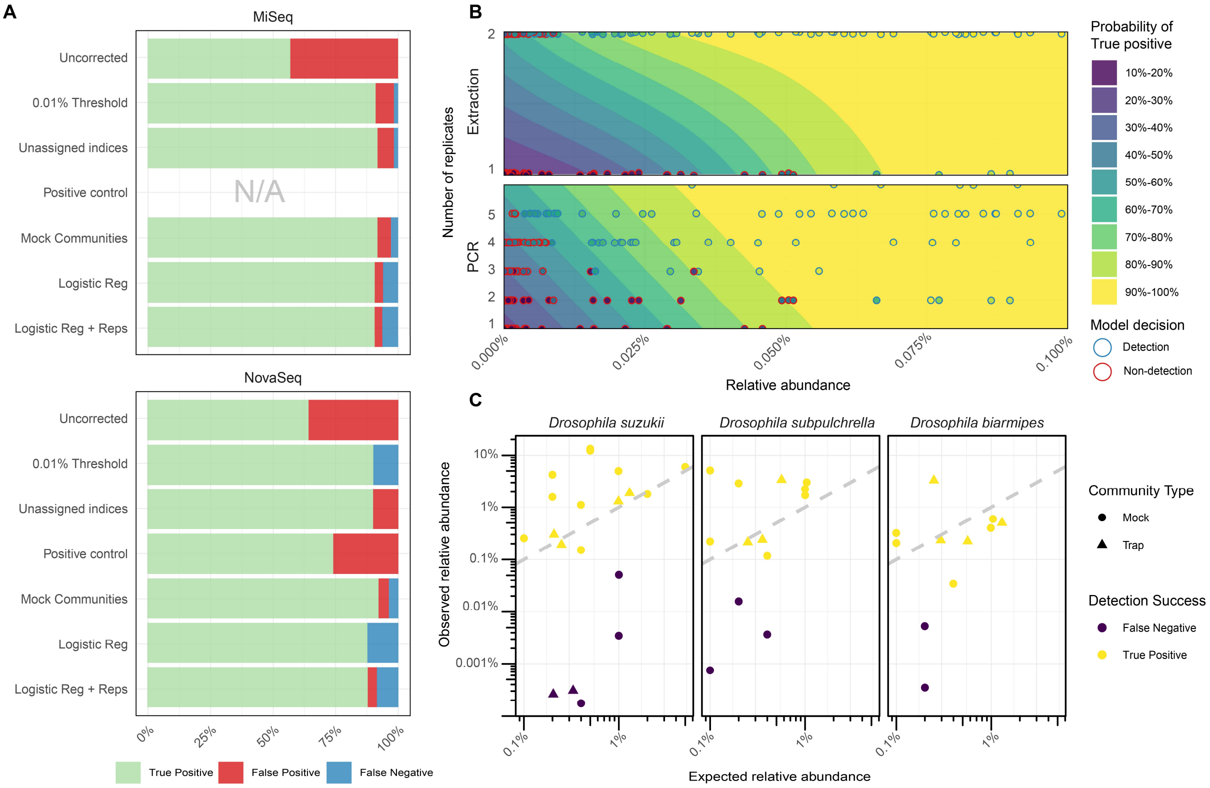


Figure 3 A) The proportion of true positives, false positives, and false negatives across both the MiSeq and NovaSeq run following application of various methods for deriving a detection threshold. **B)** Probability surface for the final logistic regression classifier shows how the likelihood of an observation being a true positive increases with higher relative abundance and the number of extraction replicates (top), and PCR replicates (bottom) it was observed in. Classification of low abundance observations from the sequenced mock and trap communities overlaid. **C)** Relationship between expected (from morphology) and observed (from sequencing) relative abundance for target species spiked into mock and trap communities, with each observation coloured by whether it was classified as a true positive or false negative by the logistic regression model. Dashed line indicates perfect relationship between expected and observed relative abundances.

ratio as a threshold, ultimately, the logistic regression classifier incorporating both abundance and replicate information was chosen for use on the field samples, as framing the trade-off between false positives and negatives in terms of the probability of an observation being a true detection provides advantages for interpretability. When this final model was trained again on the full dataset, the most important model covariates were the number of DNA extraction replicates a taxon was observed in (t = 2.11), followed by its mean relative abundance across all replicates (t = 1.44), the sequencing run (t = 1.04) and finally the total number of PCR replicates it was observed in (t = 0.680). The differing importance of extraction and PCR replicates for detection efficiency can be seen on the probability surface for the final logistic regression model (Fig. 3B).

Detection success in field and mock samples

Following application of the logistic regression model, *D. suzukii* was successfully detected in 14 of 19 positive samples with no false positives, giving a sensitivity of 73.6% and specificity of 100%. The secondary targets, *D. subpulchrella* and *D. biarmipes* were detected in 10 of 13 and 8 of 11 positive samples respectively, again with no false positive detections, resulting in a sensitivity of 76% and 81% for each and specificity of 100% for both. The accuracy of the assay was 86%, 88% and 91% for *D. suzukii*, *D. subpulchrella*, and *D. biarmipes* respectively. While all false negative results for the three targets occurred when the respective taxa were ≤1% relative abundance within the physical community, there was no clear relationship between relative abundance and detection failure (Fig. 3C). This was most apparent for the false negatives for *D. suzukii* in mock community samples D500M3 and D500M4 where they were spiked in at 1% relative abundance, but successful detection in the samples D1000M3 and D1000M4 where the overall community composition was similar but the targets were at 0.5% relative abundance (Fig. 3C, Supplementary Table 1).

Community diversity

A total of 1,281 specimens were collected from the cherry orchard, and 4,772 from the stone fruit orchard over the entirety of the 10-week trapping period (Fig. 4B). Of these, 654 specimens were caught by the apple cider vinegar traps, 1,640 by the synthetic lure (Syn), and 2,224 by the synthetic lure treatment with the propylene glycol and insecticide cube (SPD). On the other hand, fruit crushing and salt flotation (FF) collected at least 1,535 specimens, with the absolute number likely being much higher due to some larvae being too small to accurately count. Following sequencing and application of the logistic regression detection model, a total of 46 unique insect taxa were identified within the trap samples, 45 of which could be successfully assigned to species level taxonomy. This bycatch diversity included 19 Diptera (excluding the 3 spiked in targets), 10 Coleoptera, 7 Hymenoptera, 2 Lepidoptera and Hemiptera, and a single Neuropteran species (Fig. 4A). Although the PCR primers used were designed to amplify insects (Vamos et al., 2017), the spider species *Badumna longinqua* and *Tenuiphantes tenui*s were also detected in the fruit

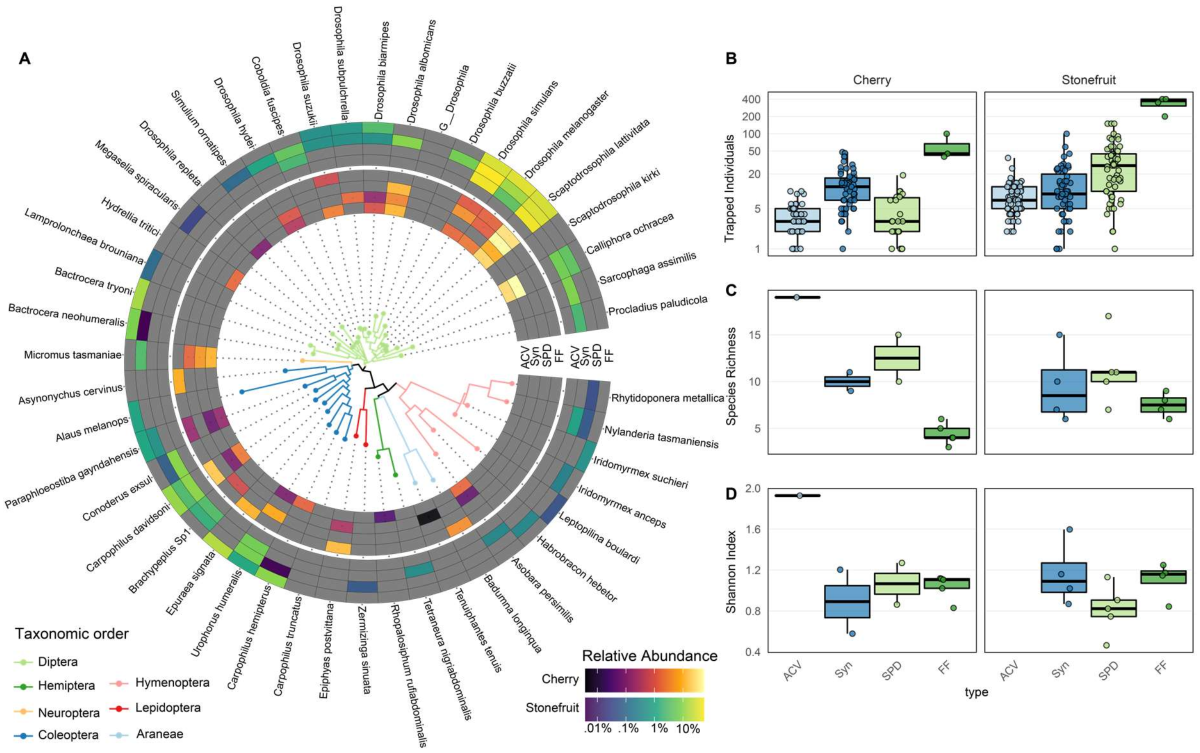


Figure 4 A) Phylogenetic relationships between detected taxa, with the mean relative abundance of each taxon displayed by sampling method for the cherry (inner heatmap) and stone fruit (outer heatmap) orchards. **B)** Number of individual specimens collected from each orchard by each sampling strategy over the total course of the 10-week trapping period, displayed on a pseudo-log scale. **C)** Species richness and **D)** Shannon index for each community following metabarcoding. * Spiked-in exotic species, not present in Australia.

crush and synthetic lure samples from the cherry orchard. When the identities of bycatch taxa were compared to species occurrence records, all were confirmed to be endemic or previously recorded in Australia.

For the field collected communities, the species richness estimated by the breakaway model matched the number of detected species, indicating that all species in the communities had been captured at that sequencing depth. For these communities, the sampling method significantly affected species richness (ANOVA, $F_{(3, 19)} = 5.19$, $p = .009$) (Fig. 4C), with the single successful ACV sample (17 species) containing significantly more species than the fruit crush treatment (mean 6.78 ± 1.15, $p = .029$), but no significant differences were found between any of the other sampling methods ($p > .05$). Significant differences in Shannon diversity were also found between sampling methods ($F_{(3, 19)} = 5.23$, $p = .008$), primarily driven by the SPD treatment having many more taxa at low abundance than the ACV treatment ($p < .001$) (Fig. 4B). In contrast, no significant differences were found between the cherry and stone fruit orchards in either of the α-diversity metrics ($F_{(1, 21)} = 0.76, 0.49$, both $p > .05$).

There were significant effects of both orchard (Likelihood Ratio Test [LRT] = 126.98, p = .002) and sampling method (LRT = 224.53, p < .001) on community composition, with no interaction effect found between the two (p > .05). Principal coordinate analysis revealed that while the SPD samples from the stone fruit orchard clustered tightly together, the 95% confidence ellipses of the synthetic lure and the fruit crush samples from the same orchard completely overlapped these (Supplementary Fig. 5). In contrast, the synthetic lure and fruit crush samples from the cherry orchard formed discrete clusters separated from each other, as well as all samples from the stone fruit orchard. Taken together, this indicates that differences in β-diversity are primarily driven by a distinct cohort of species occurring in each orchard, as well as between the synthetic lure and fruit crush samples within the cherry orchard. For the stone fruit orchard on the other hand, the differences in species occurrence and abundance between sampling methods were much less pronounced (Fig. 4A).

Discussion

Whilst originally developed for studying biodiversity, metabarcoding approaches are increasingly being applied to the detection of invasive species in aquatic and terrestrial environments (Brown et al., 2016; Piper et al., 2019; Tedersoo et al., 2019). Here we demonstrate the use of a non-destructive metabarcoding assay to detect the rapidly spreading global pest *Drosophila suzukii* and its close relatives *D. biarmipes* and *D. subpulchrella* within large unsorted trap catches. By circumventing the time-consuming and labour-intensive process of morphological sorting, adoption of metabarcoding assays by diagnostic laboratories could enable a substantial increase in the geographic scale and intensity of *D. suzukii* surveillance, and thus the likelihood of detecting a new incursion. Nevertheless, our results show that aspects of trap design and laboratory protocols may need to be reconsidered if metabarcoding is to be successfully adopted for invasive insect identification.

Apple cider vinegar is the most commonly used attractant and drowning solution for *D. suzukii* surveillance (Hamby et al., 2014; Harris et al., 2014; Landolt et al., 2012; Mazzetto et al., 2015), yet almost all communities collected using this method failed to produce a sequenceable amplicon. This limited success may be related to trapped specimens being immersed within the highly acidic and watery solution for up to two weeks between traps being set and collected, which can cause degradation of DNA molecules (Lindahl, 1993). Even so, if pH or hydrolysis mediated DNA degradation were the only factors involved, a comparable failure rate would be expected for those communities trapped in the more acidic synthetic lure (Table 1). Furthermore, even if the DNA of the trapped specimens were completely degraded, the ethanol preserved *D. suzukii*, *D. subpulchrella* and *D. biarmipes* specimens that were spiked into these samples should have produced some data. On the other hand, apple cider vinegar is a complex matrix containing various polysaccharides, polyphenolics, and tannins, all of which have PCR inhibiting properties (Jara et al., 2008). While all specimens were rinsed with ethanol, and the DNA extraction method involved two clean-up steps, carry over of some residual inhibitors may have prevented amplification for many of these samples (Martins et al., 2019). In contrast, traps employing the synthetic attractant lure but using a separate propylene glycol drowning solution adequately preserved specimens for metabarcoding analysis. Propylene glycol shows promise for use in *Drosophila* surveillance traps when molecular methods are to be used for identification, being cheap, non-flammable, non-evaporative, and able to effectively preserve DNA for up to 6 months (Martoni et al., 2021; Nakamura et al., 2020). To facilitate the use of liquid preservatives such as propylene glycol, new trap designs should physically separate the highly acidic lures from the drowning solution, either in a separate compartment within the trap or a controlled release sachet (Larson et al., 2020). Nevertheless, DNA degradation or PCR inhibition were not the only factors in play, and further laboratory optimisation may be required to resolve the seemingly random dropouts of single replicates that occurred across both the mock and field collected communities.

Early detection surveillance depends upon swift diagnostic turnaround to ensure that quarantine and intervention procedures are appropriate and effective. In light of this, our study opted for a rapid laboratory protocol that omitted any normalisation or purification of DNA between the extraction and both PCRs. While similar rapid protocols have been successfully applied to destructively homogenised specimens (Elbrecht & Steinke, 2019), the extra variability introduced by the non-destructive protocol may have contributed to the large number of replicate dropouts observed. Therefore, we suggest that future studies employing non-destructive DNA extractions normalise the resulting extracts and PCR amplicons to similar concentrations in order to increase sequencing success. While these additional steps will increase laboratory processing time, ultimately it is the sequencing process itself which represents the longest step in a metabarcoding assay, taking between 40-56 hours depending on the HTS platform (Piper et al., 2019). While the Illumina NovaSeq platform used in our study is currently the most cost-effective for large numbers of samples, drawing together hundreds of trap samples on a regular basis without in-turn increasing diagnostic turnaround times may prove a logistical challenge for smaller surveillance programmes. Therefore, lower throughput platforms such as the Illumina MiSeq will likely remain important into the future, despite their higher cost per gigabase of data and longer runtimes (Elbrecht et al., 2017).

In addition to dropouts of some replicates, there was also variability in the taxa detected between successfully sequenced replicates. While replicate dissimilarity showed a slight relationship with sequencing depth differences, with a mean 3.2 million sequence reads per replicate the sequencing depths obtained in our study were an order of magnitude higher than most metabarcoding studies (Singer et al., 2019). Conversely, all taxa that were detected in 50% or less replicates were below a 1% physical relative abundance within the respective community, suggesting that the taxonomic dropouts were not simply a product of insufficient sequencing depth as some have proposed (Smith & Peay, 2014), but instead may be due to stochastic sampling of DNA molecules from low abundance taxa as small quantities of liquid go through the metabarcoding pipeline (Leray & Knowlton, 2017). This phenomenon, also known as pipeline noise, increases dissimilarity between replicated samples (Zhou et al., 2013) and can be further exacerbated by taxonomic biases from different species traits (McLaren et al., 2019). Previous studies conducting metabarcoding on preservative ethanol have shown higher variance in taxon detections compared to tissue homogenisation, and that results are much more sensitive to exoskeleton hardness and specimen morphology, rather than just specimen biomass (Marquina, Esparza-Salas, et al., 2019; Zizka et al., 2019). The leeching of DNA from specimens into ethanol is conceptually similar to the non-destructive DNA extraction used here, and therefore we expect similar issues to have played a role in our study. Further mechanistic research will be required to better understand the specific biases of non-destructive methods; yet our results suggest that these approaches may best be considered closer to environmental DNA metabarcoding, which requires a higher level of replication to maximise species detection (Alberdi et al., 2018; Ficetola et al., 2015; Mata et al., 2019). These replicates should consist of both DNA extraction and PCR replicates, as we found both these stages introduced variation in the species detected. That said, including technical replicates should not come at the expense of reduced biological samples, as regardless of the effectiveness of metabarcoding or any other diagnostic assay, if an insect is not caught in a trap, it does not necessarily mean it is absent in the area (Low-Choy, 2015). A variety of 'occupancy' models have been developed to account for this imperfect detection through use of multiple samples, often taken at repeated visits to a site (Ji et al., 2020; Schnell et al., 2015b). By estimating the probability of true species occurrence, occupancy models provide a more accurate understanding of the distribution of pests across the landscape (Allen et al., 2021), which may prove useful as part of area-wide management programmes in regions where *D. suzukii* has already been introduced (Gilioli et al., 2013). Nevertheless, in countries such as Australia where *D. suzukii* is currently absent, any detection even in a single sample would result in an immediate management response.

Molecular recombination of oligonucleotide indices used to label samples during sequencing can cause taxa from one sample to "bleed" into others and must be controlled for using a detection threshold (Piper et al., 2019). Use of positive control samples in the form of synthetic sequences or taxa 'alien' to the study environment has previously been proposed for empirically measuring and accounting for the run-specific contamination rate (Galan et al., 2018; Palmer et al., 2017; Piper et al., 2019). In our study, however, we found this approach drastically underestimated cross contamination, underperforming compared to simply placing a minimum relative abundance threshold of 0.01% across the dataset. Indeed, none of the evaluated methods for empirically deriving a detection threshold were able to increase the proportion of true positives detections above 90%. This suggests that the similar abundances recorded for taxa close to the limits of detection and false positive observations introduced by index switching means there will always be a trade-off between type I and II error (Alberdi et al., 2019). Despite this, the ability of the logistic regression model to frame this trade-off in terms of a probability that an observation is a true detection (determined by the abundance of each ASV and the number of replicates it was detected in) provides benefits for practical interpretation of the results. The coefficients of the logistic regression model highlight that true sequences are likely to be present in more replicates at higher abundance, a concept that has previously been formalised by Zepeda-Mendoza et al. (2016) and integrated into the software package *begum* (Yang et al., 2021). While in our study a simple probability threshold of 50% was used to consider an observation a true detection or not, this threshold could be further tailored to the specific goals and statistical power desired by the surveillance programme (Whittle et al., 2013). For instance, a biodiversity survey may prefer a stricter threshold to ensure only the most robust detections are recorded (Alberdi et al., 2019), whilst an invasive species surveillance programme may opt for a more lenient threshold to maximise sensitivity, as the economic consequences of a false negative are much higher (Jarrad et al., 2011). Nevertheless, the logistic regression model, as well as the mock community and positive control methods all require a portion of each sequencing run to be pre-allocated to mock communities or positive controls, which introduces additional sequencing costs. For studies where this may not be practical, the abundance ratio of correctly assigned to unassigned index combinations allows the contamination rate to be estimated post-hoc without requiring inclusion of additional control samples. Alternatively, the twin-tagging approach used in this study to differentiate PCR replicates could be expanded to ensure every library contains a completely unique twin-tag as well as the unique Illumina indexes. The extra power to identify switched molecules enabled by this approach has recently been shown to alleviate cross contamination issues altogether (Yang et al., 2021), yet comes at the substantial upfront cost of purchasing separate primer oligos for each sample and replicate.

Besides the spiked-in target species, metabarcoding revealed the identity of diverse arthropod communities collected as bycatch through *D. suzukii* surveillance methods. Communities extracted from fallen fruit were the least diverse, but showed the most variability between samples, possibly due to limited larval dispersal creating a patchy distribution across fallen fruit. In agreement with previous comparisons of *D. suzukii* attractants, the synthetic lure outperformed the apple cider vinegar in both the number of specimens collected and selectivity for *Drosophila* species (Burrack et al., 2015; Cha et al., 2018; Tonina et al., 2018). The SPD treatment on the other hand was slightly less selective than the synthetic lure on its own, likely due to the inclusion of the dichlorvos insecticide cube, the effects of which were clearly illustrated by the presence of the ant species *Iridomyrmex suchieri* and *Iridomyrmex anceps* within this treatment. These ants may have entered traps to prey on already collected insects and would have been able to safely exit those traps which solely relied upon flying insects drowning in the attractant solution. Predation of trapped specimens by ants and other species has been documented by a number of studies (Armstrong & Richman, 2007; De Groot & Nott, 2003; Lynegaard et al., 2014; Martín et al., 2013), and this raises an intriguing question about whether predation could be an additional source of false negatives for surveillance programmes.

Whilst all the bycatch identified in our study were either endemic or previously recorded in Australia, in other cases metabarcoding has revealed the presence of unanticipated or cryptic exotic taxa that have been missed by formerly targeted surveys (Batovska et al., 2020; Simmons et al., 2016). While metabarcoding is not the only novel high-throughput diagnostic assay, with other recent advances such as hybridisation probes (Wilcox et al., 2018), and CARMEN–Cas13 (Ackerman et al., 2020) also offering sensitive detection of multiple targets, these alternatives required the targets to be defined *a-priori*. Therefore, while the ability for metabarcoding to be conducted on unsorted trap catches is a significant advance in itself, the universal nature of metabarcoding primers could substantially expand the range of organisms within the scope of a diagnostic laboratory without having to redesign the assay for the next emerging pest (Piper et al., 2021). Adoption of high-throughput metabarcoding assays for screening of mixed trap catches therefore offers a viable method for increasing the geographic scale and intensity of insect pest surveillance that could be readily expanded to the next emerging threat.

Acknowledgments

The authors would like to thank Jessica Henneken, Shahidul Islam, Kevin Farnier and Farrukh Baig for their assistance with field work, Lea Rako and Shane McEvey for their taxonomic support, and Francesco Martoni for assisting to validate the unique-dual indices used in this study. We further thank the Cornell Drosophila Species Stock Centre, Ehime University Drosophila Species Stock Centre, and the National Institute of Agricultural Botany East Malling Research Centre for providing the exotic *Drosophila* specimens used in this work. Field work for this study was conducted on the traditional lands of the Yorta Yorta and Bunurong people and we pay our respects to elder's past and present. Financial support was provided by Agriculture Victoria's Improved Market Access for Horticulture programme (CMI105584), and 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. A.M.P. is further supported by an Australian Government Research Training Program Scholarship.

Author contributions

A.M.P., J.P.C. and M.J.B. conceptualised the study. A.M.P. performed all field sampling, laboratory procedures, bioinformatic and statistical analyses, and wrote the first draft of the manuscript with input and supervision from J.P.C., N.O.I.C. and M.J.B. All authors read and approved the final manuscript.

Data Archiving Statement:

Raw sequence reads have been uploaded to NCBI SRA acc no: (XXXXX, to be assigned later) and final OTU tables and reference sequences used to make taxonomic assignments are available from dryad reference no: (XXXXX, to be assigned later). All code required to reproduce the bioinformatic and statistical analyses presented in this manuscript is available at the following GitHub repository: https://github.com/alexpiper/Drosophila_metabarcoding

References

Ackerman, C. M., Myhrvold, C., Thakku, S. G., Freije, C. A., Metsky, H. C., Yang, D. K., Ye, S. H., Boehm, C. K., Kosoko-Thoroddsen, T. S. F., Kehe, J., Nguyen, T. G., Carter, A., Kulesa, A., Barnes, J. R., Dugan, V. G., Hung, D. T., Blainey, P. C., & Sabeti, P. C. (2020). Massively multiplexed nucleic acid detection with Cas13. *Nature*, 582(7811), 277–282. https://doi.org/10.1038/s41586-020-2279-8

Alberdi, A., Aizpurua, O., Bohmann, K., Gopalakrishnan, S., Lynggaard, C., Nielsen, M., & Gilbert, M. T. P. (2019). Promises and pitfalls of using high-throughput sequencing for diet analysis. *Molecular Ecology Resources*, 19(2), 327–348. https://doi.org/10.1111/1755-0998.12960

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

Allen, M. C., Nielsen, A. L., Peterson, D. L., & Lockwood, J. L. (2021). Terrestrial eDNA survey outperforms conventional approach for detecting an invasive pest insect within an agricultural ecosystem. *Environmental* DNA, 00, 1–11. https://doi.org/10.1002/edn3.231

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

Anderson, C., Low-Choy, S., Whittle, P., Taylor, S., Gambley, C., Smith, L., Gillespie, P., Löcker, H., Davis, R., & Dominiak, B. (2017). Australian plant biosecurity surveillance systems. *Crop Protection*, *100*, 8–20. https://doi.org/10.1016/j.cropro.2017.05.023

Armstrong, J. S., & Richman, D. B. (2007). Interference of Boll Weevil Trapping by Spiders (Araneida) and an Evaluation of Trap Modification to Reduce Unwanted Arthropods. *Journal of Entomological Science*, *42*(3), 392–398. https://doi.org/10.18474/0749-8004-42.3.392

Asplen, M. K., Anfora, G., Biondi, A., Choi, D. S., Chu, D., Daane, K. M., Gibert, P., Gutierrez, A. P., Hoelmer, K. A., Hutchison, W. D., Isaacs, R., Jiang, Z. L., Kárpáti, Z., Kimura, M. T., Pascual, M., Philips, C. R., Plantamp, C., Ponti, L., Vétek, G., … Desneux, N. (2015). Invasion biology of spotted wing Drosophila (Drosophila suzukii): a global perspective and future priorities. *Journal of Pest Science*, *88*(3), 469–494. https://doi.org/10.1007/s10340-015-0681-z

Axtner, J., Crampton-platt, A., Lisa, A. H., Mohamed, A., Xu, C. C. Y., Yu, D. W., & Wilting, A. (2019). An efficient and robust laboratory workflow and tetrapod database for larger scale environmental DNA studies. *GigaScience*, *8*(4), giz029. https://doi.org/10.1093/gigascience/giz029

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

Bock, I. R., & Parsons, P. A. (1980). Culture methods for species of the Drosophila (Scaptodrosophila) coracina group. *Drosophila Information Service*, *55*, 147–148.

Boughdad, A., Haddi, K., El Bouazzati, A., Nassiri, A., Tahiri, A., El Anbri, C., Eddaya, T., Zaid, A., & Biondi, A. (2021). First record of the invasive spotted wing Drosophila infesting berry crops in Africa. *Journal of Pest Science*, *94*, 261–271. https://doi.org/10.1007/s10340-020-01280-0

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. https://doi.org/10.1111/ddi.12465

Burrack, H. J., Asplen, M., Bahder, L., Collins, J., Drummond, F. A., Guédot, C., Isaacs, R., Johnson, D., Blanton, A., Lee, J. C., Loeb, G., Rodriguez-Saona, C., Timmeren, S. Van, Walsh, D., & McPhie, D. R. (2015). Multistate comparison of attractants for monitoring Drosophila suzukii (Diptera: Drosophilidae) in blueberries and caneberries. *Environmental Entomology*, *44*(3), 704–712. https://doi.org/10.1093/ee/nvv022

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLoS ONE*, *12*(10), e0185056. https://doi.org/10.1371/journal.pone.0185056

Calabria, G., Máca, J., Bächli, G., Serra, L., & Pascual, M. (2012). First records of the potential pest species Drosophila suzukii (Diptera: Drosophilidae) in Europe. *Journal of Applied Entomology*, 136, 139–147. https://doi.org/10.1111/j.1439-0418.2010.01583.x

Callahan, B. J. (2019). *Consequences of using dada2 on NovaSeq data [Discussion post]*. https://github.com/benjjneb/dada2/issues/791#issuecomment-502256869

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7), 581–583. https://doi.org/10.1038/nmeth.3869

Carew, M. E., Coleman, R. A., & Hoffmann, A. A. (2018). Can non-destructive DNA extraction of bulk invertebrate samples be used for metabarcoding? *PeerJ*, 6, e4980. https://doi.org/10.7717/peerj.4980

Cha, D. H., Adams, T., Rogg, H., & Landolt, P. J. (2012). Identification and Field Evaluation of Fermentation Volatiles from Wine and Vinegar that Mediate Attraction of Spotted Wing Drosophila, Drosophila suzukii. *Journal of Chemical Ecology*, 38(11), 1419–1431. https://doi.org/10.1007/s10886-012-0196-5

Cha, D. H., Adams, T., Werle, C. T., Sampson, B. J., Adamczyk, J. J., Rogg, H., & Landolt, P. J. (2014). A four-component synthetic attractant for Drosophila suzukii (Diptera: Drosophilidae) isolated from fermented bait headspace. *Pest Management Science*, 70(2), 324–331. https://doi.org/10.1002/ps.3568

Cha, D. H., Hesler, S. P., Wallingford, A. K., Zaman, F., Jentsch, P., Nyrop, J., & Loeb, G. M. (2018). Comparison of Commercial Lures and Food Baits for Early Detection of Fruit Infestation Risk by Drosophila suzukii (Diptera: Drosophilidae). *Journal of Economic Entomology*, 111(2), 645–652. https://doi.org/10.1093/jee/tox369

Cini, A., Ioriatti, C., & Anfora, G. (2012). A review of the invasion of Drosophila suzukii in Europe and a draft research agenda for integrated pest management. *Bulletin of Insectology*, 65(1), 149–160.

Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., Granger, B., Green, L., Howd, T., Mason, T., Vicente, G., Dasilva, M., Brodeur, W., DeSmet, T., Dodge, S., Lennon, N. J., & Gabriel, S. (2018). Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*, 19, 332. https://doi.org/10.1186/s12864-018-4703-0

Coughlin, S. S., Trock, B., Criqui, M. H., Pickle, L. W., Browner, D., & Tefft, M. C. (1992). The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. *Journal of Clinical Epidemiology*, 45(1), 1–7. https://doi.org/10.1016/0895-4356(92)90180-U

Darling, J. A., Pochon, X., Abbott, C. L., Inglis, G. J., & Zaiko, A. (2020). The risks of using molecular biodiversity data for incidental detection of species of concern. *Diversity and Distributions*, 26(9), 1116–1121. https://doi.org/10.1111/ddi.13108

David, J. R., Gibert, P., Legout, H., Pétavy, G., Capy, P., & Moreteau, B. (2005). Isofemale lines in Drosophila: An empirical approach to quantitative trait analysis in natural populations. *Heredity*, 94(1), 3–12. https://doi.org/10.1038/sj.hdy.6800562

De Groot, P., & Nott, R. W. (2003). Response of Monochamus (Col., Cerambycidae) and some Buprestidae to flight intercept traps. *Journal of Applied Entomology*, 127(9–10), 548–552. https://doi.org/10.1046/j.1439-0418.2003.00799.x

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biology Letters*, *10*(9), 20140562. https://doi.org/10.1098/rsbl.2014.0562

Dhami, M. K., & Kumarasinghe, L. (2014). A HRM real-time PCR assay for rapid and specific identification of the emerging pest spotted-wing Drosophila (Drosophila suzukii). *PLoS ONE*, 9(6). https://doi.org/10.1371/journal.pone.0098934

Dos Santos, L. A., Mendes, M. F., Krüger, A. P., Blauth, M. L., Gottschalk, M. S., & Garcia, F. R. M. (2017). Global potential distribution of Drosophila suzukii (Diptera, Drosophilidae). *PLoS ONE*, *12*(3), 1–13. https://doi.org/10.1371/journal.pone.0174318

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*(9), 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Edgar, R. C., & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, *31*(21), 3476–3482. https://doi.org/10.1093/bioinformatics/btv401

Elbrecht, V., & Leese, F. (2017). Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. *Frontiers in Environmental Science*, 5, 11. https://doi.org/10.3389/fenvs.2017.00011

Elbrecht, V., & Steinke, D. (2019). Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring. *Freshwater Biology*, 64, 380–387. https://doi.org/10.1111/fwb.13220

Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., & Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, 8(10), 1265–1275. https://doi.org/10.1111/2041-210X.12789

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., Rayé, G., & Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, 15(3), 543–556. https://doi.org/10.1111/1755-0998.12338

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.

Galan, M., Pons, J. B., Tournayre, O., Pierre, É., Leuchtmann, M., Pontier, D., & Charbonnel, N. (2018). Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. *Molecular Ecology Resources*, 18(3), 474–489. https://doi.org/10.1111/1755-0998.12749

Goodhue, R. E., Bolda, M., Farnsworth, D., Williams, J. C., & Zalom, F. G. (2011). Spotted wing drosophila infestation of California strawberries and raspberries: Economic analysis of potential revenue losses and control costs. *Pest Management Science*, 67(11), 1396–1402. https://doi.org/10.1002/ps.2259

Hamby, K. A., Bolda, M. P., Sheehan, M. E., & Zalom, F. G. (2014). Seasonal monitoring for drosophila suzukii (Diptera: Drosophilidae) in California commercial raspberries. *Environmental Entomology*, 43(4), 1008–1018. https://doi.org/10.1603/EN13245

Hardulak, L. A., Morinière, J., Hausmann, A., Hendrich, L., Schmidt, S., Doczkal, D., Müller, J., Hebert, P. D. N., & Haszprunar, G. (2020). DNA metabarcoding for biodiversity monitoring in a national park: screening for invasive and pest species. *Molecular Ecology Resources*, 20, 1542– 1557. https://doi.org/10.1111/1755-0998.13212

Harris, D. W., Hamby, K. A., Wilson, H. E., & Zalom, F. G. (2014). Seasonal monitoring of Drosophila suzukii (Diptera: Drosophilidae) in a mixed fruit production system. *Journal of Asia-Pacific Entomology*, 17(4), 857–864. https://doi.org/10.1016/j.aspen.2014.08.006

Hauser, M. (2011). A historic account of the invasion of Drosophila suzukii (Matsumura) (Diptera: Drosophilidae) in the continental United States, with remarks on their identification. *Pest Management Science*, 67(11), 1352–1357. https://doi.org/10.1002/ps.2265

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de La Société Vaudoise Des Sciences Naturelles.*

Jara, C., Mateo, E., Guillamón, J. M., Torija, M. J., & Mas, A. (2008). Analysis of several methods for the extraction of high quality DNA from acetic acid bacteria in wine and vinegar for characterization by PCR-based methods. *International Journal of Food Microbiology*, 128(2), 336–341. https://doi.org/10.1016/j.ijfoodmicro.2008.09.008

Jarrad, F. C., Barrett, S., Murray, J., Stoklosa, R., Whittle, P., & Mengersen, K. (2011). Ecological aspects of biosecurity surveillance design for the detection of multiple invasive animal species. *Biological Invasions*, 13(4), 803–818. https://doi.org/10.1007/s10530-010-9870-0

Ji, Y., Baker, C. C. M., Li, Y., Popescu, V. D., Wang, Z., Wang, J., Wang, L., Wu, C., Hua, C., Yang, Z., Yang, C., Xu, C. C. Y., Wen, Q., Pierce, N. E., & Yu, D. W. (2020). Large-scale quantification of vertebrate biodiversity in Ailaoshan nature reserve from leech iDNA. *BioRxiv*. https://doi.org/10.1101/2020.02.10.941336

Kanzawa, T. (1939). *Studies on Drosophila suzukii Mats.*

Kim, S. S., Tripodi, A. D., Johnson, D. T., & Szalanski, A. L. (2014). Molecular Diagnostics of Drosophila suzukii (Diptera : Drosophilidae) using PCR-RFLP. *Journal of Economic Entomology*, 107(3), 1292–1294. https://doi.org/10.1603/ec13389

Kim, Y. H., Hur, J. H., Lee, G. S., Choi, M. Y., & Koh, Y. H. (2016). Rapid and highly accurate detection of Drosophila suzukii, spotted wing Drosophila (Diptera: Drosophilidae) by loop-mediated isothermal amplification assays. *Journal of Asia-Pacific Entomology*, 19(4), 1211–1216. https://doi.org/10.1016/j.aspen.2016.10.015

Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* https://www.tidymodels.org

Kwadha, C. A., Okwaro, L. A., Kleman, I., Rehermann, G., Revadi, S., Ndlela, S., Khamis, F. M., Nderitu, P. W., Kasina, M., George, M. K., Kithusi, G. G., Mohamed, S. A., Lattorff, H. M. G., & Becher, P. G. (2021). Detection of the spotted wing drosophila, Drosophila suzukii, in continental sub-Saharan Africa. *Journal of Pest Science*, 94, 251–259. https://doi.org/10.1007/s10340-021-01330-1

Landolt, P. J., Adams, T., Davis, T. S., & Rogg, H. (2012). Spotted Wing Drosophila, Drosophila suzukii (Diptera: Drosophilidae), trapped with combinations of wines and vinegars. *Florida Entomologist*, 95(2), 326–332. https://doi.org/10.1653/024.095.0213

Larson, N. R., Strickland, J., Shields, V. D. C., & Zhang, A. (2020). Controlled-Release Dispenser and Dry Trap Developments for Drosophila suzukii Detection. *Frontiers in Ecology and Evolution*, 8, 45. https://doi.org/10.3389/fevo.2020.00045

Lee, J. C., Burrack, H. J., Barrantes, L. D., Beers, E. H., Dreves, A. J., Hamby, K. A., Haviland, D. R., Isaacs, R., Richardson, T. A., Shearer, P. W., Stanley, C. A., Walsh, D. B., Walton, V. M., Zalom, F. G., & Bruck, D. J. (2012). Evaluation of monitoring traps for Drosophila suzukii (Diptera: Drosophilidae) in North America. *Journal of Economic Entomology*, 105(4), 1350–1357. https://doi.org/10.1603/EC12132

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

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

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

Low-Choy, S. (2015). Getting the Story Straight: Laying the Foundations for Statistical Evaluation of the Performance of Surveillance. In F. Jarrad, S. Low-Choy, & K. Mengersen (Eds.), *Biosecurity Surveillance. Quantitative approaches* (6th ed., pp. 43–73). CABI.

Lozupone, C., & Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12), 8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005

Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D., & Dangl, J. L. (2013). Practical innovations for high-throughput amplicon sequencing. *Nature Methods*, 10(10), 999–1002. https://doi.org/10.1038/nmeth.2634

Lynegaard, G. K., Offenberg, J., Fast, T. S., Axelsen, J. A., Mwatawala, M. W., & Rwegasira, G. M. (2014). Using insect traps to increase weaver ant (Oecophylla longinoda) prey capture. *Journal of Applied Entomology*, 138(7), 539–546. https://doi.org/10.1111/jen.12108

MacConaill, L. E., Burns, R. T., Nag, A., Coleman, H. A., Slevin, M. K., Giorda, K., Light, M., Lai, K., Jarosz, M., McNeill, M. S., Ducar, M. D., Meyerson, M., & Thorner, A. R. (2018). Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics*, 19, 30. https://doi.org/10.1186/s12864-017-4428-5

Maino, J. L., Schouten, R., & Umina, P. (2020). Predicting the global invasion of Drosophila suzukii to improve Australian Biosecurity Preparedness. *Journal of Applied Ecology*, 00, 1–12. https://doi.org/10.1111/1365-2664.13812

Marquina, D., Andersson, A. F., & Ronquist, F. (2019). New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. *Molecular Ecology Resources*, 19(1), 90–104. https://doi.org/10.1111/1755-0998.12942

Marquina, D., Esparza-Salas, R., Roslin, T., & Ronquist, F. (2019). Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. *Molecular Ecology Resources*, 19(6), 1516–1530. https://doi.org/10.1111/1755-0998.13071

Martín, A., Etxebeste, I., Pérez, G., Álvarez, G., Sánchez, E., & Pajares, J. (2013). Modified pheromone traps help reduce bycatch of bark-beetle natural enemies. *Agricultural and Forest Entomology*, 15(1), 86–97. https://doi.org/10.1111/j.1461-9563.2012.00594.x

Martins, F. M. S., Galhardo, M., Filipe, A. F., Teixeira, A., Pinheiro, P., Paupério, J., Alves, P. C., & Beja, P. (2019). Have the cake and eat it: Optimizing nondestructive DNA metabarcoding of macroinvertebrate samples for freshwater biomonitoring. *Molecular Ecology Resources*, 19(4), 863–876. https://doi.org/10.1111/1755-0998.13012

Martoni, F., Nogarotto, E., Piper, A. M., Mann, R., Valenzuela, I., Eow, L., Rako, L., Rodoni, B. C., & Blacket, M. J. (2021). Propylene Glycol and Non-Destructive DNA Extractions Enable Preservation and Isolation of Insect and Hosted Bacterial DNA. *Agriculture*, 11(1), 77. https://doi.org/10.3390/agriculture11010077

Mata, V. A., Rebelo, H., Amorim, F., Mccracken, G. F., Jarman, S., & Beja, P. (2019). How much is enough? Effects of technical and biological replication on metabarcoding dietary analysis. *Molecular Ecology*, 28, 165–175. https://doi.org/10.1111/mec.14779

Mazzetto, F., Pansa, M. G., Ingegno, B. L., Tavella, L., & Alma, A. (2015). Monitoring of the exotic fly Drosophila suzukii in stone, pome and soft fruit orchards in NW Italy. *Journal of Asia-Pacific Entomology*, 18(2), 321–329. https://doi.org/10.1016/j.aspen.2015.04.001

McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing measurements. *ELife*, 8, e46923. https://doi.org/10.7554/eLife.46923

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, 8(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, 6(1), 140. https://doi.org/10.1186/s40168-018-0521-5

Nakamura, S., Tamura, S., Taki, H., & Shoda-Kagaya, E. (2020). Propylene glycol: a promising preservative for insects, comparable to ethanol, from trapping to DNA analysis. *Entomologia Experimentalis et Applicata*, 168(2), 158–165. https://doi.org/10.1111/eea.12876

Nielsen, M., Gilbert, M. T. P., Pape, T., & Bohmann, K. (2019). A simplified DNA extraction protocol for unsorted bulk arthropod samples that maintains exoskeletal integrity. *Environmental DNA*, 1(2), 144–154. https://doi.org/10.1002/edn3.16

Palmer, J. M., Jusino, M. A., Banik, M. T., & Lindner, D. L. (2017). Non-biological synthetic spike-in controls and the AMPtk software pipeline improve fungal high throughput amplicon sequencing data. *PeerJ*, 213470. https://doi.org/10.1101/213470

Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *GigaScience*, 8(8), giz092. https://doi.org/10.1093/gigascience/giz092

Piper, A. M., Cogan, N. O. I., Cunningham, J. P., & Blacket, M. J. (2021). Computational Evaluation of DNA Metabarcoding for Universal Diagnostics of Invasive Insect Pests. *BioRxiv*. https://doi.org/10.1101/2021.03.16.435710

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26(7), 1641–1650. https://doi.org/10.1093/molbev/msp077

Quinn, T. P., Le, V., & Cardilini, A. P. A. (2021). Test set verification is an essential step in model building. *Methods in Ecology and Evolution*, 12, 127–129. https://doi.org/10.1111/2041-210X.13495

R Core Team. (2019). *R: A Language and Environment for Statistical Computing.* https://www.r-project.org/

Reed, M. R. (1938). The Olfactory Reactions of Drosophila Melanogaster Meigen to the products of Fermenting Banana. *Physiological Zoology*, 11(3), 317–325. https://doi.org/10.1086/physzool.11.3.30151465

Rennstam Rubbmark, O., Sint, D., Horngacher, N., & Traugott, M. (2018). A broadly applicable COI primer pair and an efficient single-tube amplicon library preparation protocol for metabarcoding. *Ecology and Evolution*, 8(24), 12335–12350. https://doi.org/10.1002/ece3.4520

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. https://doi.org/10.1016/j.ympev.2006.12.005

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology and Evolution*, 3(3), 430–439. https://doi.org/10.1038/s41559-018-0793-y

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015a). Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, 15(6), 1289–1303. https://doi.org/10.1111/1755-0998.12402

Schnell, I. B., Sollmann, R., Calvignac-Spencer, S., Siddall, M. E., Yu, D. W., Wilting, A., & Gilbert, M. T. P. (2015b). iDNA from terrestrial haematophagous leeches as a wildlife surveying and monitoring tool - prospects, pitfalls and avenues to be developed. Frontiers in Zoology, 12, 24. https://doi.org/10.1186/s12983-015-0115-z

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

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. https://doi.org/10.1139/cjfas-2015-0262

Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A., & Hajibabaei, M. (2019). Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *Scientific Reports*, 9, 5991. https://doi.org/10.1038/s41598-019-42455-9

Smith, D. P., & Peay, K. G. (2014). Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS ONE*, 9(2), e90234. https://doi.org/10.1371/journal.pone.0090234

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, 17(2), 57–86.

Tedersoo, L., Drenkhan, R., Anslan, S., Morales-Rodriguez, C., & Cleary, M. (2019). High-throughput identification and diagnostics of pathogens and pests: overview and practical recommendations. *Molecular Ecology Resources*, 19, 47–76. https://doi.org/10.1111/1755-0998.12959

Tonina, L., Grassi, A., Caruso, S., Mori, N., Gottardello, A., Anfora, G., Giomi, F., Vaccari, G., & Ioriatti, C. (2018). Comparison of attractants for monitoring Drosophila suzukii in sweet cherry orchards in Italy. *Journal of Applied Entomology*, 142, 18–25. https://doi.org/10.1111/jen.12416

Vamos, E. E., Elbrecht, V., & Leese, F. (2017). Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcoding and Metagenomics*, 1, e14625. https://doi.org/10.3897/mbmg.1.14625

Van Timmeren, S., Diepenbrock, L. M., Bertone, M. A., Burrack, H. J., & Isaacs, R. (2017). A Filter Method for Improved Monitoring of Drosophila suzukii (Diptera:

Drosophilidae) Larvae in Fruit. *Journal of Integrated Pest Management*, 8(1), 23. https://doi.org/10.1093/jipm/pmx019

Walsh, D. B., Bolda, M. P., Goodhue, R. E., Dreves, A. J., Lee, J., Bruck, D. J., Walton, V. M., O'Neal, S. D., & Zalom, F. G. (2011). Drosophila suzukii (Diptera: Drosophilidae): Invasive Pest of Ripening Soft Fruit Expanding its Geographic Range and Damage Potential. *Journal of Integrated Pest Management*, 2(1), G1–G7. https://doi.org/10.1603/ipm10010

Wang, Y., Naumann, U., Wright, S. T., & Warton, D. I. (2012). Mvabund- an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, 3(3), 471–474. https://doi.org/10.1111/j.2041-210X.2012.00190.x

Whittle, P. J. L., Stoklosa, R., Barrett, S., Jarrad, F. C., Majer, J. D., Martin, P. A. J., & Mengersen, K. (2013). A method for designing complex biosecurity surveillance systems: Detecting non-indigenous species of invertebrates on Barrow Island. *Diversity and Distributions*, 19, 629–639. https://doi.org/10.1111/ddi.12056

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wilcox, T. M., Zarn, K. E., Piggott, M. P., Young, M. K., McKelvey, K. S., & Schwartz, M. K. (2018). Capture enrichment of aquatic environmental DNA: A first proof of concept. *Molecular Ecology Resources*, 18(6), 1392–1401. https://doi.org/10.1111/1755-0998.12928

Willis, A. D. (2019). Rarefaction, alpha diversity, and statistics. *Frontiers in Microbiology*, 10, 2407. https://doi.org/10.3389/fmicb.2019.02407

Willis, A. D. & Bunge, J. (2015). Estimating diversity via frequency ratios. *Biometrics*, 71(4), 1042–1049. https://doi.org/10.1111/biom.12332

Wilkinson, S. (2019). aphid: an R package for analysis with profile hidden Markov models. *Bioinformatics*, 35(19), 3829–3830. https://doi.org/10.1093/bioinformatics/btz159

Yang, C., Bohmann, K., Wang, X., Cai, W., Wales, N., Ding, Z., Gopalakrishnan, S., & Yu, D. W. (2021). Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. *Methods in Ecology and Evolution*, 12, 1252–1264. https://doi.org/10.1111/2041-210X.13602

Yu, G., Lam, T. T. Y., Zhu, H., & Guan, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular Biology and Evolution*, 35(12), 3041–3043. https://doi.org/10.1093/molbev/msy194

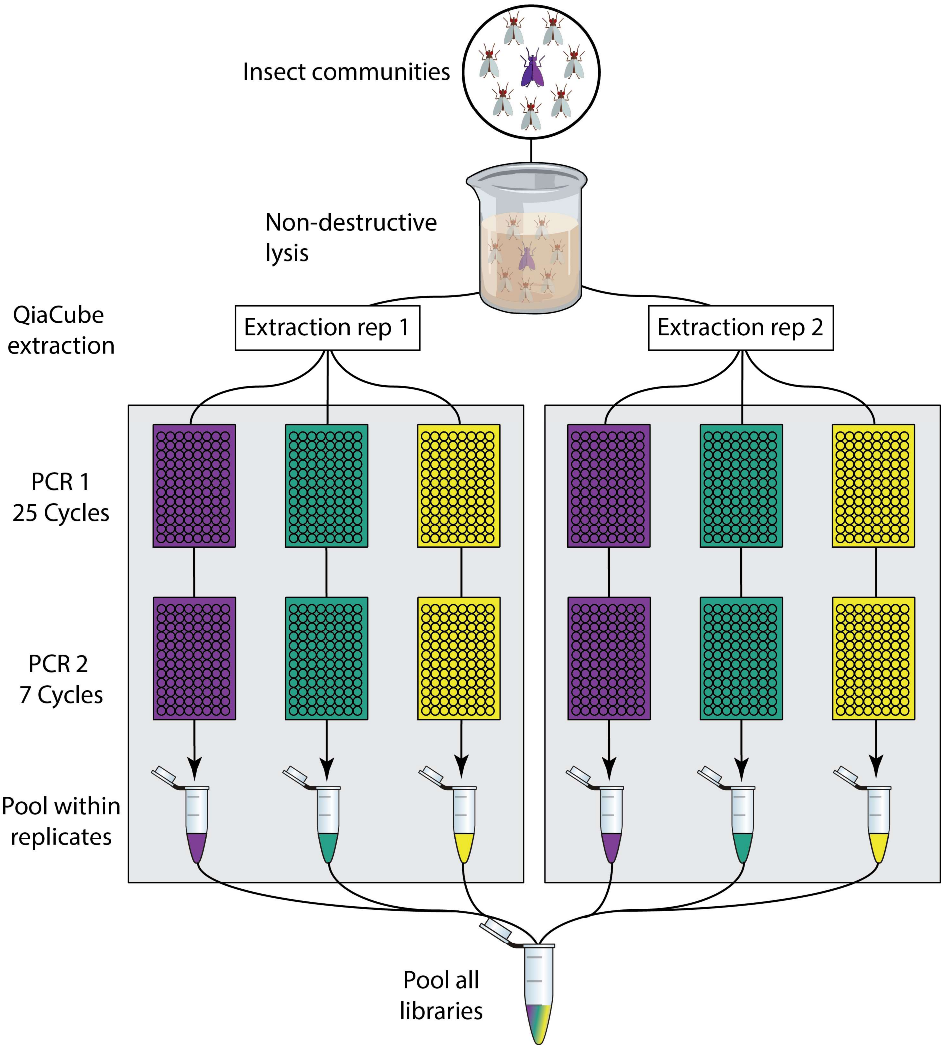
Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods in Ecology and Evolution*, 8(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Zhou, J., Jiang, Y. H., Deng, Y., Shi, Z., Zhou, B. Y., Xue, K., Wu, L., He, Z., & Yang, Y. (2013). Random sampling process leads to overestimation of β-diversity of microbial communities. *MBio*, 4(3), e00324-13. https://doi.org/10.1128/mBio.00324-13

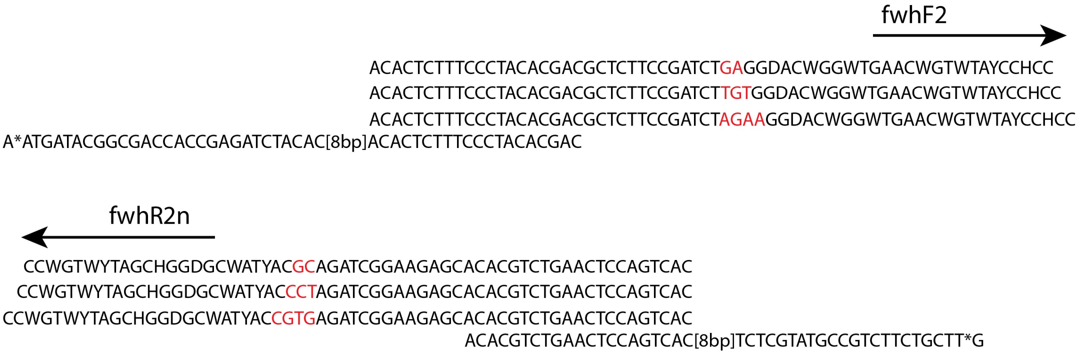
Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., … Taberlet, P. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, 28(8), 1857–1862. https://doi.org/10.1111/mec.15060

Zizka, V. M. A., Leese, F., Peinert, B., & Geiger, M. F. (2019). DNA metabarcoding from sample fixative as a quick and voucher-preserving biodiversity assessment method. *Genome*, 62(3), 122–136. https://doi.org/10.1139/gen-2018-0048

4.5 Supplementary Information



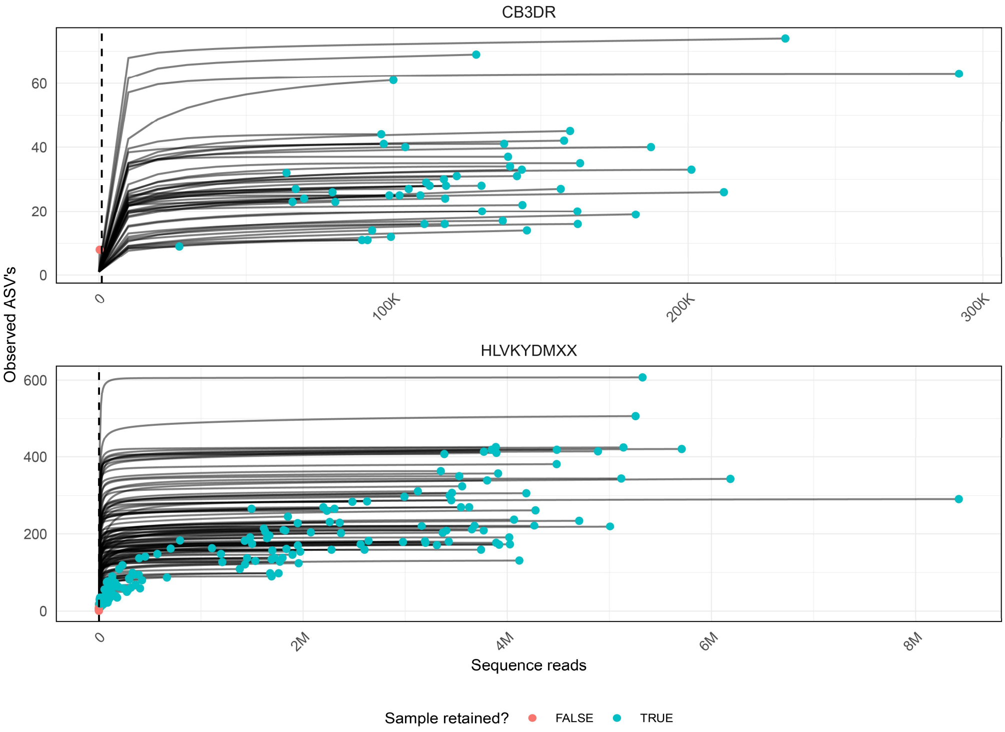
Supplementary Figure 1: ‘Twin-tagging’ replication strategy used to prepare metabarcoding libraries for this study.



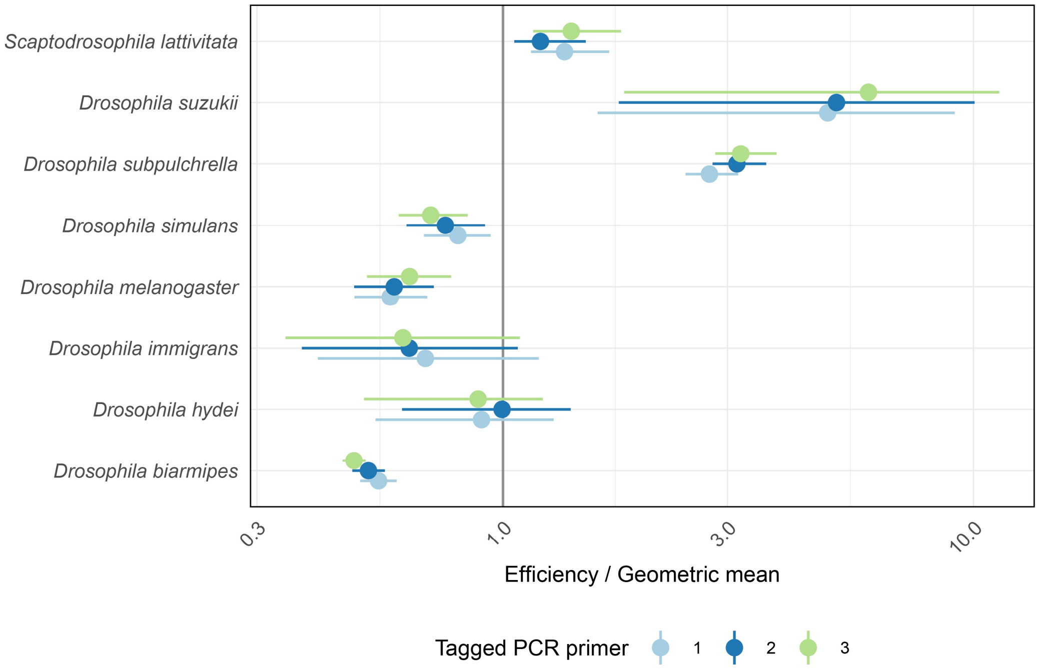
```
                                   ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGGDACWGGWTGAACWGTWTAYCCHCC
                                   ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGTGGDACWGGWTGAACWGTWTAYCCHCC
                                   ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGAAGGDACWGGWTGAACWGTWTAYCCHCC
A*ATGATACGGCGACCACCGAGATCTACAC[8bp]ACACTCTTTCCCTACACGAC

  CCWGTWYTAGCHGGDGCWATYACGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
 CCWGTWYTAGCHGGDGCWATYACCCTAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
CCWGTWYTAGCHGGDGCWATYACCGTGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
                                        ACACGTCTGAACTCCAGTCAC[8bp]TCTCGTATGCCGTCTTCTGCTT*G
```

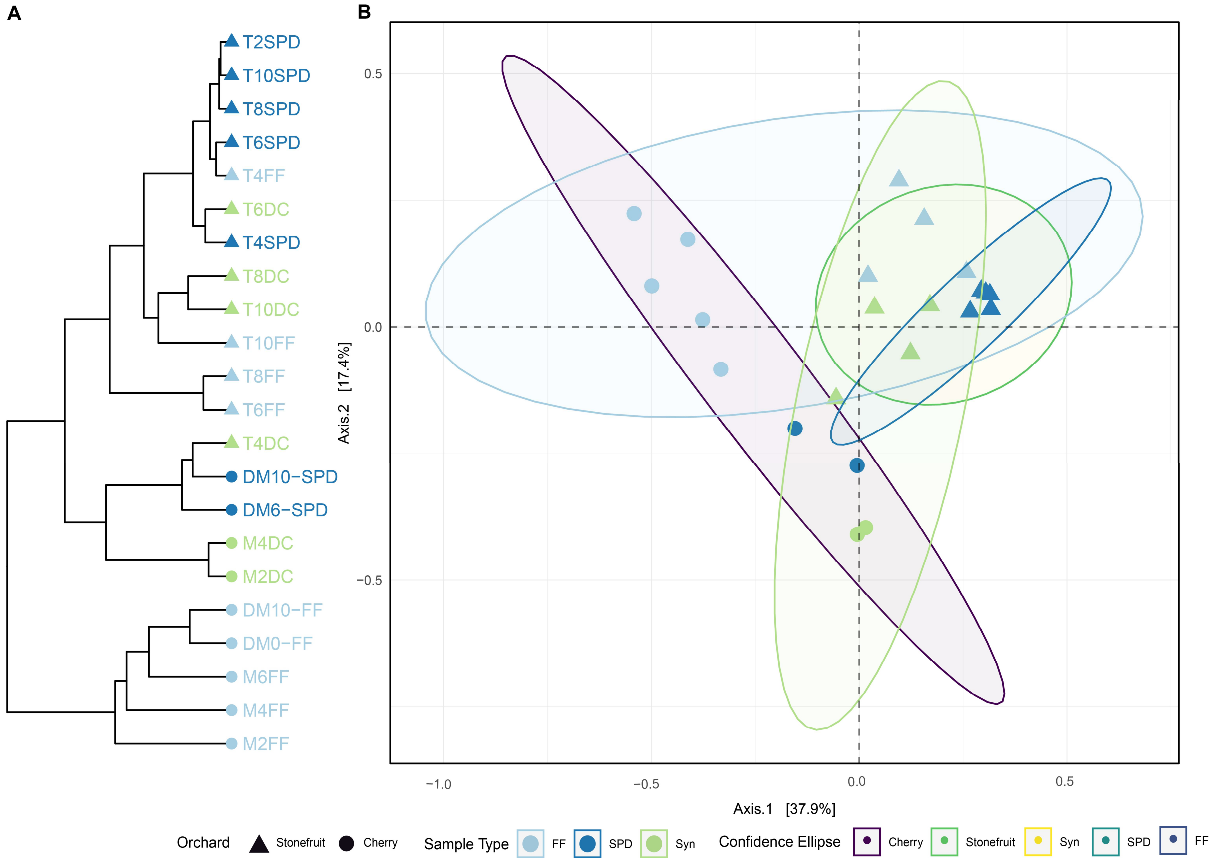
Supplementary Figure 2: Design of in-line tagged PCR primers to facilitate the twin-tagging strategy. Bases highlighted in red represent the inline tags, which incorporate length variation to improve phasing during the critical first cycles of the sequencing process.



Supplementary Figure 3: Rarefaction curves displaying the accumulation of unique ASVs within each sample as sequencing depth increases, displayed on a pseudo-log scale. All samples that received <1000 sequence reads were removed from subsequent analyses.



Supplementary Figure 4: Quantitative performance of the three ‘tagged’ fwhF2-fwhR2n primer sets across all mock community members.



Supplemetary Figure 5: A) Hierarchial clustering of weightedUniFrac distances between samples using the ward.D2 method. **B)** Principal coordinate analysis of weighted UniFrac distances between samples with 95% confidence ellipses drawn using a multivariate t-distribution.

Supplementary Note 1: *Design of synthetic positive controls*

To design synthetic positive controls, alignments of 13 major insect families; *Drosophilidae*, *Tephritidae*, *Culicidae*, *Crambidae*, *Tortricidae*, *Apidae*, *Siricidae*, *Aphididae*, *Triozidae*, *Cerambycidae*, *Nitidulidae*, *Thripidae* and *Acrididae* were extracted from the curated COI reference database of Piper et al., (2021). Profile Hidden Markov Models (Eddy, 1998) were then derived separately for each family using the aphid R package (Wilkinson, 2019), and novel 658 bp sequences generated from the per-site nucleotide base probabilities described by each profile. Each synthetic sequence was checked for absence of stop codons and verified via a BLAST search to be >8% diverged from any sequence on GenBank for both the full 658 bp sequence and ~220 bp subregion amplified by the metabarcoding primers evaluated in this study. To increase the GC content and allow further differentiation from biological sequences, the letters PAC for ‘Positive Amplification Control’ was spelt in amino acids (CCT GCC TGC) at each end of the synthetic sequence, and then synthesised as gBlocks fragments (Integrated DNA

Technologies, USA). The final 676 bp synthetic sequences were equimolarly pooled to form a mixed template positive control sample included as a separate library in all sequencing runs.

Supplementary Table 1: Composition of all mock communities used in this study. * Target exotic species, not present in Australia

	Drosophila melanogaster	***Drosophila simulans***	***Drosophila hydei***	***Scaptodrosophila lattivittata***	***Drosophila Immigrans***	***Drosophila subpulchrella ****	***Drosophila suzukii ****	***Drosophila biarmipes ****	**Total individuals**
D100M1	30	50	10	10	0	0	0	0	100
D100M2	40	9	30	10	10	1	0	0	100
D100M3	14	10	40	30	0	1	0	1	96
D100M4	40	10	0	39	5	1	5	0	100
D100M5	10	55	14	0	20	0	1	1	101
D250M1	75	125	25	25	0	0	0	0	250
D250M2	100	25	75	25	24	1	0	0	250
D250M3	37	25	100	82	0	1	5	1	251
D250M4	100	25	0	104	15	1	5	0	250
D250M5	25	145	29	0	50	0	1	1	251
D500M1	150	250	50	50	0	0	0	0	500
D500M2	200	50	150	50	49	1	0	0	500
D500M3	70	50	220	154	0	1	5	1	501
D500M4	210	50	0	210	24	1	5	0	500
D500M5	54	275	70	0	100	0	1	1	501
D1000M1	300	500	100	100	0	0	0	0	1000
D1000M2	400	100	300	100	99	1	0	0	1000
D1000M3	194	100	400	300	0	1	5	1	1001
D1000M4	444	100	0	400	50	1	5	0	1000
D1000M5	100	550	149	0	200	0	1	1	1001
DLarv1	75	125	49	0	0	0	1	0	250
DLarv2	100	25	100	25	0	0	0	0	250
DLarv3	40	29	100	80	0	0	1	0	250
DLarv4	100	50	0	100	0	0	0	0	250
DLarv5	0	190	19	40	0	0	1	0	250

5
Quantification of Insect Pests Within Mixed Trap Samples Using a Bias-Corrected Metabarcoding Assay

5.1 Chapter preface:

The preceding chapters focussed on the qualitative application of metabarcoding to early detection of invasive insects, yet for other biosurveillance activities, such as population monitoring to support pest eradication or suppression efforts, accurate quantitative measurements of species abundance are required. In order to refine the quantitative performance of the metabarcoding assay developed in chapters 3 and 4, this chapter evaluates the use of statistical and machine learning models to actively correct for taxonomic bias during data analysis. This bias-correction approach is then validated on the case study of pheromone trapped Carpophilus beetles, endemic pests of almonds and stone fruit in Australia. This chapter is presented as a self-contained manuscript in the final stages of preparation, with intended submission to the journal *Pest Management Science*, and includes supplementary material at the end.

5.2 Publication details:

Quantification of insect pests within mixed trap samples using a bias-corrected metabarcoding assay

Stage of publication: In Preparation

Journal details: Pest Management Science

Authors: Alexander M. Piper, Lea Rako, Linda Semeraro, Noel O.I. Cogan, Mark J. Blacket, John Paul Cunningham.

5.3 Statement of joint authorship:

A.M.P, J.P.C. and M.J.B. conceptualised the study, A.M.P. performed all molecular laboratory procedures, bioinformatic and statistical analyses. L.R. performed all morphological identification of trap samples. L.R. and L.S. generated the Carpophilus reference sequences used for identification. A.M.P. wrote the first draft of the manuscript with input and supervision from J.P.C., N.O.I.C., and M.J.B. All authors contributed to the editing of the final manuscript and approved the version presented here.

Statement from co-author confirming the contribution of the PhD candidate:

“As co-author of the manuscript ‘Piper, A. M., Rako, L., Semeraro, L., Cogan N.O.I., Blacket M.J. & Cunningham, J. P. (In preparation). Quantification of insect pests within mixed trap samples using a bias-corrected metabarcoding assay, *Pest Management Science*’, I confirm that Alexander M. Piper has made the contributions listed above.”

Associate Professor John Paul Cunningham

30/03/2021

5.4 Manuscript

Quantification of insect pests within mixed trap samples using a bias-corrected metabarcoding assay

Alexander M. Piper[1,2], Lea Rako[1], Linda Semeraro[1], Noel O.I. Cogan[1,2], Mark J. Blacket[1], John Paul Cunningham[1,2]

[1] Agriculture Victoria Research, AgriBio, Centre for AgriBioscience, Bundoora, Victoria, 3083, Australia.

[2] School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, 3086, Australia.

Keywords: DNA barcoding, sap beetles, biosurveillance, biomonitoring, agriculture, non-destructive extraction, NGS sequencing

Running title: Correcting metabarcoding bias for insect population monitoring

Corresponding author:

Alexander M. Piper

Email: alexander.piper@agriculture.vic.gov.au

Abstract

Monitoring of pest populations forms a cornerstone of integrated pest management (IPM) programmes, informing the timely application of control measures before widespread crop damage can occur. While monitoring traps aim to be species-specific for the target pest, this is not always possible, and the use of traps with a lower specificity can require extensive specimen sorting. DNA metabarcoding offers a high-throughput molecular method for simultaneously identifying multiple species within unsorted trap samples, but taxonomic bias currently limits its ability to provide quantitative data. Here we compare six statistical and machine learning models for their ability to accurately estimate and correct for taxonomic bias in a metabarcoding assay targeting mixed communities of horticultural pest *Carpophilus* beetles (Coleoptera: Nitidulidae). All six models substantially improved the concordance between expected and observed relative abundances, and the bias-corrected relative abundances could be translated back into counts of specimens using an independent measurement of the sample. However, none of the models, nor data transformations, could reduce the root mean squared error below 11%, regardless of the number of samples used for model training. While taxonomic bias was found to act consistently between DNA extraction replicates, PCR replicates, and samples; intra-specimen variability in morphological traits likely imposes this limit on bias correctability. Despite this, bias-calibrated and absolute abundance adjusted metabarcoding datasets are comparable to those derived from traditional morphological identification, while requiring substantially less time and personnel to obtain. Bias-calibrated metabarcoding may therefore provide an approach for significantly scaling up the identification and quantification of trapped specimens collected through IPM population monitoring activities.

Introduction

Monitoring of pest populations is a foundation of integrated pest management (IPM) programmes, allowing precise and targeted control measures to be applied before widespread crop damage can occur (Gray et al., 2008; Kogan, 1988). Monitoring traps baited with synthetic pheromone lures enable selective capture of target pest species (Witzgall et al., 2010), but where a species-specific pheromone is not available a more general lure must be used, for example based on host plant semiochemicals (Cha et al., 2014; Cunningham et al., 2016) or pheromone blends that attract several related species (Hossain et al., 2006; Tan et al., 2014). Use of these more broadly tuned lures can lead to substantial bycatch that must be sorted through to locate and quantify the target pests (Digirolomo & Dodds, 2014; Spears et al., 2016; Weber & Ferro, 1991), a process which can be further complicated by closely related non-pest species that share morphological features with the pest taxa (Blacket et al., 2012; Hossain & Bartelt, 2010).

DNA metabarcoding is a recently developed high-throughput sequencing (HTS) based assay for simultaneously identifying multiple species that can be applied directly to unsorted trap catches (Liu et al., 2020; Piper et al., 2019). To date, use of metabarcoding in pest management has centred upon its qualitative application within early detection surveillance (Brown et al., 2016; Piper et al., 2019), yet metabarcoding also holds the potential for measuring species abundance within the sample; if taxon-specific quantitative biases can be overcome (Deagle et al., 2019; Lamb et al., 2019). Taxonomic bias in metabarcoding assays primarily arises through mismatch between PCR primers and the various template molecules released by a mixed community (Clarke et al., 2014; Piñol et al., 2015). This presents a particular problem for the widely adopted mitochondrial cytochrome c oxidase I (COI) barcoding gene where there are no strictly conserved nucleotide sites for design of universal primers (Deagle et al., 2014). While inclusion of degenerate nucleotide bases within primers can account for template variation between targets (Elbrecht et al., 2019; Piñol et al., 2019), any other trait that alters the number of molecules released by a specimen, such as its biomass (Elbrecht et al., 2017), exoskeleton hardness (Marquina et al., 2019), or mitochondrial copy number (Krehenwinkel et al., 2017; Wilcox et al., 2018), will further bias the results.

While continued efforts to optimise primers and protocols will no doubt prove important for refining the quantitative performance of metabarcoding, a less explored but complementary approach is the use of statistical models to actively correct for taxonomic bias during analysis. Metabarcoding bias can be modelled as a multiplicative effect, where each consecutive step of the laboratory and bioinformatic protocol distorts the starting abundances by a taxon-specific multiplicative factor (McLaren et al., 2019). Under this model, any bias introduced throughout the entire protocol should be ameliorable by simply dividing the final abundances of each taxon by an appropriate correction factor (Krehenwinkel et al., 2017; McLaren et al., 2019). These correction factors can be obtained by measuring the deviation between the expected and observed abundances in morphologically identified, or artificially assembled 'mock' communities, and then used to calibrate further samples (Krehenwinkel et al., 2017). Stochastic variation or 'pipeline noise' introduced throughout the laboratory process can, however, impact the accuracy of these measurements, propagating error into the final calibrated results. In light of this, the most effective modelling approach and minimum number of observations required to accurately capture the taxonomic bias must be determined before bias-calibration can be integrated into metabarcoding analysis pipelines.

Deriving correction factors from previously identified communities can be framed as a predictive modelling problem, for which the field of supervised machine learning provides a collection of eminently suitable techniques (Crisci et al., 2012; Lucas, 2020). However, the datasets generated by HTS platforms have nuances that can compromise the inference and interpretation of predictive models if not appropriately considered (Quinn et al., 2018). Metabarcoding and other HTS assays provide compositional data (sometimes called relative, or proportions data), where the sequence read counts returned for each taxon are conditionally dependent on the counts of all other taxa within the sample (Gloor et al., 2017; Quinn et al., 2019). Therefore, if the representation of molecules from one taxon increases due to taxonomic bias the measured counts of other taxa will appear to decrease, even if their absolute abundances in the original sample remain unchanged. When analyses fail to take this compositionality into account it may appear that bias does not consistently act across samples, limiting its correctability (McLaren et al., 2019). Compositional Data Analysis (CODA) approaches deal with this by transforming the raw variables into a set of log-ratios in which the denominator is an internal reference within each sample, such as a specific control taxon or the per-sample geometric mean (Quinn et al., 2019). By framing the analysis in this way, log-ratio transformations map compositional data into conventional Euclidean geometry, thereby enabling the use of many statistical methods without violating underlying assumptions (Aitchison, 1982). Alternatively, compositional constraints can be removed by adjusting samples back to absolute abundances using an independent measurement of the total specimens or biomass in the sample (Harrison et al., 2021). While measuring absolute abundance presents a challenge in itself for the microbiome studies which pioneered this technique (Morton et al., 2019; Props et al., 2017), it may be more tractable for insect metabarcoding where specimens can be seen with the naked eye. Translating metabarcoding-provided relative abundances back to absolute abundances also provides benefits for data interpretation and practical implementation, as many economic thresholds for IPM are assessed using absolute numbers of insects per unit area, per plant, or per part of plant (Ramsden et al., 2017). This means that traditional economic threshold models can be applied directly, without having to be reformulated to account for the compositional data returned by metabarcoding.

In this study, we compare a series of predictive models and data transformations for their ability to estimate and correct for taxonomic bias in metabarcoding assays, and determine whether the relative abundances provided by metabarcoding can be transformed back to absolute counts of insects using independent measurements of each sample. This approach is then applied to metabarcoding based identification of trap-caught *Carpophilus* beetles (Coleoptera: Nitidulidae), a genus containing several economically important pests of fruit and nut crops (Hossain, 2018; Hossain et al., 2006). In Australia, the pheromone baited traps employed in *Carpophilus* surveillance can capture hundreds to thousands of insects from up to 12 different species, and assessing the abundance of each constituent species—only some of which are pests—is limited by the requirement for trained entomological diagnosticians (Hossain, 2018; James et al., 1995). Following selection of the best performing model, we measure how closely the corrected bias reflects the results of morphological sorting to determine whether datasets of interest to IPM population monitoring can be provided by high-throughput metabarcoding assays.

Methods

Samples

15 adult and 6 larval mock communities with total abundances of approximately 300 individuals were assembled from laboratory reared colonies of *Carpophilus davidsoni*, *C. hemipterus*, and *C. truncatus*, supplemented with field collected specimens of *C. nepos*, *C. marginalus*, *Urophorus humeralis*, and *Brachypeplus sp.* (Supplementary Table 1). In addition, 12 field samples were collected from traps baited with a commercial Carpophilus aggregation pheromone and food lure (Hossain et al., 2006), deployed as part of regular monitoring activities in almond orchards located near Sunraysia, Victoria, Australia (Supplementary Table 1). All trapped Nitidulid beetles were morphologically identified using the taxonomic key of Leschen & Marris (2005) and stored dry at 4 °C for 1 year before DNA extraction.

Metabarcoding

DNA was extracted from mock and field collected communities using the non-destructive Qiagen DNeasy based protocol presented in Chapter 4. In brief, ethanol was removed from the insect communities using a 1000 µL pipette and specimens dried overnight to ensure all residual ethanol was evaporated. Dried specimens were suspended in a 10:1 mix of Qiagen ATL lysate buffer and proteinase K in 15 mL falcon tubes, with the total volume of buffer increased proportionally to the size of the insect community to ensure all specimens were fully immersed, then incubated for 24 hours at 56 °C and 220 rpm in a shaking incubator. Following incubation, lysate was removed from the specimens and split into two separate replicate aliquots per community, each of which were manually pipetted into a separate Qiagen 96 well DNeasy extraction plate using a multichannel pipette. The remainder of the Qiagen DNeasy Blood & Tissue protocol was then followed within the QiaCube automated DNA purification workstation (Qiagen, Germany).

Metabarcoding libraries were prepared from the non-destructively extracted DNA using a two-step PCR approach. First, 3 replicated PCRs were used to amplify the COI locus from each DNA extract using the fwhf2-fwhR2n primer pairs (Vamos et al., 2017), modified with 2-4 bp inline tags at the 5'- terminus to differentiate each PCR replicate as per Chapter 4. Each 25 µL reaction consisted of 5 µL 5X MyFi reaction buffer (Bioline, USA), 1

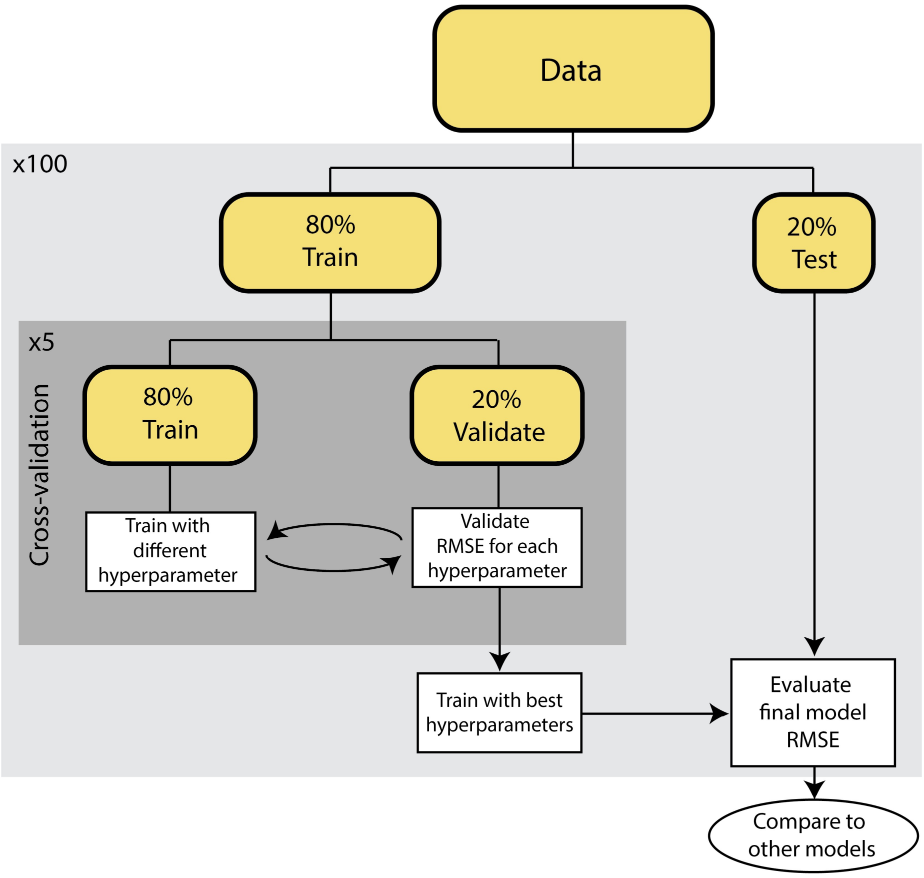

Data
x100
80%
Train
20%
Test
x5
Cross-validation
80%
Train
20%
Validate
Train with different hyperparameter
Validate RMSE for each hyperparameter
Train with best hyperparameters
Evaluate final model RMSE
Compare to other models

Figure 1: Flow diagram of modelling workflow used in this study. First the entire dataset was split into training (80%) and held-out test sets (20%), then the training set was further split into five cross-validation folds, on which models were iteratively trained and evaluated to select the best hyper-parameters. Following selection of optimal hyper-parameters, the finalised models were compared on the original held-out data set. The entire process was repeated across 100 random training/test splits to ensure robustness of comparisons to whichever samples were assigned to each set.

μL of 10 nM forward and reverse primers, 0.8 μL MyFi DNA polymerase, 11.2 μL bovine serum albumin (BSA) and 2 μL of variable concentration template DNA. Cycling conditions were 94 °C for 2 min, 30 cycles of 94 °C for 30 sec, 50 °C for 45 sec, and 72 °C for 45 sec, followed by a final extension step of 2 min at 72 °C, with each set of PCR replicates amplified in a separate thermocycler. Successful amplification was verified on a 2% w/v agarose gel, and amplicons diluted 1:10 in ddH20 with no further clean-up step. 1 μL of the diluted COI amplicons were then amplified again using 7 cycles of real-time PCR in order to attach 8 bp unique dual indexes and Illumina sequencing adapters (Costello et al., 2018). Cycling conditions for the second PCR were 98 °C for 10 sec, 65 °C for 30 sec, and 72 °C for 30 sec, with each cycle followed by a SYBR Green fluorescence read. Melt curve analysis was used to quantify the concentrations of successfully indexed libraries, and these measurements used to pool all libraries in equimolar ratios using a Biomek FX^{P}

liquid handling robot (Beckman Coulter, USA). Pooled libraries were purified using a 0.8:1 ratio of AMPure XP beads (Beckman Coulter, USA) then sized and quantified using a 2200 TapeStation (Agilent Technologies, USA) and Qubit 3.0 Fluorometer (Thermo Fisher, USA). Final pooled and cleaned libraries were either diluted to 7 pM, spiked with 5% PhiX, and sequenced on an Illumina MiSeq, or diluted to 100 pM, spiked with 1% PhiX and sequenced an Illumina NovaSeq6000 S2 flow cell, both using 2 x 150 bp reads. In order to minimise the risk of contamination from the laboratory environment, DNA extraction, preparation of PCR master-mixes, PCR amplification, and library preparation were each performed in separate rooms using dedicated equipment and pipettes.

Bioinformatics analysis

Sequence reads were demultiplexed using *bcl2fastq* allowing for no mismatches to the expected index combinations, followed by a second round of demultiplexing for the inline tags using *Seal* in *BBTools* v38 (Bushnell et al., 2017). Demultiplexed sequencing reads (NCBI SRA acc no: xxxxxxx, to be assigned later) were trimmed of PCR primer sequences using *BBDuK* in *BBTools* v38 and any sequences with >1 expected error (Edgar & Flyvbjerg, 2015), <8 unique 2-mers or any ambiguous 'N' bases were removed. Remaining sequences were denoised with *DADA2* v1.16 (Callahan et al., 2016), using the "pseudo-pooling" mode for increased sensitivity to rare variants, and the error matrix modified to enforce monotonicity in order to deal with the binned quality scores produced by the NovaSeq as per Chapter 4. Following denoising, the Amplicon Sequence Variants (ASVs) inferred separately from each sequencing run were merged into a single table and chimeric sequences removed using the *removeBimeraDenovo* function in DADA2. The remaining ASVs were aligned to a Profile Hidden Markov Model (PHMM) of the COI barcode region (Chapter 3) using the *aphid* R package (Wilkinson, 2019) in order to filter out pseudogenes and non-specific amplification products.

To assign taxonomy to the filtered ASVs, the IDTAXA algorithm (Murali et al., 2018) was trained for 5 iterations on the curated insect reference database generated in Chapter 3 supplemented with 91 additional Australian Nitidulid sequences. Hierarchical taxonomy was assigned to the lowest rank attainable with a minimum 60% bootstrap support, and additional species level assignments obtained using a nucleotide BLAST search (Altschul et al., 1990) against the same reference database. As bias correction models can only be trained for those taxa for which abundance was measured a-priori, all ASVs corresponding to species that were not identified during the initial morphological sorting were removed, and only replicates with >1000 total sequence reads retained. All remaining ASVs were then agglomerated by species, and their associated sequence read counts transformed into per-sample relative abundances.

Table 1: Classes of statistical and machine learning models evaluated for bias estimation and calibration.

Method	**Linearity**	**Tuned hyper-parameters**	**Fixed hyper-parameters**	**References**
metacal	Linear	• N/A	• N/A	McLaren et al., (2019)
Linear regression	Linear	• N/A	• N/A	N/A
LASSO regression	Linear	• *penalty*: Regularization penalty • *mixture*: proportion of L1 regularization	• N/A	Tibshirani, (1996)
Polynomial support vector machine (SVM)	Non-linear	• *cost*: The cost of predicting a sample within or on the wrong side of the margin. • *degree*: The polynomial degree. • *scale_factor*: A scaling factor for the kernel. • *margin*: The epsilon in the SVM insensitive loss function	• N/A	Cortes & Vladimir, (1995)
Random forest	Non-linear	• *mtry*: Number of Predictors randomly sampled at each split • *min_n*: number of data points per node for the node to be split	• *trees*: number of trees contained in the ensemble = 1000	Breiman, (2001)
XGBoost	Non-linear	• *min_n*: number of data points per node for the node to be split • *tree_depth*: maximum depth of the tree (i.e. number of splits) • *learn_rate*: Rate at which the boosting algorithm adapts from iteration-to-iteration.	• *trees*: number of trees contained in the ensemble = 1000	Chen & Guestrin, (2016)

Comparing correction models

Five linear and non-linear predictive models (Table 1) were fit to the expected and observed relative abundances as is, or transformed using the natural log, log-odds (logit), or into absolute abundances by multiplying each by the total number of morphologically counted individuals in the respective sample. Each model was constructed so that the proportions of reads produced by metabarcoding was a function of the proportion of starting individuals, species, sequencing run, and the type of community (mock adult, mock larval, or field collected adult) (Supplementary equation 1). In addition, the same 5 model types were fit to the compositional error, or the ratio between expected and observed abundances (McLaren et al., 2019), transformed to be relative to C. *hemipterus* (additive log-ratio [ALR], Supplementary equation 3), or the per-sample geometric mean (centred log-ratio [CLR], Supplementary equation 4). Again, the taxon, sequencing run, and community type were included as covariates (Supplementary equation 2), and the final calibrated abundances obtained by dividing the observed proportions by the correction factors predicted by the model. Models fit to the compositional error were also compared to metacal, a published model specifically designed to correct for taxonomic bias using a CODA framework (McLaren et al., 2019). To accurately evaluate the predictive performance of a model, it must be tested on data it has not seen before (Quinn et al., 2021; Topçuoğlu et al., 2020). Therefore, a random 80% of the previously identified communities were assigned to a training set on which each model was fit, with the remaining 20% held aside as a test set (Figure 1). Dataset splitting was conducted in a stratified manner in order to maintain similar proportions of mock adult, mock larval, and field collected communities within the training and test sets. Model selection for the machine learning algorithms requires tuning of hyper-parameters, which are those input parameters that need to be specified by the user rather than learned directly from the data (Table 1). To determine the best performing hyper-parameters, the training set was further split 80/20 into five separate cross-validation (CV) folds, again in a stratified manner (Figure 1), then a grid search following a Latin-hypercube design (Sacks et al., 1989) was used to explore the possible hyper-parameter space. Final hyper-parameter values were selected from those that led to the lowest average Root Mean Squared Error (RMSE) across the 5 CV folds, then each model was re-trained on the complete training

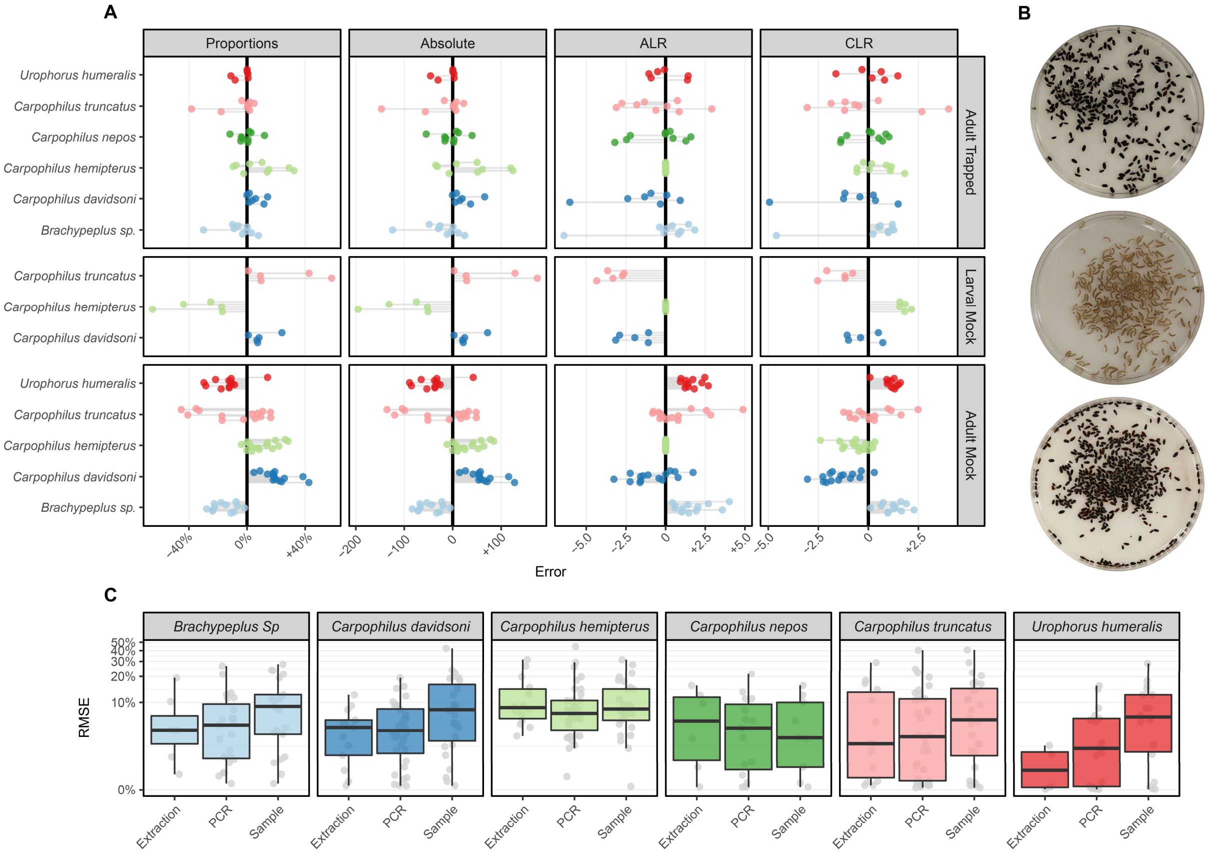


Figure 2: A) Quantitative error between expected and observed abundance, displayed as; relative abundances (proportions), absolute abundances, additive log-ratios relative to C. *hemipterus* (ALR), and centred log-ratios (CLR) for each taxon across the three community types **B)** Photos of morphological species contained within a typical adult trapped, larval mock, and adult mock community **C)** Consistency of taxonomic bias between extraction replicates of the same sample, PCR replicates of the same DNA extraction, and different samples (biological replicates).

set and applied to the withheld test dataset in order evaluate its predictive performance. Dataset splitting, hyper-parameter tuning, final model training, and test set evaluation were repeated 100 times using a different random initiation seed to ensure that comparisons were robust to whichever samples were assigned to the training or test sets (Figure 1). The performance of models across all 100 training/test splits was compared against the uncorrected data, as well as the baseline linear regression model using t-tests with a Benjamini-Hochberg correction. In order to determine the minimum number of training samples required to obtain a good fit for the final model, samples were iteratively removed from the training set and the change in test set RMSE measured for each taxon. All statistical analysis and model training procedures were conducted using the *tidymodels* (Kuhn & Wickham, 2020) and *tidyverse* (Wickham et al., 2019) packages within the R 4.1 statistical programming environment (R Core Team, 2019), and all figures plotted using ggplot2 (Wickham, 2016).

Results

Sequencing results

All mock and field collected communities were sequenced across a portion of a MiSeq V2, and NovaSeq 6000 S2 lane, yielding 1,067,524 and 172,221,696 filtered reads respectively (mean: 213,505 ± 26,187 per sample for MiSeq, mean: 3,131,304 ± 418,767 per replicate for NovaSeq), however, a large number of replicate dropouts were seen across both the mock and trap communities sequenced on the NovaSeq (Supplementary Figure 1). For the mock communities, 26% of the replicates from both the adults and the larval samples did not produce sufficient data to pass quality control steps, while for the field collected communities only 50% were successful. Most replicate dropouts occurred in PCR replicate set 2 of extraction replicate 1, where all replicates failed, and PCR replicate set 1 of extraction replicate 2 where 27 of 36 replicates failed (Supplementary Figure 1). These samples were processed in the same batch as those in Chapter 4 which saw similar replicate dropouts, and likely indicates a systematic failure during PCR amplification or library pooling. Despite the failure of these replicates, all of the adult and larval mock communities and 82% of the trap samples had at least one replicate that was successfully sequenced and could therefore be analysed further. From these samples, a total of 32 unique taxa were identified at above 0.01% relative abundance, substantially higher than the 6 distinct species recorded by morphological sorting. The majority of these additional taxa were from the orders Coleoptera (13 species) and Diptera (7 Species) (Supplementary Figure 2) and represented low abundance specimens that were likely overlooked or left unidentified while searching for the target *Carpophilus* species. As correction factors can only be calculated for taxa which abundance was measured in advance, all species that were not identified during the morphological sorting were removed from the samples, reducing the mean reads per sample to 1,978,250 (±368,804; range: 22,496-8,310,225).

Bias acts consistently across samples and replicates

When the expected relative abundances from morphological counting were compared to the metabarcoding results, substantial taxonomic bias was seen (Figure 2A). The mock larval and adult communities showed the strongest deviation from expected abundances, with an RMSE of 14% and 13% respectively, while the trapped mock communities had an RMSE of 7%. The magnitude and direction of the quantitative bias was relatively

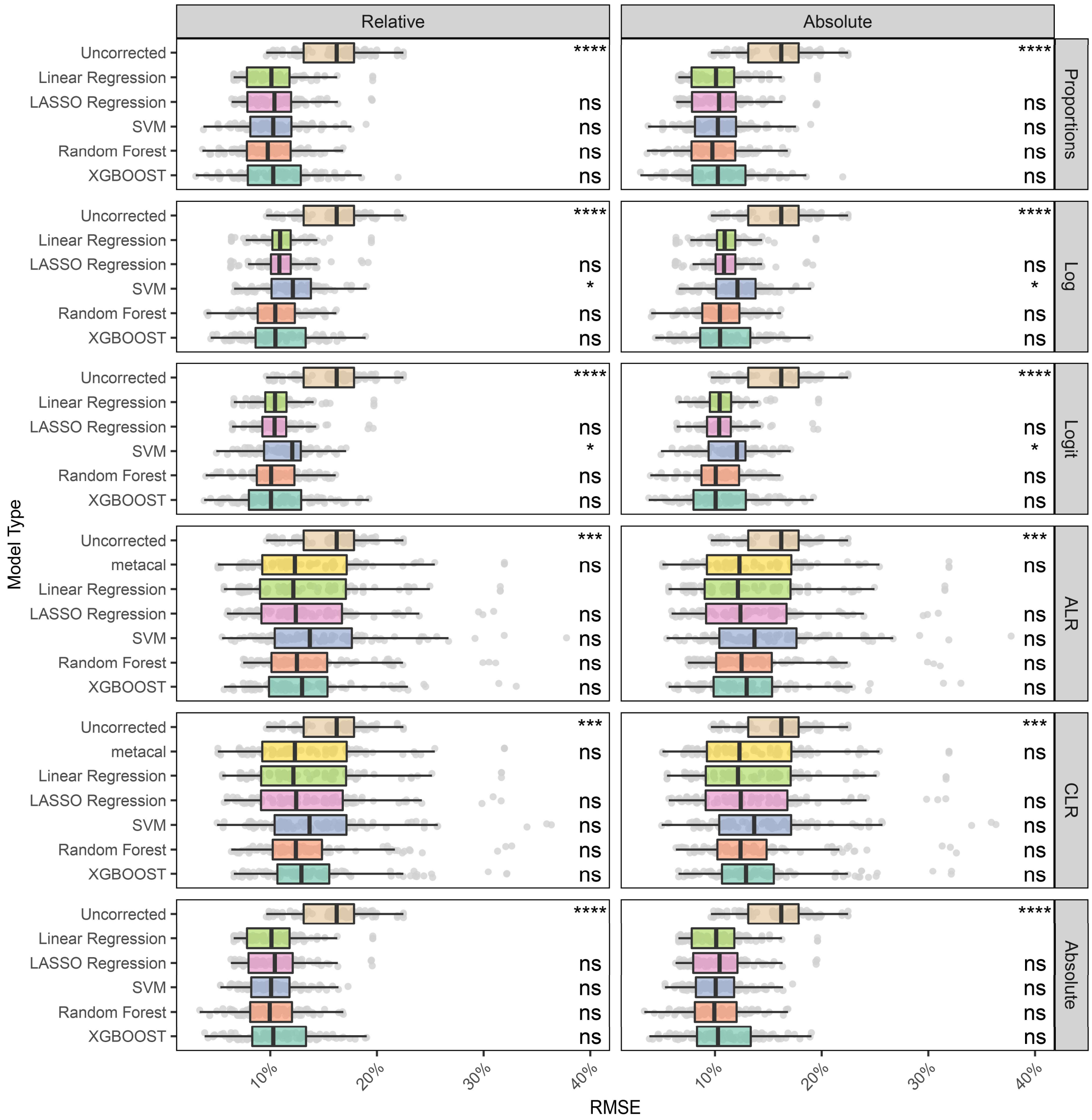


Figure 3: Comparison of Root Mean Square Error (RMSE) for each model and data transformation, calculated across 100 training/test dataset splits. Significance of pairwise t-test comparisons of each model against the baseline linear regression indicated on right of each box. Left panel indicates the relative abundances, while right panel indicates absolute abundances obtained by multiplying the bias corrected relative abundances by the total number of morphologically counted individuals in each community. Abbreviations: ALR; additive log-ratio, CLR; centred log-ratio

consistent between DNA extraction replicates of the same samples and PCR replicates of the same extractions (Figure 2C), while more variability was seen between different trap samples, particularly for *Urophorus humeralis*.

Bias can be corrected in relative and absolute abundances

All 6 predictive models significantly improved the relationship between the expected relative abundances from morphological counting and those observed from metabarcoding (t-tests, p < .001), and this remained consistent after the data was transformed to absolute abundances (Figure 3, Supplementary Figures 3-6). However, there was no significant improvement in quantitative performance for any of the more complex statistical or machine learning models compared to the baseline linear regression model across all data transformations (t-tests, p > .05). On the other hand, the SVM model showed slightly worse performance than the linear regression across both the log and log-odds transformed relative (p = .01) and absolute abundances (p = .04). Ultimately none of the log, log-odds, or CLR or ALR transformations improved the predictive performance of any model over the untransformed relative abundance data, with the latter two CODA transformations producing slightly worse predictions overall (Figure 3). In addition to the CLR or ALR transformations providing no appreciable benefit, the metacal model which was specifically designed for correcting metabarcoding bias using the same CODA framework also showed no improvement in performance over the baseline regression model (p > .05). While not statistically significant, the Random Forest model showed the smallest median RMSE for the 100 random dataset splits across most data transformations (Figure 3). Therefore, the Random Forest model fit directly to the relative abundances with no additional transformations was selected as the final model for correcting taxonomic bias in this dataset.

Performance of final random forest model

For the final Random Forest model, the median RMSE was relatively consistent across all taxa included in the adult mock communities, ranging from 2% for *Brachypeplus sp.* to 6% for *C. davidsonii* (Figure 4A). For the trapped communities on the other hand, *C. hemipterus* showed a median RMSE of 12.2%, significantly higher than all the other taxa (Tukey HSD, p < .001), while *Urophorus humeralis* was significantly lower at 1% (p < .001). The Random Forest model fit to the larval mock communities showed a higher RMSE than both the mock and field collected adults, with *C. davidsonii* having a median RMSE of 10%, and *C. hemipterus* and *C. truncatus* showing a median RMSE of 15% and 15.1% respectively (Figure 4A). Across all taxa the variance in the RMSE between the different training/test

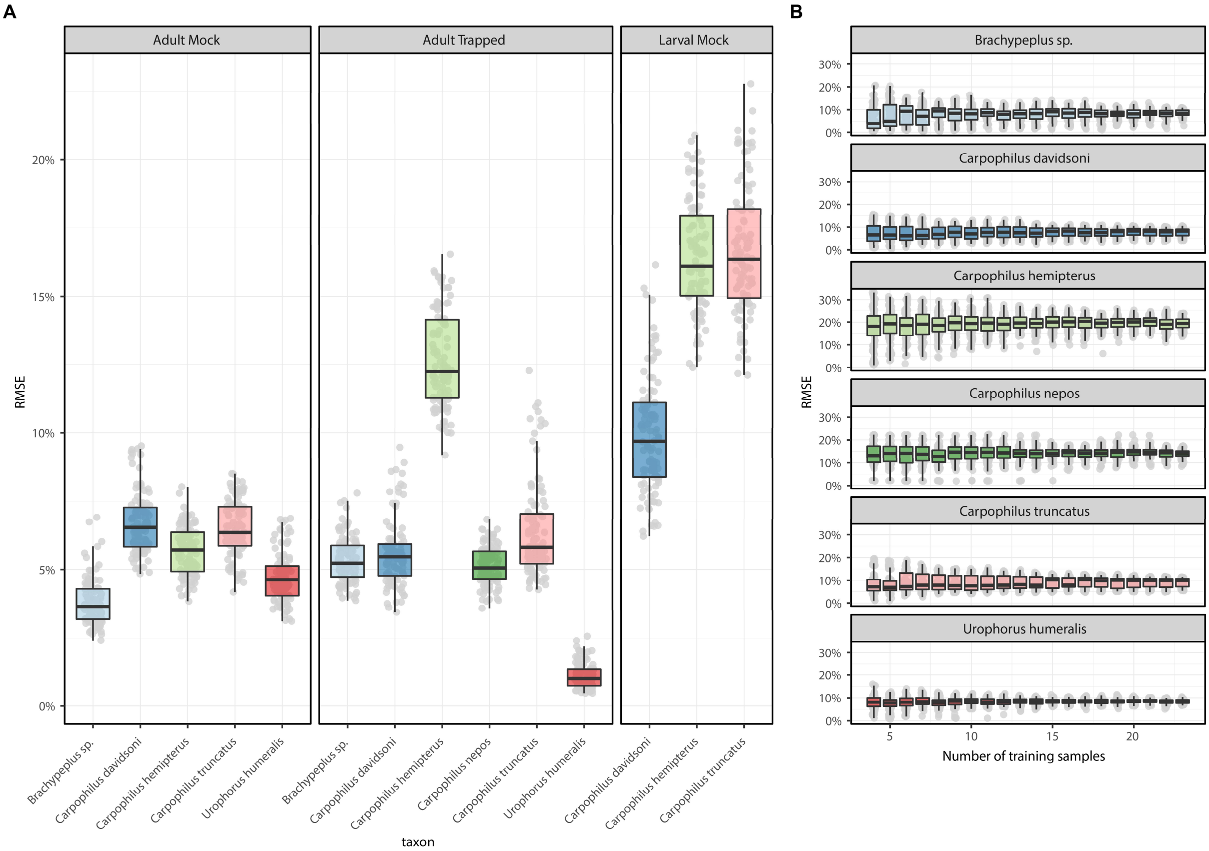


Figure 4: A) Root Mean Square Error (RMSE) for the final Random Forest model across the 100 random test sets, displayed by taxon and community type. **B)** Change in test set RMSE for each taxon as the number of samples used to train the model is increased, repeated across 100 random dataset splits.

splits steadily reduced as more samples were used to train the model, but the median RMSE did not decrease beyond 5 samples used for training (Figure 4B).

Bias calibration makes metabarcoding comparable to morphological counting

Across all taxa, a substantial divergence was seen between the abundances obtained through morphological sorting and those returned by metabarcoding (Figure 5A, 5B). However, when these same metabarcoding results were calibrated using the final Random Forest model and then translated back to absolute abundances, the resulting specimen counts were significantly closer to the morphological count (Figure 5C). This was similarly reflected in multivariate space, where the bias calibration procedure recovered most of the compositional error between the morphological and metabarcoding results (Figure 6). Furthermore, the similarity between the multivariate clustering across the relative and absolute abundances reinforces that analyses based on relative abundances will reflect those made on absolute abundances if a compositionally appropriate distance such as the Aitchison distance is used (Figure 6, Supplementary Figure 7).

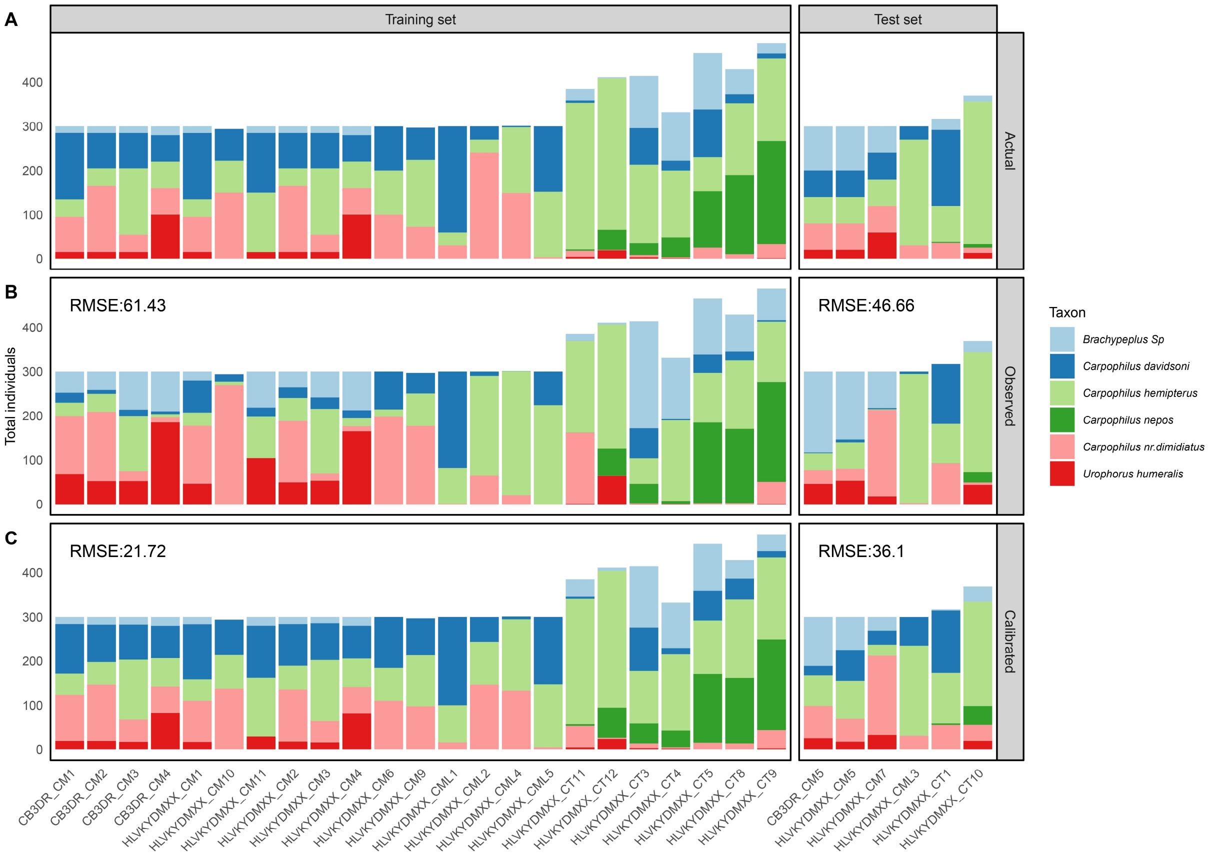


Figure 5: Comparison of measured abundances across each sample for a single random training and test set split. **A)** Actual abundances from morphological counting. **B)** Observed abundances after sequencing transformed to absolute abundances by multiplying by the total number of morphologically counted individuals in the sample. **C)** Observed absolute abundances following calibration with the final Random Forest model.

Discussion

Statistical correction models are shown here to successfully ameliorate taxonomic biases introduced during metabarcoding protocols, enabling accurate estimates of both relative abundance and absolute counts of insect specimens. Following bias-calibration and absolute abundance adjustment, metabarcoding datasets become comparable to those derived from morphological sorting, while requiring substantially less time and personnel to obtain. Importantly, our study began with the use of highly degenerate primers that showed almost no mismatch to the target species, which led to the initial deviation between expected and observed relative abundances being substantially lower than many other metabarcoding studies (Lamb et al., 2019). When fit to this already low taxonomic bias, all evaluated statistical and machine learning models were similarly effective in correcting for it, with neither CODA transformations nor absolute abundance adjustment further improving correctability. This equivalent performance of models with different mechanisms and complexity is likely due to a limit in correctability being reached, where intraspecific variation between specimens begins to outweigh the systematic taxon bias. Intraspecific nucleotide variation can occur in primer sites and affect PCR amplification (Piper et al., 2021), and differences in morphological traits such as cuticle hardness can be the product of developmental environment or exposure to different weather conditions during trapping (Hopkins & Kramer, 1992; Krehenwinkel et al., 2018), with impacts on non-destructive DNA extraction efficiency (Marquina et al., 2019). Furthermore, the "pipeline noise" that is introduced through stochastic sampling of molecules throughout the metabarcoding laboratory workflow (Leray & Knowlton, 2017) may also contribute additional variance to the model fits. Finally, because models were trained on 'ground truth' data obtained through morphological counting, any human error during this process would introduce additional variance into the bias estimates (Culverhouse et al., 2014; MacLeod et al., 2010). Using more pre-identified samples to train the correction models increases the robustness to both intraspecific variation and occasional large random errors, yet this only marginally increased the overall correctability and comes at the expense of reduced sequencing effort applied to real samples. Therefore, we suggest that including a minimum of 5 pre-identified samples per sequencing run should allow taxonomic bias to be sufficiently captured, and future studies should determine whether previously obtained bias estimates can be accurately extrapolated across batches and sequencing runs.

The similar performance of the non-linear machine learning models to the simple linear regression across all data transformations suggests that metabarcoding taxonomic bias can be accurately captured with log-linear relationships, reinforcing the multiplicative model proposed by McLaren et al., (2019). Although it should be noted that neither the *metacal* model nor CODA transformations proposed by the same study provided any appreciable benefits over ignoring the compositional nature of the data and fitting models directly to the relative abundances. While this seems to conflict with previous theory (Gloor et al., 2017; Quinn et al., 2018), this discrepancy may be due to the simple and consistent species composition of the samples analysed here. The main justification for use of CODA transformations is to increase consistency of bias estimates across differently composed samples (McLaren et al., 2019), yet the pheromone lure used to collect samples for our study meant that after the necessary removal of the low-abundance bycatch taxa, the analysed communities were almost taxonomically identical.

Therefore, while we did not see any benefits of CODA transformations for estimating bias, they may be more applicable when analysing communities collected through less targeted sampling methods, such as traps containing lures derived from host odours (Chapter 4) or passively collected wind-borne insects (Batovska et al., 2020).

Similar to the CODA transformations, there was also no appreciable gain in accuracy when samples were transformed to absolute abundances either prior to, or following, model training and bias calibration. Despite this, a metabarcoding assay which provides absolute counts of specimens or biomass rather than sequence reads is an important step towards easing interpretation by non-specialists, as well as integration into economic injury threshold models. Our study used the total number of individuals counted during morphological sorting to adjust the metabarcoding data into absolute abundances, but a higher throughput method for obtaining this independent measurement would be desirable for practical application: for instance, the total weight or volume of each community could be measured prior to DNA extraction (although this may introduce additional biases due to not all species being the same physical size (Elbrecht et al., 2017)), or specimen counts could be obtained from photographs of trap catches using image analysis techniques (Mele, 2013). For studies where an independent measurement of absolute abundance cannot be obtained, such as when samples have already gone through destructive DNA extraction, taxonomic bias calibration performs just as well on relative abundances alone (Supplementary Figure 7). So, while incorporating absolute abundance information aids interpretation, conducting analyses on bias-calibrated relative abundances remains valid approach as long as the statistical challenges of comparing relative abundances across-samples are appropriately considered (Gloor et al., 2017).

The commercially available Carpophilus lure used in this study is a synthetic blend derived from the pheromones of three *Carpophilus* species (Bartelt et al., 1995), combined with a synergistic "co-attractant" of fermenting fruit volatiles (Bartelt & Hossain, 2006). Attraction to fermentation volatiles is common across diverse taxonomic groups (Davis et al., 2013), and it was therefore unsurprising when an additional 26 species were recorded in the metabarcoding analysis compared to the morphological sorting, which aimed to identify only the high priority pests. As the correction models could only be

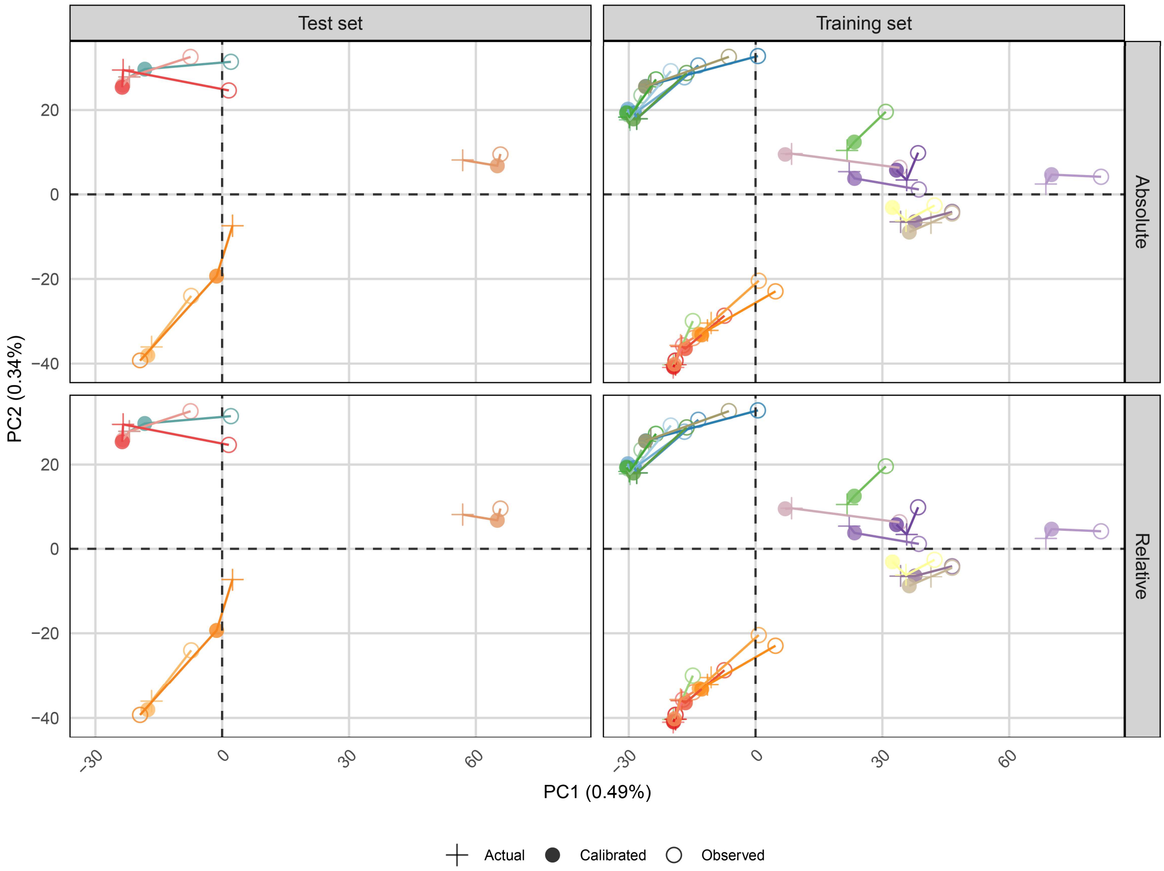


Figure 6: Principal coordinate analysis of Aitchison distance between actual, observed, and calibrated abundances using the final Random Forest model for each sample within in the test and training sets. Bias calibration produces metabarcoding results that more closely reflect the actual morphological content for both relative and absolute abundances.

trained on species for which measurements were obtained in advance, these taxa had to be excluded from further analysis, an aspect which remains a major limitation for application of correction factors to more diverse communities. To address this, McLaren et al. (2019) highlighted the potential use of phylogenetic imputation methods to extrapolate corrections factors from measurable species to closely related taxa which share similar traits, an approach that is commonly employed to predict bacterial genome content from 16S metabarcoding sequences (Goberna & Verdú, 2016; Zaneveld & Thurber, 2014). While a promising avenue for future research, this approach may be complicated by metabarcoding bias being the joint product of multiple laboratory steps, each interacting with various species traits (Martoni et al., in prep). Yet even if metabarcoding must be targeted to a smaller cohort of measurable species in order to be quantitative, this is no different to alternative quantitative molecular assays such as qPCR, while still allowing substantially more species to be identified in a single diagnostic test. On this note it is worth mentioning recent statistical advances which use a quantitative measurement for a subset of taxa (commonly obtained through qPCR) to estimate taxonomic bias for the remainder of the species in the sample (Williamson et al., 2021). This alternative framework avoids the complications involved in assembling mock communities and could conceivably be integrated with "spike-in" internal standards (Harrison et al., 2021) to provide a more streamlined quantitative metabarcoding workflow.

When crop managers do not have access to accurate and timely pest abundance information the quality of management decisions can be affected, leading to an over-reliance on damaging insecticides which increase input costs and eliminate populations of natural enemies (Peterson et al., 2018). Bias-calibrated metabarcoding poses a promising method for scaling up the identification and quantification of pest taxa by insect diagnostic laboratories to support IPM population monitoring. While wider adoption of metabarcoding faces challenges related to diagnostic turnaround time, cost of platforms, and technological access (Piper et al., 2019), the rapid and ongoing evolution of this technology (e.g. low-cost and real time nanopore sequencing) will most certainly circumvent much of this in the near future (Baloğlu et al., 2021; Krehenwinkel et al., 2019). Importantly, the model-based bias calibration approach presented here is independent of the target taxonomic groups, laboratory protocol, and HTS platform used, and may therefore prove an effective and readily adoptable method for increasing the quantitative accuracy of metabarcoding assays for insect population monitoring.

Acknowledgments

The authors would like to thank Mofakhar Hossein, Kevin Farnier, Daniel Lai, and David Madge for providing mixed *Carpophilus* samples from colonies and population monitoring activities. Field work for this study was conducted on traditional lands of the First Peoples of the Millewa-Mallee, the Latji Latji, Ngintait, and Nyeri Nyeri, and we pay our respects to their Elders past and present. Financial support was provided by Horticulture Innovation Australia grant AL16009 "An integrated pest management program for the Australian almond industry" and Agriculture Victoria's Improved Market Access for Horticulture program (CMI105584). A.M.P. is supported by an Australian Government Research Training Program Scholarship.

Availability of data and materials:

Raw sequence reads have been uploaded to NCBI SRA acc no: (XXXXX, to be assigned later) and final OTU tables and reference sequences used to make taxonomic assignments are available from dryad reference no: (XXXXX, to be assigned later). All code required to reproduce the statistical analyses and generate all figures presented in this paper is contained in the following GitHub repository: https://github.com/alexpiper/carpophilus_metabarcoding

Author contributions

A.M.P., J.P.C. and M.J.B. conceptualised the study, A.M.P. performed all molecular laboratory procedures, bioinformatic and statistical analyses. L.R. performed all morphological identification of trap samples. L.R. and L.S. generated the Carpophilus reference sequences used for identification. A.M.P. wrote the first draft of the manuscript with input and supervision from J.P.C., N.O.I.C., and M.J.B. All authors read and approved the final version of the manuscript.

References

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)*. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

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

Baloğlu, B., Chen, Z., Elbrecht, V., Braukmann, T., MacDonald, S., & Steinke, D. (2021). A workflow for accurate metabarcoding using nanopore MinION sequencing. *Methods in Ecology and Evolution*, 00, 1–11. https://doi.org/10.1111/2041-210x.13561

Bartelt, R. J., & Hossain, M. S. (2006). Development of synthetic food-related attractant for Carpophilus davidsoni and its effectiveness in the Stone Fruit orchards in Southern Australia. *Journal of Chemical Ecology*, 32(10), 2145–2162. https://doi.org/10.1007/s10886-006-9135-7

Bartelt, R. J., Vetter, R. S., Carlson, D. G., Petroski, R. J., & Baker, T. C. (1995). Pheromone combination lures for Carpophilus (Coleoptera: Nitidulidae) species. *Journal of Economic Entomology*, 88(4), 864–869.

Batovska, J., Piper, A. M., Valenzuela, I., Cunningham, J. P., & Blacket, M. J. (2020). Developing a Non-destructive Metabarcoding Protocol for Detection of Pest

Insects in Bulk Trap Catches. *Research Square*. https://doi.org/10.21203/rs.3.rs-125070/v1

Blacket, M. J., Semeraro, L., & Malipatil, M. B. (2012). Barcoding Queensland Fruit Flies (Bactrocera tryoni): Impediments and improvements. *Molecular Ecology Resources*, 12(3), 428–436. https://doi.org/10.1111/j.1755-0998.2012.03124.x

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1201/9780429469275-8

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. https://doi.org/10.1111/ddi.12465

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLoS ONE*, 12(10), e0185056. https://doi.org/10.1371/journal.pone.0185056

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7), 581–583. https://doi.org/10.1038/nmeth.3869

Cha, D. H., Adams, T., Werle, C. T., Sampson, B. J., Adamczyk, J. J., Rogg, H., & Landolt, P. J. (2014). A four-component synthetic attractant for Drosophila suzukii (Diptera: Drosophilidae) isolated from fermented bait headspace. *Pest Management Science*, 70(2), 324–331. https://doi.org/10.1002/ps.3568

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

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. https://doi.org/10.1111/1755-0998.12265

Cortes, C., & Vladimir, V. (1995). Support-Vector Networks. *Machine Learning*, 20, 273–297. https://doi.org/10.1109/64.163674

Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., Granger, B., Green, L., Howd, T., Mason, T., Vicente, G., Dasilva, M., Brodeur, W., DeSmet, T., Dodge, S., Lennon, N. J., & Gabriel, S. (2018). Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*, 19, 332. https://doi.org/10.1186/s12864-018-4703-0

Crisci, C., Ghattas, B., & Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling*, 240, 113–122. https://doi.org/10.1016/j.ecolmodel.2012.03.001

Culverhouse, P. F., Macleod, N., Williams, R., Benfield, M. C., Lopes, R. M., & Picheral, M. (2014). An empirical assessment of the consistency of taxonomic identifications. *Marine Biology Research*, 10(1), 73–84. https://doi.org/10.1080/17451000.2013.810762

Cunningham, J. P., Carlsson, M. A., Villa, T. F., Dekker, T., & Clarke, A. R. (2016). Do Fruit Ripening Volatiles Enable Resource Specialism in Polyphagous Fruit Flies? *Journal of Chemical Ecology*, *42*(9), 931–940. https://doi.org/10.1007/s10886-016-0752-5

Davis, T. S., Crippen, T. L., Hofstetter, R. W., & Tomberlin, J. K. (2013). Microbial Volatile Emissions as Insect Semiochemicals. *Journal of Chemical Ecology*, 39(7), 840–859. https://doi.org/10.1007/s10886-013-0306-z

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biology Letters*, *10*(9), 20140562. https://doi.org/10.1098/rsbl.2014.0562

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., Kartzinel, T. R., & Eveson, J. P. (2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology*, *28*(2), 391–406. https://doi.org/10.1111/mec.14734

Digirolomo, M. F., & Dodds, K. J. (2014). Cerambycidae Bycatch from Asian Longhorned Beetle Survey Traps Placed in Forested Environs. *Northeastern Naturalist*, 21(3), 28–34. https://doi.org/10.1656/045.021.0310

Edgar, R. C., & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, *31*(21), 3476–3482. https://doi.org/10.1093/bioinformatics/btv401

Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Hajibabaei, M., Wright, M., Zakharov, E. V., Hebert, P. D. N., & Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, 7, e7745. https://doi.org/10.7717/peerj.7745

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

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8, 2224. https://doi.org/10.3389/fmicb.2017.02224

Goberna, M., & Verdú, M. (2016). Predicting microbial traits with phylogenies. *ISME Journal*, 10(4), 959–967. https://doi.org/10.1038/ismej.2015.171

Gray, M. E., Ratcliffe, S. T., & Ratcliffe, S. T. (2008). The IPM paradigm: Concepts, strategies and tactics. In *Integrated Pest Management: Concepts, Tactics, Strategies and Case Studies*. https://doi.org/10.1017/CBO9780511626463.002

Harrison, J., Calder, W. J., Shuman, B., & Buerkle, C. A. (2021). The quest for absolute abundance: the use of internal standards for DNA-barcoding in microbial ecology. *Molecular Ecology Resources*, 21, 30–43. https://doi.org/10.32942/osf.io/q7gy6

Hopkins, T. L., & Kramer, K. J. (1992). Insect cuticle sclerotization. *Annual Review of Entomology*, 37(1), 273–302. https://doi.org/10.1146/annurev.en.37.010192.001421

Hossain, M. S., & Bartelt, R. (2010). Chemical ecology of Carpophilus sap beetles (Coleoptera: Nitidulidae) and development of an environmentally friendly method of crop protection. *Terrestrial Arthropod Reviews*, 3(1), 29–61. https://doi.org/10.1163/187498310X489981

Hossain, M. S., Williams, D. G., Mansfield, C., Bartelt, R. J., Callinan, L., & Il'Ichev, A. L. (2006). An attract-and-kill system to control Carpophilus spp. in Australian stone fruit orchards. *Entomologia Experimentalis et Applicata*, 118(1), 11–19. https://doi.org/10.1111/j.1570-7458.2006.00354.x

Kogan, M. (1988). Integrated pest management theory and practice. *Entomologia Experimentalis et Applicata*, 49, 59–70. https://doi.org/10.1111/j.1570-7458.1988.tb02477.x

Krehenwinkel, H., Fong, M., Kennedy, S., Huang, E. G., Noriyuki, S., Cayetano, L., & Gillespie, R. (2018). The effect of DNA degradation bias in passive sampling devices on metabarcoding studies of arthropod communities and their associated microbiota. *PLoS ONE*, 13(1), e0189188. https://doi.org/10.1371/journal.pone.0189188

Krehenwinkel, H., Pomerantz, A., Henderson, J. B., Kennedy, S. R., Lim, Y., Swamy, V., Shoobridge, J. D., Patel, N. H., Rosemary, G., Prost, S., Lim, J. Y., Swamy, V., Shoobridge, J. D., Graham, N., Patel, N. H., Gillespie, R. G., & Prost, S. (2019). Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *GigaScience*, 8(6), giz006. https://doi.org/10.1093/gigascience/giz006

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. https://doi.org/10.1038/s41598-017-17333-x

Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* https://www.tidymodels.org

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How quantitative is metabarcoding: a meta-analytical approach. *Molecular Ecology*, 28(2), 420–430. https://doi.org/10.1111/mec.14920

Leray, M., & Knowlton, N. (2017). Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ*, 3, e3006. https://doi.org/10.7717/peerj.3006

Leschen, R. A. B., & Marris, J. W. M. (2005). Carpophilus (Coleoptera: Nitidulidae) of New Zealand with notes on Australian species. *Landcare Research Contract Report*: LC0405/153.

Liu, M., Clarke, L. J., Baker, S. C., Jordan, G. J., & Burridge, C. P. (2020). A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology*, 45(3), 373–385. https://doi.org/10.1111/een.12831

Lucas, T. C. D. (2020). A translucent box: interpretable machine learning in ecology. *Ecological Monographs*, 90(4), e01422. https://doi.org/10.1002/ecm.1422

MacLeod, N., Benfield, M., & Culverhouse, P. (2010). Time to automate identification. *Nature*, 467(7312), 154–155. https://doi.org/10.1038/467154a

Marquina, D., Esparza-Salas, R., Roslin, T., & Ronquist, F. (2019). Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. *Molecular Ecology Resources*, 19(6), 1516–1530. https://doi.org/10.1111/1755-0998.13071

Martoni, F., Piper, A. M., Rodoni, B. C., & Blacket, M. J. (2021). Disentangling bias for non-destructive metabarcoding of insects. *In Preparation.*

McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing measurements. *ELife*, 8, e46923. https://doi.org/10.7554/eLife.46923

Mele, K. (2013). Insect soup challenge: Segmentation, counting, and simple classification. *Proceedings of the IEEE International Conference on Computer Vision*, 168–171. https://doi.org/10.1109/ICCVW.2013.28

Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., Zengler, K., & Knight, R. (2019). Establishing microbial composition measurement standards with reference frames. *Nature Communications*, 10, 2719. https://doi.org/10.1038/s41467-019-10656-5

Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, 6(1), 140. https://doi.org/10.1186/s40168-018-0521-5

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. https://doi.org/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(2), 407– 419. https://doi.org/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), giz092. https://doi.org/10.1093/gigascience/giz092

Piper, A. M., Cogan, N. O. I., Cunningham, J. P., & Blacket, M. J. (2021). Computational Evaluation of DNA Metabarcoding for Universal Diagnostics of Invasive Insect Pests. *BioRxiv*. https://doi.org/10.1101/2021.03.16.435710

Props, R., Kerckhof, F. M., Rubbens, P., Vrieze, J. De, Sanabria, E. H., Waegeman, W., Monsieurs, P., Hammes, F., & Boon, N. (2017). Absolute quantification of microbial taxon abundances. *ISME Journal*, *11*(2), 584–587. https://doi.org/10.1038/ismej.2016.117

Quinn, T. P., Erb, I., Gloor, G., Notredame, C., Richardson, M. F., & Crowley, T. M. (2019). A field guide for the compositional analysis of any-omics data. *GigaScience*, 8(9), giz107. https://doi.org/10.1093/gigascience/giz107

Quinn, T. P., Erb, I., Richardson, M. F., & Crowley, T. M. (2018). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16), 2870–2878. https://doi.org/10.1093/bioinformatics/bty175

Quinn, T. P., Le, V., & Cardilini, A. P. A. (2021). Test set verification is an essential step in model building. *Methods in Ecology and Evolution*, 12, 127–129. https://doi.org/10.1111/2041-210X.13495

R Core Team. (2019). *R: A Language and Environment for Statistical Computing.* https://www.r-project.org/

Ramsden, M. W., Kendall, S. L., Ellis, S. A., & Berry, P. M. (2017). A review of economic thresholds for invertebrate pests in UK arable crops. *Crop Protection*, 96, 30–43. https://doi.org/10.1016/j.cropro.2017.01.009

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science.* https://doi.org/10.1214/ss/1177012413

Spears, L. R., Looney, C., Ikerd, H., Koch, J. B., Griswold, T., Strange, J. P., & Ramirez, R. A. (2016). Pheromone Lure and Trap Color Affects Bycatch in Agricultural Landscapes of Utah. *Environmental Entomology*, 45(4), 1009–1016. https://doi.org/10.1093/ee/nvw085

Tan, K. H., Nishida, R., Jang, E. B., & Shelly, T. E. (2014). Pheromones, male lures, and trapping of tephritid fruit flies. In *Trapping and the Detection, Control, and Regulation of Tephritid Fruit Flies: Lures, Area-Wide Programs, And Trade Implications.* https://doi.org/10.1007/978-94-017-9193-9_2

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Topçuoğlu, B. D., Lesniak, N. A., Ruffin IV, M. T., Wiens, J., & Schloss, P. D. (2020). A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. *MBio*, 11(3), e00434-20. https://doi.org/10.1101/816090

Vamos, E. E., Elbrecht, V., & Leese, F. (2017). Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcoding and Metagenomics*, 1, e14625. https://doi.org/10.3897/mbmg.1.14625

Weber, D. C., & Ferro, D. N. (1991). Nontarget Noctuids Complicate Integrated Pest Management Monitoring of Sweet Corn with Pheromone Traps in Massachusetts. *Journal of Economic Entomology*, 84(4), 1364–1369. https://doi.org/10.1093/jee/84.4.1364

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. http://ggplot2.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wilcox, T. M., Zarn, K. E., Piggott, M. P., Young, M. K., McKelvey, K. S., & Schwartz, M. K. (2018). Capture enrichment of aquatic environmental DNA: A first proof of concept. *Molecular Ecology Resources*, 18(6), 1392–1401. https://doi.org/10.1111/1755-0998.12928

Wilkinson, S. (2019). aphid: an R package for analysis with profile hidden Markov models. *Bioinformatics*, 35(19), 3829–3830. https://doi.org/10.1093/bioinformatics/btz159

Williamson, B. D., Hughes, J. P., & Willis, A. D. (2021). A multiview model for relative and absolute microbial abundances. *Biometrics*, 1–14. https://doi.org/10.1111/biom.13503

Witzgall, P., Kirsch, P., & Cork, A. (2010). Sex pheromones and their impact on pest management. *Journal of Chemical Ecology*, 36(1), 80–100. https://doi.org/10.1007/s10886-009-9737-y

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

5.5 Supplementary Information

Equation 1: Proportions regression model

$$t\left(\frac{exp_{ij}}{\sum exp_j}\right) = \beta_1\left(t\left(\frac{obs_{ij}}{\sum obs_j}\right)\right) + \beta_2(Taxon_i) + \beta_3(seq\ run_j) + \beta_4(comm\ type_j)$$

Where i is the respective taxon, j is the respective sample, and t is either the natural log, log-odds, or absolute abundance transformation.

Equation 2: Compositional regression model

$$t\left(\frac{obs_{ij}}{exp_{ij}}\right) = \beta_1(Taxon_i) + \beta_2(seq\ run_j) + \beta_3(seq\ run_j)$$

Where i is the respective taxon, j is the respective sample, and t is either the CLR or ALR transformation.

Equation 3: Additive log-ratio transform (ALR)

$$alr(x) = \left[log\frac{x_i}{x_r}, \dots, log\frac{x_I}{x_r}\right]$$

Where r is an arbitrary reference taxon, i is the first taxa and I is the last taxa.

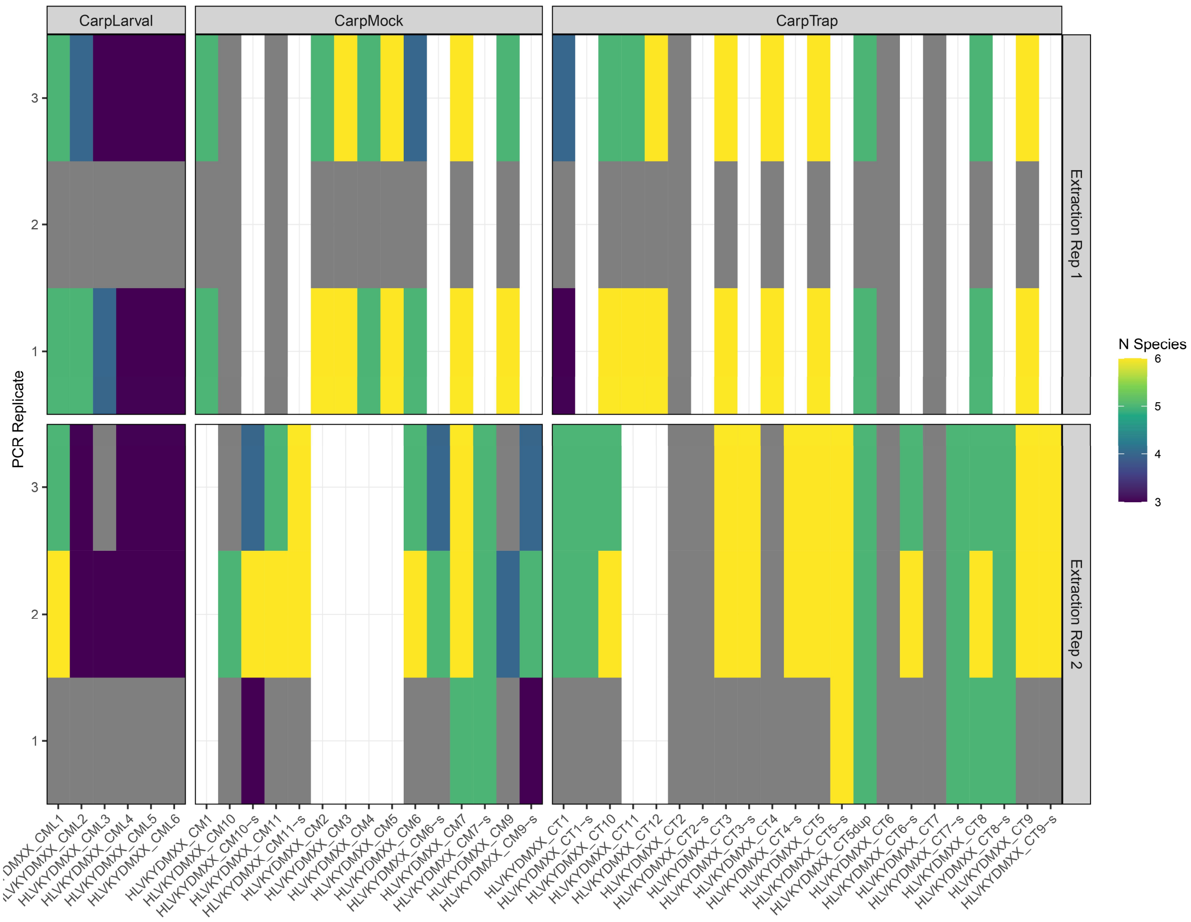
Equation 4: Centred log-ratio transform (CLR)

$$clr(x) = \left[log \frac{x_i}{g(x)}, \dots, log \frac{x_I}{g(x)}\right]$$

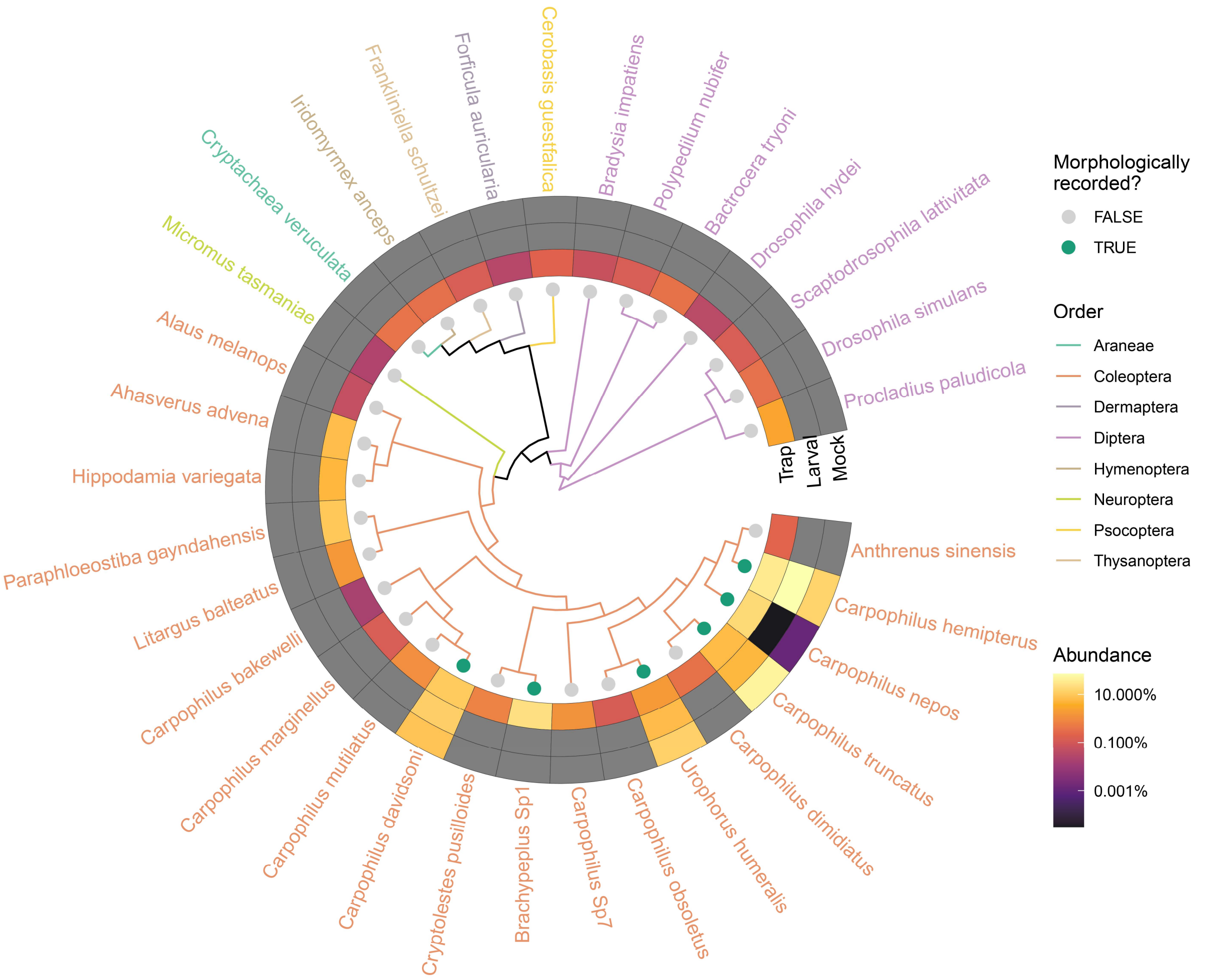
Where g is the geometric mean, i is the first taxa and I is the last taxa.

Supplementary Table 1: Number of individual specimens from each species included within each mock community, or morphologically identified within each trapped community.

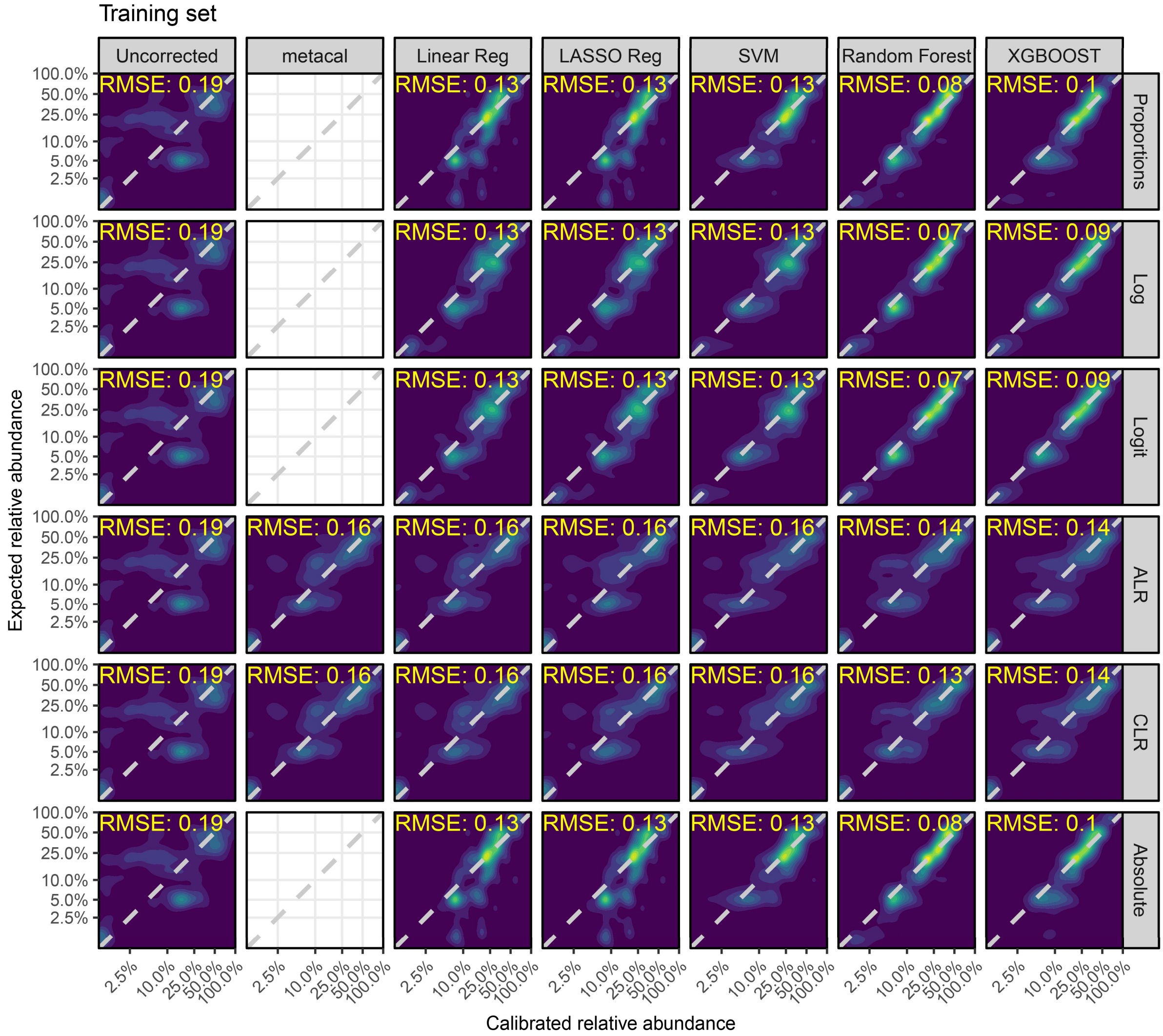
Community Type	Sample Name	Carpophilus davidsoni	Carpophilus truncatus	Carpophilus hemipterus	Urophorus humeralis	Brachypeplus Sp.	Carpophilus nepos	Carpophilus marginalus
Mock	CM1	150	80	40	15	15		
Mock	CM2	80	150	40	15	15		
Mock	CM3	80	40	150	15	15		
Mock	CM4	60	60	60	100	20		
Mock	CM5	60	60	60	20	100		
Mock	CM6	100	100	100	0	0		
Mock	CM7	60	60	60	60	60		
Mock	CM8	100	3	100	49	48		
Mock	CM9	73	73	151	0	0		
Mock	CM10	72	150	72	0	0		
Mock	CM11	135	0	135	15	15		
Larvae	CML1	240	30	30				
Larvae	CML2	30	240	30				
Larvae	CML3	30	30	240				
Larvae	CML4	3	149	149				
Larvae	CML5	148	3	149				
Larvae	CML6	148	149	3				
Trapped	CT1	172	36	82	0	25	2	0
Trapped	CT2	32	25	123	1	97	51	1
Trapped	CT3	83	5	178	3	118	27	3
Trapped	CT4	22	2	151	1	110	46	1
Trapped	CT5	108	25	77	0	128	128	3
Trapped	CT6	40	0	118	3	143	135	3
Trapped	CT7	25	22	184	2	51	51	6
Trapped	CT8	20	10	162	0	57	180	1
Trapped	CT9	11	32	187	1	23	234	5
Trapped	CT10	0	12	323	13	13	8	2
Trapped	CT11	5	14	332	4	27	3	1
Trapped	CT12	0	1	343	19	2	46	88



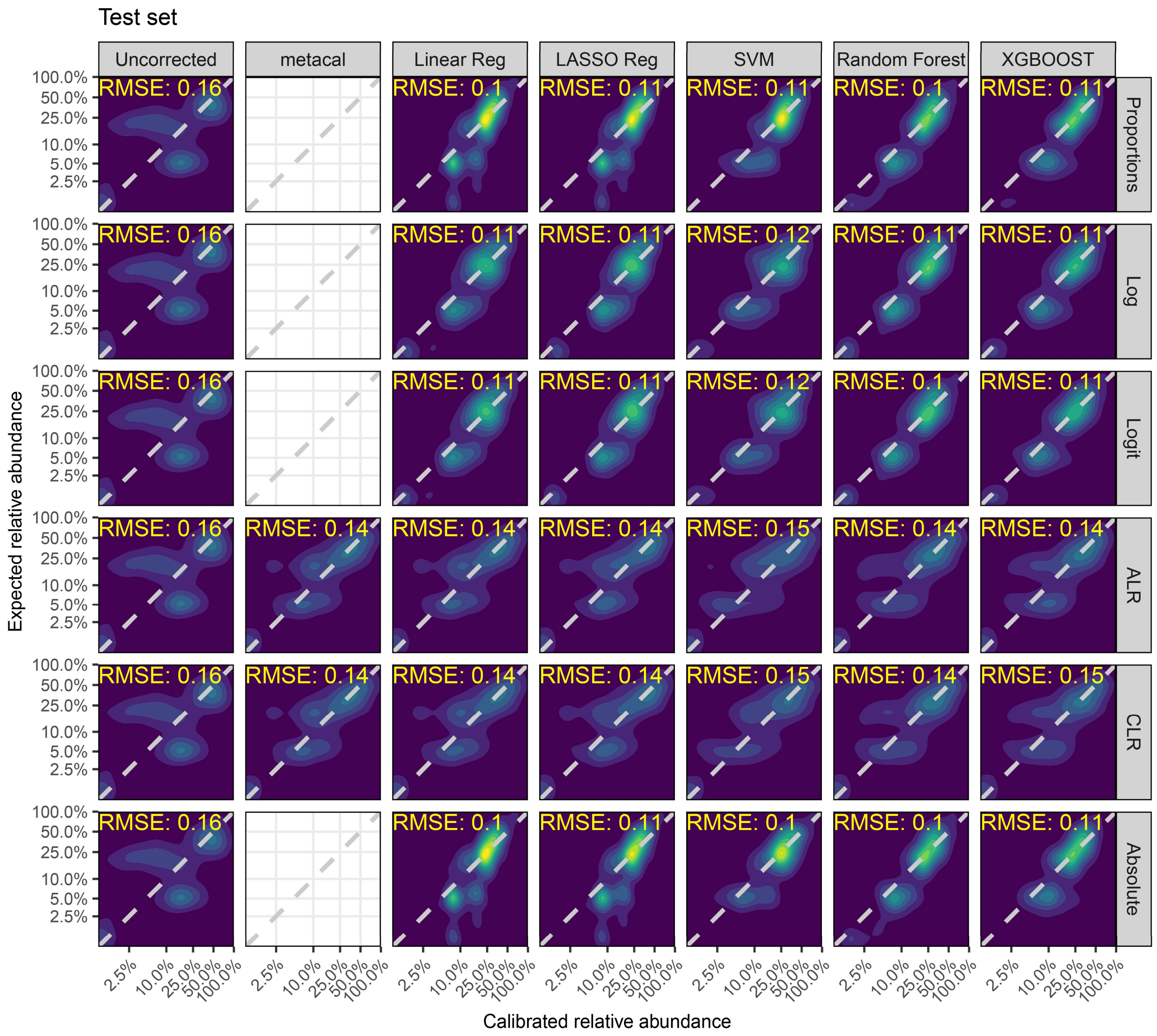
Supplementary Figure 1: Number of mock community taxa detected across each DNA extraction and PCR replicate. Grey indicates dropouts of all taxa within a sequenced replicate, while white indicates samples that were not replicated.



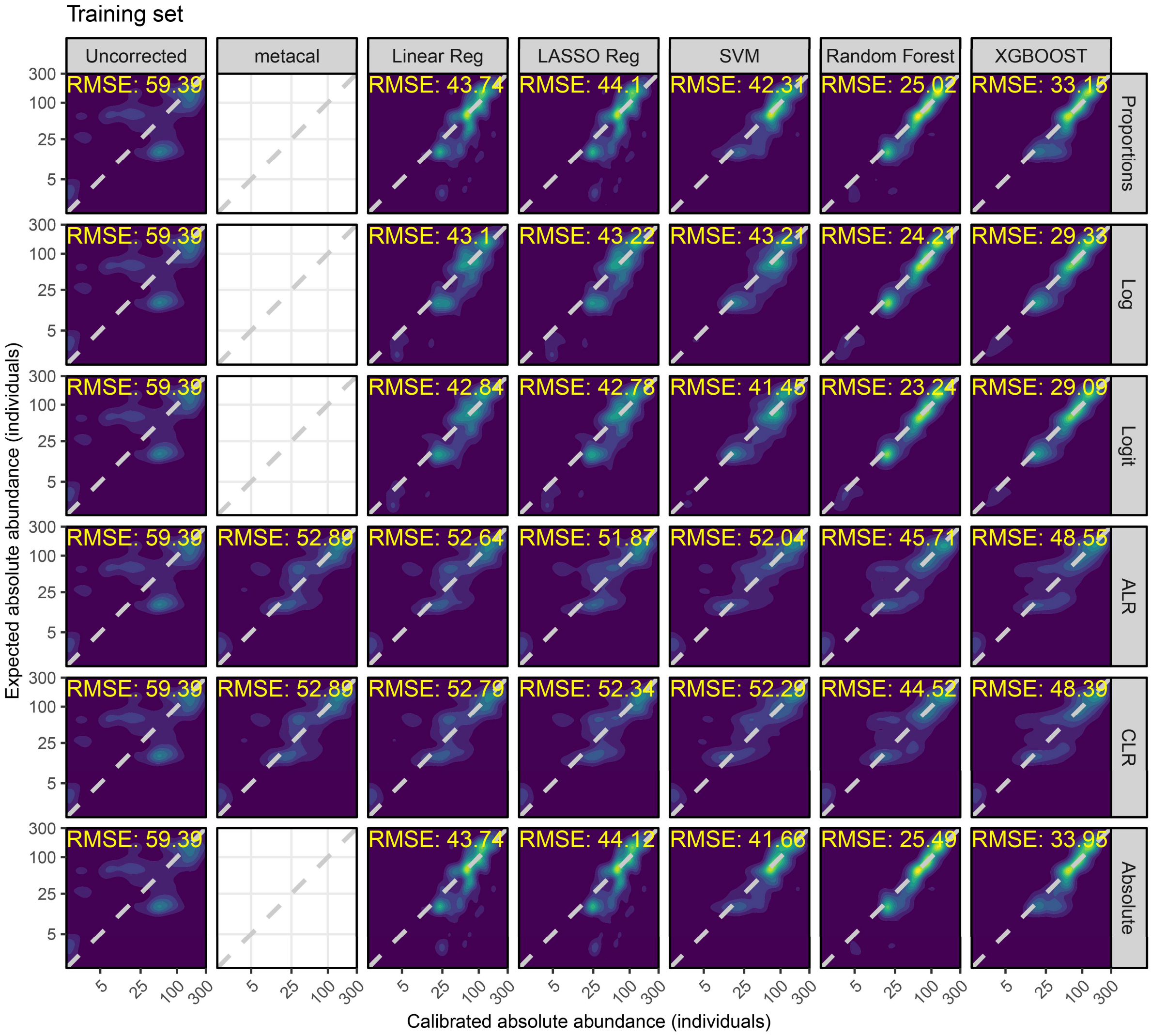
Supplementary Figure 2: Phylogenetic relationships between all species detected within the trap samples, with those detected in both morphological sorting and metabarcoding highlighted. The mean relative abundance of the respective taxa across each community type is displayed as a heatmap.



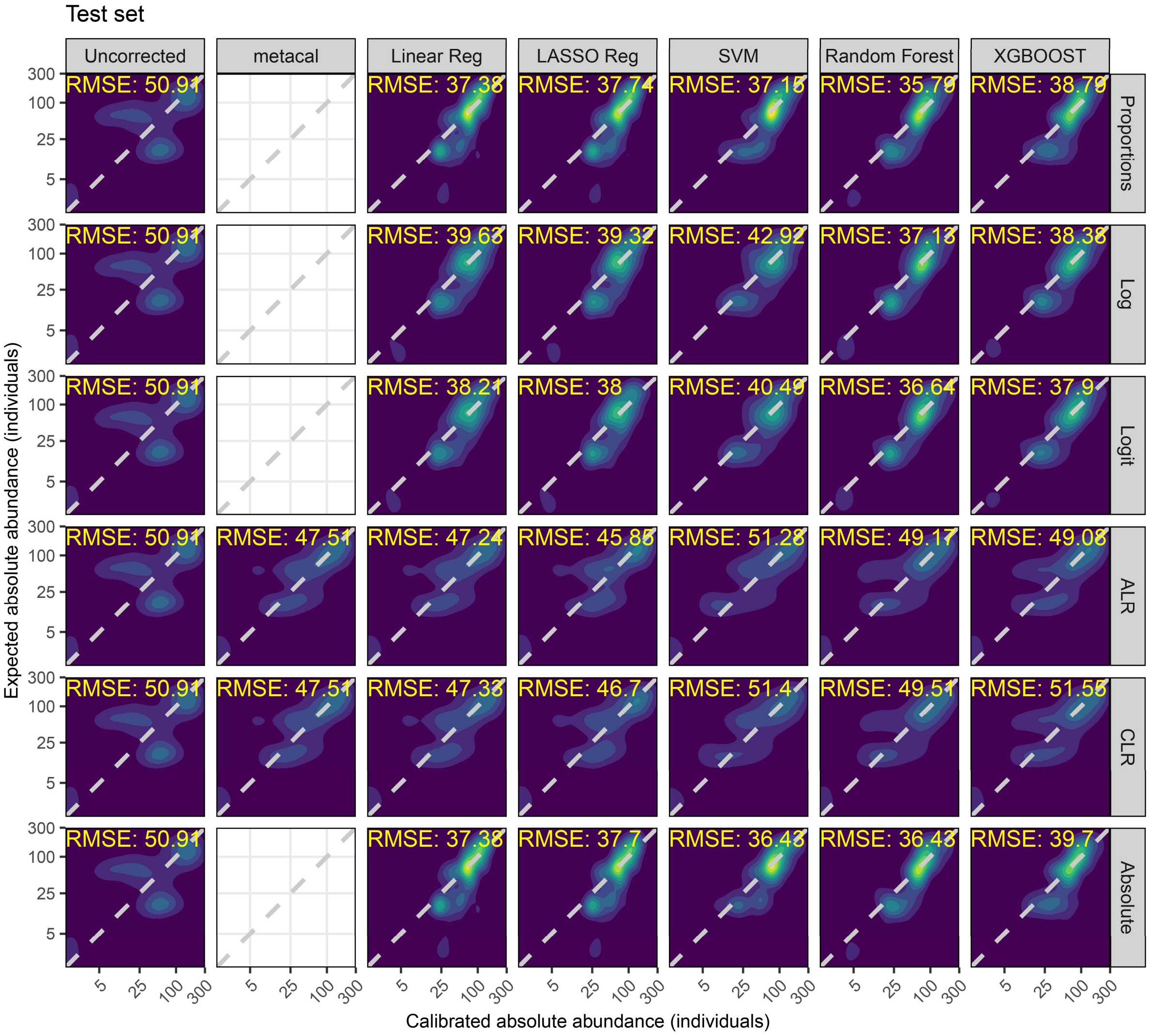
Supplementary Figure 3: Density plot of expected relative abundances compared to model calibrated relative abundances across each model tyle and data transformation, displayed for the training sets from all 100 data splits. Data are displayed on a pseudo-log scale to avoid compressing variation near zero.



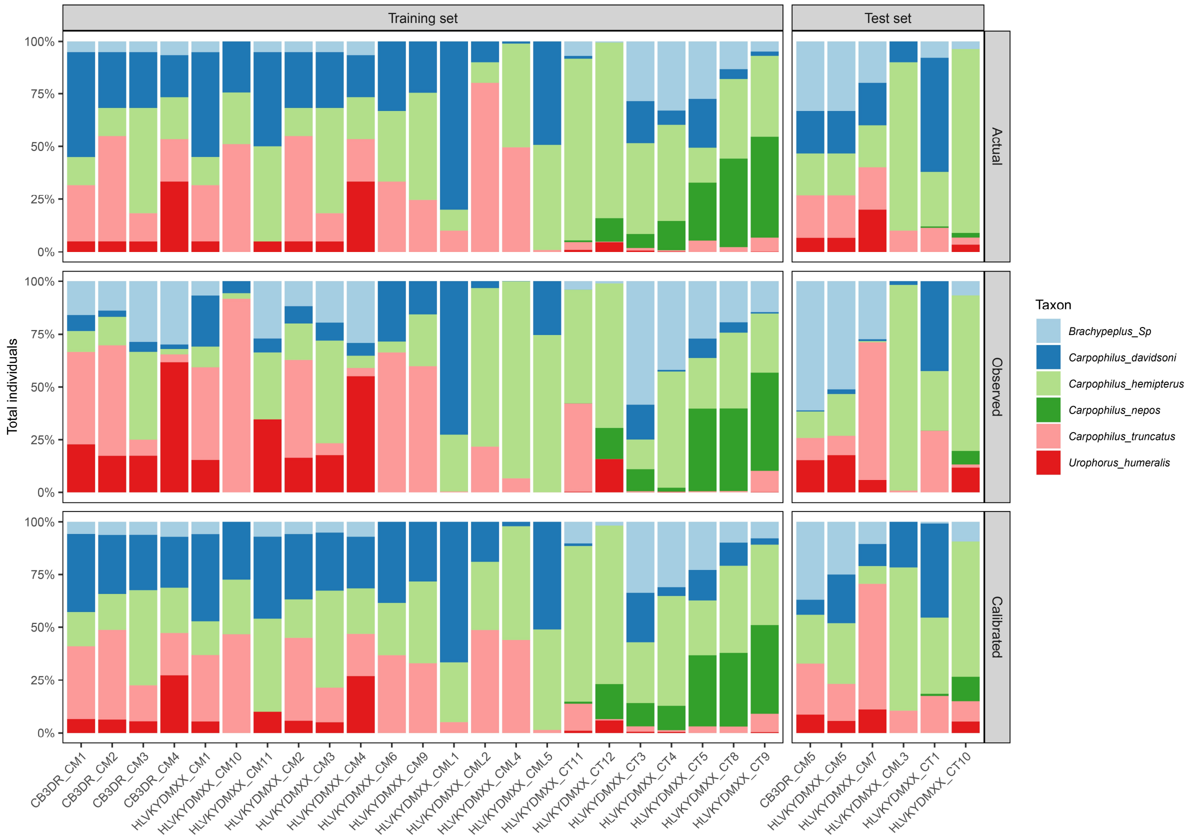
Supplementary Figure 4: Density plot of expected relative abundances compared to model calibrated relative abundances across each model type and data transformation, displayed for the test sets from all 100 data splits. Data are displayed on a pseudo-log scale to avoid compressing variation near zero.



Supplementary Figure 5: Density plot of expected absolute abundances compared to model calibrated absolute abundances across each model type and data transformation, displayed for the training sets from all 100 data splits. Absolute abundances were obtained by multiplying the corrected relative abundances by the total number of individuals in the sample. Data are displayed on a pseudo-log scale to avoid compressing variation near zero.



Supplementary Figure 6: Density plot of expected absolute abundances compared to model calibrated absolute abundances across each model type and data transformation, displayed for the test sets from all 100 data splits. Absolute abundances were obtained by multiplying the corrected relative abundances by the total number of individuals in the sample. Data are displayed on a pseudo-log scale to avoid compressing variation near zero.



Supplementary Figure 7: Comparison of relative abundances for each sample in a single training and test set split, from top to bottom: Actual relative abundances from morphological counting, observed relative abundances after sequencing, and observed relative abundances after calibration using the final random forest model.

6
Exploring the Genomic Consequences of Range Expansion in an Invasive Tephritid Fruit Fly

6.1 Chapter preface:

While the metabarcoding assay developed in chapters 3-5 was shown to successfully detect and quantify abundances of insect pests, the limited nucleotide variation contained within the COI mini-barcode does not provide sufficient resolution for tracing the source of new outbreaks. Therefore, this chapter develops a complementary low-coverage whole genome sequencing (lcWGS) assay to predict the geographic origin of intercepted specimens and explore patterns of genetic diversity during colonisation and establishment. This approach is developed and validated on the range expansion of the Queensland fruit fly (*Bactrocera tryoni*), a highly polyphagous pest endemic to Australia but only recently established in the temperate fruit growing regions of Victoria. The population structure of *B. tryoni* is characterised through sequencing of specimens collected from across the entire endemic and invasive range, then used as a reference panel to train a deep-learning model for predicting geographic origin. The accuracy of this model is evaluated through cross-validation with samples of known origin, then used to trace the source of recent outbreaks in Tasmania, the Yarra Valley, and Auckland, New Zealand. This chapter is presented as a self-contained manuscript in the final stages of preparation, with intended submission to the journal *Evolutionary Applications*, and includes supplementary material at the end.

6.2 Publication details:

Exploring the genomic consequences of range expansion in an invasive Tephritid fruit fly

Stage of publication: In Preparation

Journal details: Evolutionary Applications

Authors: Alexander M. Piper, Noel O.I. Cogan, John Paul Cunningham, Mark J. Blacket

6.3 Statement of authorship:

A.M.P., N.O.I.C. and M.J.B. conceptualised the study, A.M.P. performed all molecular laboratory procedures, bioinformatic and statistical analyses and wrote the first draft of the manuscript with input and supervision from J.P.C., N.O.I.C., and M.J.B. All authors contributed to the editing of the final manuscript and approved the version presented here.

Statement from co-author confirming the contribution of the PhD candidate:

"As co-author of the manuscript 'Piper, A.M., Cogan N.O.I, Cunningham, J. P. & Blacket M.J. (In preparation). Exploring the genomic consequences of range expansion in an invasive Tephritid fruit fly, *Evolutionary Applications*', I confirm that Alexander M. Piper has made the contributions listed above."

Associate Professor John Paul Cunningham

30/03/2021

6.4 Manuscript

Exploring the genomic consequences of range expansion in an invasive Tephritid fruit fly

Alexander M. Piper[1,2], Noel O.I. Cogan[1,2], John Paul Cunningham[1,2], Mark J. Blacket[1]

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

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

Keywords: Invasive species, genome skimming, spatial population genomics, assignment tests, pathway analysis, machine learning

Running title: Queensland fruit fly invasion genomics

Corresponding author:

Alexander M. Piper

Email: alexander.piper@agriculture.vic.gov.au

Abstract

Over the past century the geographic range of the polyphagous plant pest, the Queensland fruit fly (*Bactrocera tryoni*, Diptera: Tephritidae), has expanded from its endemic coastal tropical and subtropical forests into temperate fruit growing regions of Australia. Outbreaks and specimen interceptions within previously pest-free areas have become regular occurrences, impacting horticultural exports, and incurring costly quarantine and intervention procedures. Determining whether these outbreaks arise from long-distance dispersal events, or resurgent local populations which have evaded eradication is a priority for coordinating biosecurity responses. In this study we use genome-wide markers, obtained through low-coverage whole genome sequencing (lcWGS), to characterise the genetic structure of endemic and invasive *B. tryoni* populations and predict the geographic origin of recent outbreaks. We find the demographic history of *B. tryoni* is defined by two major endemic populations on the East- and North-coasts of Australia, with the former being the source of both the southwards range expansion and colonisation of island populations in Melanesia. These endemic populations are genetically homogenous over large distances with no isolation-by-distance, while geographically isolated populations are highly differentiated following population bottlenecks. Temporal sampling within the southernmost invasive populations, in Victoria, revealed a broad increase in genetic diversity over the past decade, with early founder populations being replaced by further immigration from the expansion front. Evidence for recent population bottlenecks suggests, however, that regular die-off and recolonisation is ongoing in some areas, despite *B. tryoni* now being declared endemic in Victoria. Specimens intercepted from outbreaks in Tasmania, the Yarra Valley, and Auckland, New Zealand, were tentatively assigned to the south-eastern invasive range, but weak concordance between genetic and geographic structure led to confidence intervals encompassing a considerably larger geographic area. The importance of these results in context of area wide management of *B. tryoni*, and the use of lcWGS approaches for exploring insect invasions are discussed.

Introduction

Many species experience periods of range expansion or contraction throughout their evolutionary history in response to shifts in the geographic distribution of suitable habitats (Holt, 2003). More recently, anthropogenic activity has become the primary driver of range expansions, through the artificial introduction of individuals into new environments and homogenisation of habitats by agricultural and urban development (Elton, 1958; Lodge, 2003). The highly polyphagous Tephritid fruit fly *Bactrocera tryoni* (Queensland fruit fly), presents a striking example of rapid human-mediated range expansion, leading it to become one of Australia's most damaging horticultural pests (Clarke et al., 2011). Considered endemic to tropical and subtropical rainforests along the eastern coast of Queensland (QLD) and New South Wales (NSW) (Drew, 1989; Meats, 1981), within the last century the geographic distribution of *B. tryoni* has expanded south into temperate fruit growing regions of Victoria (VIC), west into the Northern Territory (NT) and overseas to various Melanesian islands (Clarke et al., 2011; Dominiak & Mapson, 2017; May, 1962). Several important fruit growing regions depend upon Pest Free Area (PFA) status to ensure access to valuable interstate and international markets (Dominiak et al., 2015), however, interceptions of *B. tryoni* specimens within these areas have become a regular occurrence, pausing fruit exports and incurring costly quarantine and intervention procedures (Florec et al., 2013; Suckling et al., 2016). Determining whether these outbreaks arise through new incursions from the endemic range or resurgent local populations which have evaded eradication efforts is a priority for coordinating biosecurity response (McInnis et al., 2016; Sved et al., 2003).

Tephritid fruit flies do not naturally disperse great distances, and instead commonly spread via long distance human-assisted movements followed by local diffusion through natural insect flight (Dominiak, 2012; Sadler et al., 2011). When specimens are intercepted while still associated with a human vector, such as within a fruit shipment or at a vehicle inspection checkpoint, their introduction pathway may be readily identified from cargo manifests or interviews with drivers. When specimens are instead intercepted within surveillance traps placed near agricultural, urban, or natural areas, confidently establishing the pathway and timeframe of introduction can prove impossible using conventional approaches (Barr et al., 2014). In these cases, molecular genetic techniques can be used to place outbreaks back to their source population, even when multiple lifecycles have progressed since the initial introduction event (Cristescu, 2015; Estoup & Guillemaud, 2010).

Previous studies of *B. tryoni* genetic structure have revealed that populations along the east coast native range are homogenous over large distances (Gilchrist et al., 2006; Yu et al., 2001), but distinct from those in NT and northern Western Australia (WA) (Cameron et al., 2010; Popa-Báez et al., 2020), which some authors accord full species status as *B. aquilonis* (Drew & Lambert, 1986; Morrow et al., 2000). In contrast, incipient (small, localised) populations within the south-eastern invasive range show multiple distinct origins, with greatly reduced gene flow compared to established regions (Blacket et al., 2017; Gilchrist & Meats, 2010). These studies used microsatellites or mitochondrial haplotypes to genotype specimens (Blacket et al., 2017; Gilchrist et al., 2006; Gilchrist & Meats, 2010; Yu et al., 2001), which were then matched against potential source populations (Gilchrist & Meats, 2010; Sved et al., 2003), or used to detect kinship groups within outbreak zones (Gilchrist et al., 2004). However, reliable population assignment using these limited loci, which represent a very small fraction of the genome, requires source populations to be relatively old and not share gene pools (Barr et al., 2014), and thus may be unsuitable when regional co-ancestry is already high such as when tracing localised incursions following an initial invasion event (Fitzpatrick et al., 2012; Schmidt et al., 2021). More recently, Popa-Báez and colleagues (2020) applied genome wide single nucleotide polymorphisms (SNPs) obtained through reduced-representation high-throughput sequencing (HTS) to examine *B. tryoni* population structure, yet placing outbreak specimens to anywhere more fine-scale than 'the east coast of Australia' remained a challenge, despite using an order of magnitude more loci than previous investigations (Popa-Báez et al., 2021). Nevertheless, given that genetic homogeneity cannot be fully achieved by any population where individual dispersal is smaller than its geographic range (Bradburd & Ralph, 2019; Cristescu, 2015), improved resolution may be obtainable using a method of genotyping that samples an even larger portion of the genome across many individuals.

As an alternative to common reduced-representation HTS approaches which sample small portions of the genome at high sequencing coverage, outlined above, low-coverage whole genome sequencing (lcWGS) instead samples the entire genome at low coverage, providing substantially more SNPs for similar costs (Lou et al., 2020; Therkildsen & Palumbi, 2017). At this low depth of coverage reliably differentiating real mutations from sequencing error becomes challenging (Nielsen et al., 2011, 2012), however, for many questions relevant to invasion biology it is not the genotype at any particular site that matters, but rather the pattern of variation across the genome (North et al., 2021). Therefore, use of probabilistic analysis frameworks which take uncertainty about individual SNP genotypes into account can still provide reliable inference about that individual's overall genetic signature (Korneliussen et al., 2014; Meisner & Albrechtsen, 2018). Simulation studies have demonstrated that when following this approach, sequencing more individuals at lower depths (0.5-2× coverage of the genome) can maximise the information obtained and provide more accurate estimates of population level parameters than sequencing fewer individuals to greater depths (Alex Buerkle & Gompert, 2013; Fumagalli, 2013; Lou et al., 2020). The size of these datasets can further enable the use of supervised machine learning methods for population assignment, a process-agnostic framework that may be particularly suited for range expanding species which, by definition, do not conform to typical assumptions of discrete, well-mixed populations (Battey et al., 2020; Schmidt et al., 2021).

In this study we use genome-wide markers, obtained through lcWGS of specimens collected from the entire endemic and invasive range, to describe the contemporary population structure of the Queensland fruit fly (*Bactrocera tryoni*). Through fine-scale sampling of the Victorian invasive range during the initial invasion in 2011/12, then later in 2017/18, we further explore spatial patterns of genetic diversity occurring during colonisation and establishment. Using this continental-scale genotype dataset as a reference panel, a spatially-explicit deep learning approach is used to predict the geographic origin of specimens intercepted during recent outbreaks from Tasmania, the Yarra Valley, and Auckland, New Zealand (NZ). The importance of these results in context of area wide management of *B. tryoni*, and the use of lcWGS approaches for understanding the historical dynamics of insect invasions are discussed.

Methods:

Sample collection, DNA Extraction and Library preparation

Male B. *tryoni* were collected from 43 endemic or recently invaded locations around Australia, New Caledonia, and French Polynesia (Table 1) using cue-lure baited traps (Meats & Hartland, 1999), and morphologically identified using standard characters (Plant Health Australia, 2018). Additional 'outbreak' specimens were collected from Auckland, NZ, in 2015, Perth, WA, in 2011, northern Tasmania in both 2011 and 2018, and the Yarra Valley, VIC in 2018. At the time of processing, all specimens had been stored either dry or in absolute ethanol at −20 °C for between 1 and 7 years. Genomic DNA was extracted from each specimen using the Qiagen DNeasy 96 blood and tissue kit within the QIACube automated sample preparation system (Qiagen, Germany), and integrity of resulting DNA evaluated using 2% w/v agarose gel electrophoresis. 605 individual DNA extracts with visible high molecular weight bands and concentrations >10 ng/µL as measured by a Qubit 2.0 fluorometer (Thermo Fisher, USA) were selected for library preparation. Genomic DNA was enzymatically sheared using the method described by Shinozuka et al. (2015). In brief, 1 mM of 5-methyl-dCTP (New England Biolabs, USA) was randomly incorporated into 1 µL of genomic DNA via whole genome amplification with the REPLI-g UltraFast mini kit (Qiagen, Germany), then digested using the MspJI restriction enzyme (New England Biolabs, USA), which recognises the modified 5-methylcytosine bases. Digested DNA underwent an end-filling and dA-tailing reaction using the JetSeq flex DNA library preparation kit (Bioline, USA), followed by ligation of in-house adapters. Adapter ligated libraries were double sided size selected to retain products of approximately 280-429 bp, first using a 0.2:1 ratio of Agencourt AMPure XP beads (Beckman Coulter, USA) to DNA and discarding the beads, followed by a second 0.2:1 ratio discarding the supernatant. Eight (8) bp dual indexes were then attached using 7 cycles of real time PCR with cycling conditions of 98 °C for 10s, 65 °C for 30s, and 72 °C for 30s, followed by a SYBR Green fluorescence read. Indexed libraries were quantified via melt curve analysis, then equimolarly pooled in batches of 96 using a Biomek FX^P liquid handling robot (Beckman Coulter, USA). Each pooled library was purified using a 0.8:1 ratio of AMPure XP beads to DNA, then sized and quantified using a 2200 TapeStation (Agilent Technologies, USA) and Qubit 3.0 fluorometer. A pooled library containing 108 specimens representing key established populations were sequenced on an Illumina HiSeq 3000 (v4 chemistry)

Table 1: Summary of collections made for all populations analysed in this study.

Location	ID	Region	N (early)	N (late)	Year (early)	Year (late)
Alice Springs	1	Central Aus	0	8		2017
Kalumburu	2	North Coast	0	7		2017
Darwin	3	North Coast	0	4		2017
Cobourg Peninsula	4	North Coast	0	16		2017-2018
Coen	5	East Coast	12	0	2011	
Cairns	6	East Coast	8	11	2011	2017
Mourilyan Harbour	7	East Coast	0	7		2017
Townsville	8	East Coast	7	7	2011	2017
Mackay	9	East Coast	4	11	2011	2017
Gladstone	10	East Coast	9	6	2011	2017
Bundaberg	11	East Coast	9	8	2011	2017
Brisbane	12	East Coast	11	6	2011	2017
Sydney	13	East Coast	8	0	2010	
Dubbo	14	Inland NSW	5	0	2010	
Cootamundra	15	Inland NSW	5	0	2011	
Hillston	16	Inland NSW	3	0	2010	
Wodonga	17	Inland VIC	4	3	2011	2017
Rutherglen	18	Inland VIC	1	2	2011	2017
Dookie	19	Inland VIC	4	5	2011	2017
Cobram	20	Inland VIC	4	2	2011	2017
Yarrawonga	21	Inland VIC	4	4	2011	2017
Shepparton	22	Inland VIC	3	4	2011	2017
Kyabram	23	Inland VIC	3	4	2011	2017
Echuca	24	Inland VIC	6	5	2011	2017
Barham	25	Inland VIC	4	6	2011	2017
Speewa	26	Inland VIC	3	3	2011	2017
Wood Wood	27	Inland VIC	5	4	2011	2017
Boundary bend	28	Inland VIC	6	6	2011	2017
Robinvale	29	Inland VIC	1	2	2011	2017
Nichols Point	30	Inland VIC	0	4		2017
Merbein	31	Inland VIC	5	3	2011	2017
Ellerslie	32	Inland NSW	7	6	2011	2017
Orbost	33	Coastal VIC	14	0	2009-2010	
Marlo	34	Coastal VIC	15	0	2009-2011	
Lakes Entrance	35	Coastal VIC	6	5	2009	2018
Bruthen	36	Coastal VIC	0	3		2018
Upper Tambo	37	Coastal VIC	4	0	2010	
Eagle Point	38	Coastal VIC	4	3	2010	2018
Sarsfield	39	Coastal VIC	0	3		2018
Bairnsdale	40	Coastal VIC	4	1	2010	2018
Sale	41	Coastal VIC	6	4	2010	2018
Yarra Valley	42	Outbreak	0	2		2018
Tasmania	43	Outbreak	2	1	2011	2018
Auckland	44	Outbreak	0	4		2015
New Caledonia	45	Melanesia	0	5		2017
Tahiti	46	Melanesia	0	3		2016
Hakatao	47	Melanesia	0	6		2016

using 2 × 150 bp reads, aiming for ‘moderate’ 10× coverage of the genome per specimen. The remaining 377 successful libraries were pooled separately and sequenced on either a

HiSeq 3000 (94 specimens) or a NovaSeq 6000 S2 flow cell lane (283 specimens), both using 2 × 150bp reads and aiming for 2× coverage of the genome.

Bioinformatics

Sequence data from the HiSeq and NovaSeq lanes were demultiplexed using *bcl2fastq* and filtered with *fastp* (Chen et al., 2018) to only retain reads with a mean base quality >20, >50 bp in length, and containing <5 consecutive N bases, as well as remove all Illumina adapter sequences and polyG tails which can occur in NovaSeq data (Arora et al., 2019). Filtered reads were mapped to the *B. tryoni* v2.2 draft reference genome (Gilchrist et al., 2014) using BWA-*m*em (Li, 2013), retaining only properly paired reads with a mapping quality >30. PCR and optical duplicates were removed using the *markdup* function of SAM*tools* v1.9 (Li et al., 2009), and reads realigned around indels using GATK *IndelRealigner* (Depristo et al., 2011). A hard-called list of variants was generated from just the moderate coverage HiSeq data using SAMtools *mpileup*, then filtered to retain only biallelic SNPs with a minor allele frequency (MAF) >5% and <20% missing data. This list of high confidence variants was used to recalibrate base quality scores within all BAM files in order to avoid introducing systematic biases due to the different sequencing technologies used (De-Kayne et al., 2021).

Genotype calling and filtering

Following base quality score recalibration, SNPs with a p-value <1e^{-6} were identified from both the moderate and low coverage datasets, and genotype likelihoods calculated using the empirical Bayesian framework implemented in the software ANGSD (Korneliussen et al., 2014). Variants from the entire dataset were filtered to retain only biallelic SNPs with <50% missing data, however this time the MAF filter was lowered to 1% to capture alleles private to populations where few individuals were successfully sequenced (Linck & Battey, 2019). Additionally, all sites with >10,000 reads across the dataset, or >50% heterozygote genotypes were removed as they likely represented incorrectly mapped paralogs, repetitive regions, or nuclear mitochondrial pseudogenes (Blacket et al., 2012; Matz, 2018). As uncertainty in genotype calling can arise within low coverage sequencing data due to difficulties resolving true mutations from mapping and sequencing errors (Han et al., 2014), all subsequent analyses were conducted using genotype likelihoods rather than hard-called genotypes unless indicated, in order to integrate this uncertainty into inferences (Nielsen et al., 2011, 2012). Linkage disequilibrium between variant sites was calculated in windows of 50 kb using ngsLD (Fox et al., 2019), and unlinked SNPs obtained using the included network-pruning method. Per-site inbreeding coefficients were calculated taking population structure into account using PCAngsd (Meisner & Albrechtsen, 2018, 2019), with the optimal number of eigenvectors used in the model determined from Velicier's Minimum Average Partial (MAP) test (Shriner, 2011). All sites which deviated from Hardy-Weinberg Equilibrium (HWE) expectations with a p value $<1e^{-6}$ were removed as they likely represented erroneous genotypes (R. S. Waples & Allendorf, 2015). Kinship between sequenced individuals was determined from the combination of KING-robust, R0, and R1 statistics calculated using *NGSRelate* V2 (Hanghøj et al., 2019), which allows identification out to 3rd order kin, as well as distinction between parent-offspring and full-sibling relationships within 1st order kin (R. K. Waples et al., 2019). To avoid biasing estimates of genetic diversity and population structure, for all detected close-kin dyads (full-sibling, half-sibling, or parent-offspring) the individual with the lower sequencing coverage was removed from subsequent analyses.

Genetic diversity

Allele frequency likelihoods were estimated directly from BAM files using ANGSD, considering only sites with <20% missing data but not using MAF or SNP likelihood filters to avoid biasing the site frequency spectrum (Matz, 2018). A maximum likelihood estimate of the folded site frequency spectrum (SFS) was generated for each population which had >2 sequenced individuals, then the average number of pairwise differences between sequences (θ_π; Nei & Li, 1979) and total number of segregating sites (θ_w; Watterson, 1975) were calculated in windows of 15 kb using the SFS as a prior. The Tajima's (*D*) test for neutrality (Tajima, 1989) was then calculated from windowed θ_π and θ_w to determine evidence for population expansion or decline. Per-individual inbreeding coefficients (*F*) were calculated using PCAngsd, and genome wide heterozygosity (H_E) using the EM algorithm from realSFS (Nielsen et al., 2012). To test for founder effects within the southern invasive range, H_E for each VIC and NSW population was regressed against its geographic distance to either coastal Gippsland or Sydney respectively as potential source populations.

Population structure & admixture

A covariance matrix of individual genotype probabilities taking population structure into account was generated using PCAngsd, then decomposed into eigenvectors in R4.3 (R Core Team, 2019). Per-individual admixture proportions were calculated using PCAngsd (number of ancestral populations (K) equal to the number of principal components used to model the dataset +1). Relative genetic differentiation between populations (2DSFS; Korneliussen et al., 2014) was estimated between each sampling location with >2 sequenced individuals, using the same filtering parameters as the intra-population genetic diversity. The weighted pairwise Fixation Index (F_{ST}) between populations was calculated genome-wide and in 15 kb windows using the 2DSFS as a prior, according to the method of Reynolds et al., (1983). To ensure analyses of population structure were not confounded by recent selection, SNPs located within the top 5% of windows in each pairwise F_{ST} comparison were removed to produce a putatively 'neutral' dataset. To test for Isolation-by-distance (IBD; Wright, 1946), the correlation between geographic distance and PCA latent space or pairwise F_{ST} was assessed using Mantel tests (Diniz-Filho et al., 2013).

Spatial assignment

The spatially-explicit deep learning based Locator method (Battey et al., 2020) was used to assign outbreak individuals to their most probable geographic origin. As Locator does not currently operate on genotype likelihoods, genotypes with a probability >95% were hard-called from the posterior probabilities output by PCAngsd and converted to a VCF file. A 10-fold Cross Validation (CV), and spatial CV procedure (Brenning, 2012) were performed to evaluate the predictive accuracy of the Locator model, where either a random 10% of specimens, or an entire geographic region were dropped out of the training set and their locations re-predicted. To measure confidence in Locator predictions, windows of 10,000 SNPs were used for model training and prediction, and a two-dimensional kernel density surface was fit over the separate windowed predictions to derive a point estimate and associated confidence contours (Battey et al., 2020). CV error was quantified by the distance in kilometres between the point estimate and the samples' true location (km-error), as well as the number of times the true location was contained within the 95%, 50% or 10% confidence contours. As any georeferencing errors

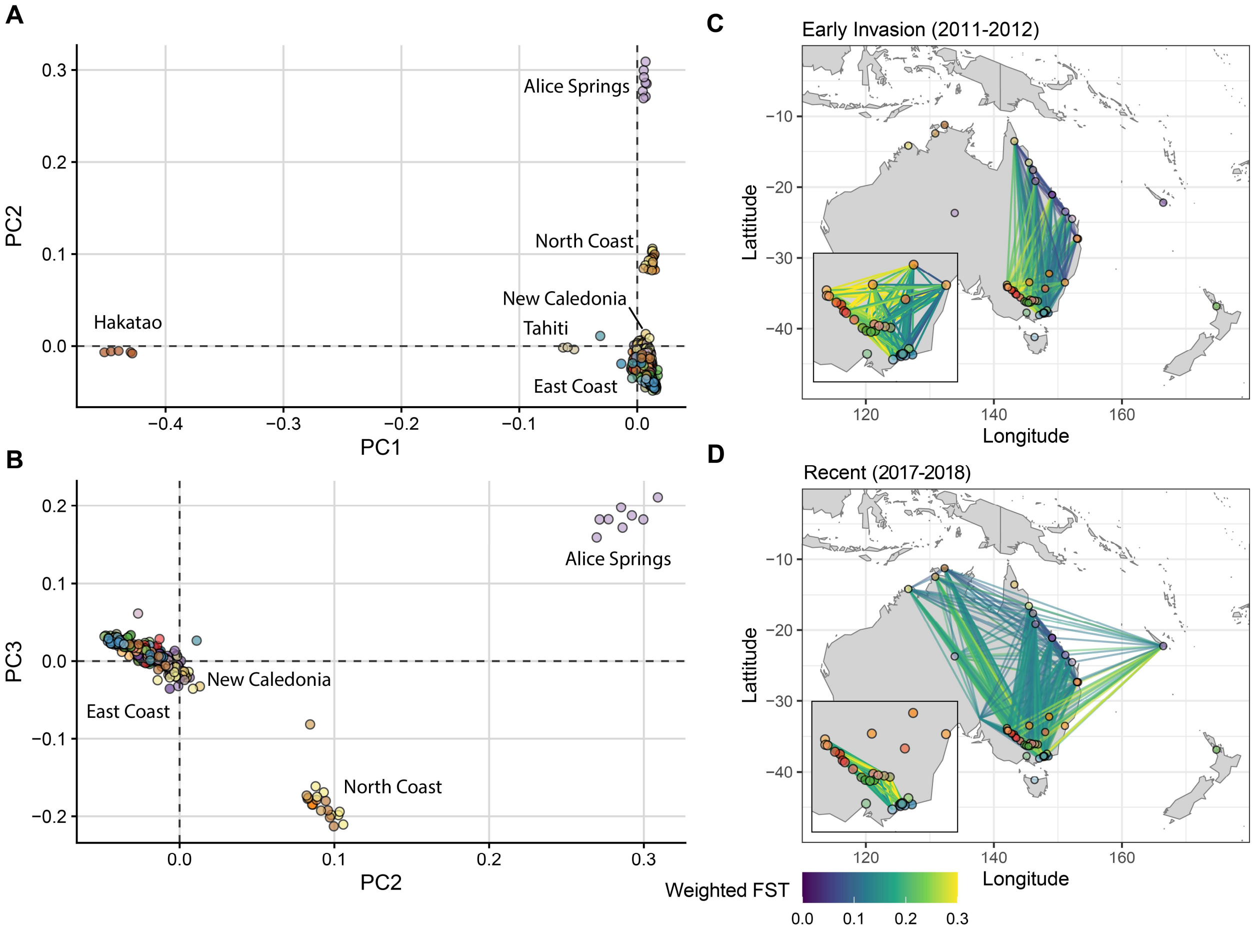


Figure 1: Individual-level principal components analysis showing samples clustering on **A)** axes 1 and 2 and **B)** axes 2 and 3, with main clusters labelled. **C)** Pairwise weighted fixation index (F_{ST}) between all sampled populations during early Victorian invasion (2011-2012) and **D)** more recently (2017-2018), displayed by geographic location.

in the training samples will propagate into model predictions, samples were considered outliers and removed if their km-error was >3 standard deviations above the mean of all other samples within a 500 km radius. The Locator model was then trained again on all remaining samples and used to predict the locations of the outbreak specimens (Table 1). All statistical analyses were conducted within R4.3 (R Core Team, 2019) using the *tidyverse* (Wickham et al., 2019) and *tidymodels* (Kuhn & Wickham, 2020) packages, with figures plotted using ggplot2 (Wickham, 2016).

Results

Sequencing results

Sequencing of the 108 specimens in the initial 'moderate coverage' dataset yielded a mean 12.6× depth of coverage per individual (± 0.786, range 0.55×-32.6×) across the 31,960 contigs of the B. *tryoni* v2.2 draft reference genome. 1,272,540 biallelic SNPs were called from this initial dataset and stringently filtered down to a set of 340,612 high confidence variants that were used for BSQR (Supplementary Fig. 1). The further 377 specimens sequenced in the 'low coverage' dataset yielded a mean 2.03× depth of coverage (± 0.147, range 0.005×-21×), but only the 316 samples that obtained >0.2× mean depth were retained. Genotype likelihoods were calculated for all variant sites with <50% missing data across both low and moderate coverage datasets, yielding 2,462,560 biallelic SNPs with a MAF >1%, of which 1,730,779 SNPs were found to be unlinked. Removing variants within the top 5% of windowed F_{ST} comparisons left 607,012 SNPs in the putatively 'neutral' dataset (Supplementary Fig. 2). Two full-sibling and 9 half-sibling dyads were identified within this dataset (Supplementary Fig. 3), and within each dyad the individual with the lowest sequencing coverage was removed to avoid biasing population structure inferences.

Principal component analyses

Principal component analysis of the remaining 347 individuals revealed specimens broadly clustered by their collection location (Fig. 1A), rather than year of collection or other technical factors (Supplementary Fig. 3). All specimens collected from both the endemic East Coast and Victorian invasive range formed a single large cluster, while those collected from the North Coast (NT and WA) formed a discrete smaller cluster separated along PC2 (Fig. 1A). The geographically isolated populations of Hakatao and Alice springs showed a large degree of separation from the larger coastal populations along PC1 and PC2 respectively, with specimens from Tahiti clustering intermediately between the East Coast and Hakatao (Fig. 1A). While the North Coast samples clustered intermediately between the East Coast and Alice springs specimens on PC2, these locations were separated along PC3 (Fig. 1B). In contrast to the other isolated island populations, specimens collected from New Caledonia clustered with specimens from the East Coast across PC axes 1, 2 and 3 (Fig. 1A, B).

Differentiation between early populations

For the specimens collected during 2011/12, genetic differentiation between sites along the East Coast endemic range was low, even between the most geographically separated collection locations of Coen and Sydney, ~2800 km apart (Fig. 1C, Supplementary Table 1).

Substantial genetic differentiation was, however, seen between the East Coast endemic range and inland locations within the Victorian invasive range (F_{ST} 0.2-0.3), but not coastal Victorian populations in Gippsland (F_{ST} 0.07-0.09) (Fig. 1C). Many of the inland Victorian sites were also differentiated from each other (F_{ST} 0.04-0.22), particularly at the western fringe of the invasive range where even the adjacent Ellerslie and Merbein sites were distinct (F_{ST} 0.19). Most inland Victorian populations were also highly differentiated from the inland NSW populations of Hillston, Dubbo, and Cootamundra (F_{ST} 0.17-0.37), as well as coastal Gippsland populations (F_{ST} 0.2-0.3), the latter being more similar to NSW populations (F_{ST} 0.02-0.05; Fig. 1C). Mantel tests found no significant correlation between geographic distance and distance in PC1 and PC2 latent space (r = 0.08, p = .053), or pairwise F_{ST} (r = -0.3, p > .05) across all early samples. When just the early samples from the NSW and Victorian invasive range were considered, a small but statistically significant correlation was found between geographic distance and both PCA latent space (r = 0.17, p = .012) and pairwise F_{ST} (r = 0.305, p = .018).

Differentiation between recent populations

Samples collected more recently from the Victorian invasive range (2017-2018) were more similar to the East Coast native range than at the earlier timepoint (F_{ST} 0.03-0.11), and comparable to that seen between the major East coast and North coast populations (F_{ST} 0.02-0.09) (Fig. 1D, Supplementary Table 2). The differentiation seen between early inland Victorian and NSW populations has since reduced considerably (F_{ST} 0.01-0.16), with the earlier distinction between Gippsland and inland populations also no longer present (F_{ST} 0.03-0.12). Despite this broad homogenisation of genetic diversity across the invasive range, populations collected from the VIC-NSW border sites of Wodonga, Yarrawonga and Rutherglen showed increased differentiation with the rest of inland Victoria (F_{ST} 0.14-0.3) compared to the earlier collections (F_{ST} 0.08-0.16; Fig. 1C, D). For the isolated invasive populations only sampled at the later timepoint, both the central Australian population of Alice springs as well as the French Polynesian island population of Hakatao showed the greatest differentiation from all other locations (F_{ST} 0.25-0.35; Fig. 1C, D). Specimens from the other French Polynesian island of Tahiti showed less differentiation with the east coast endemic range (F_{ST} 0.12-0.15), but similarly high with the VIC invasive range (F_{ST} 0.25-0.35), while the closer island of New Caledonia showed relatively little differentiation

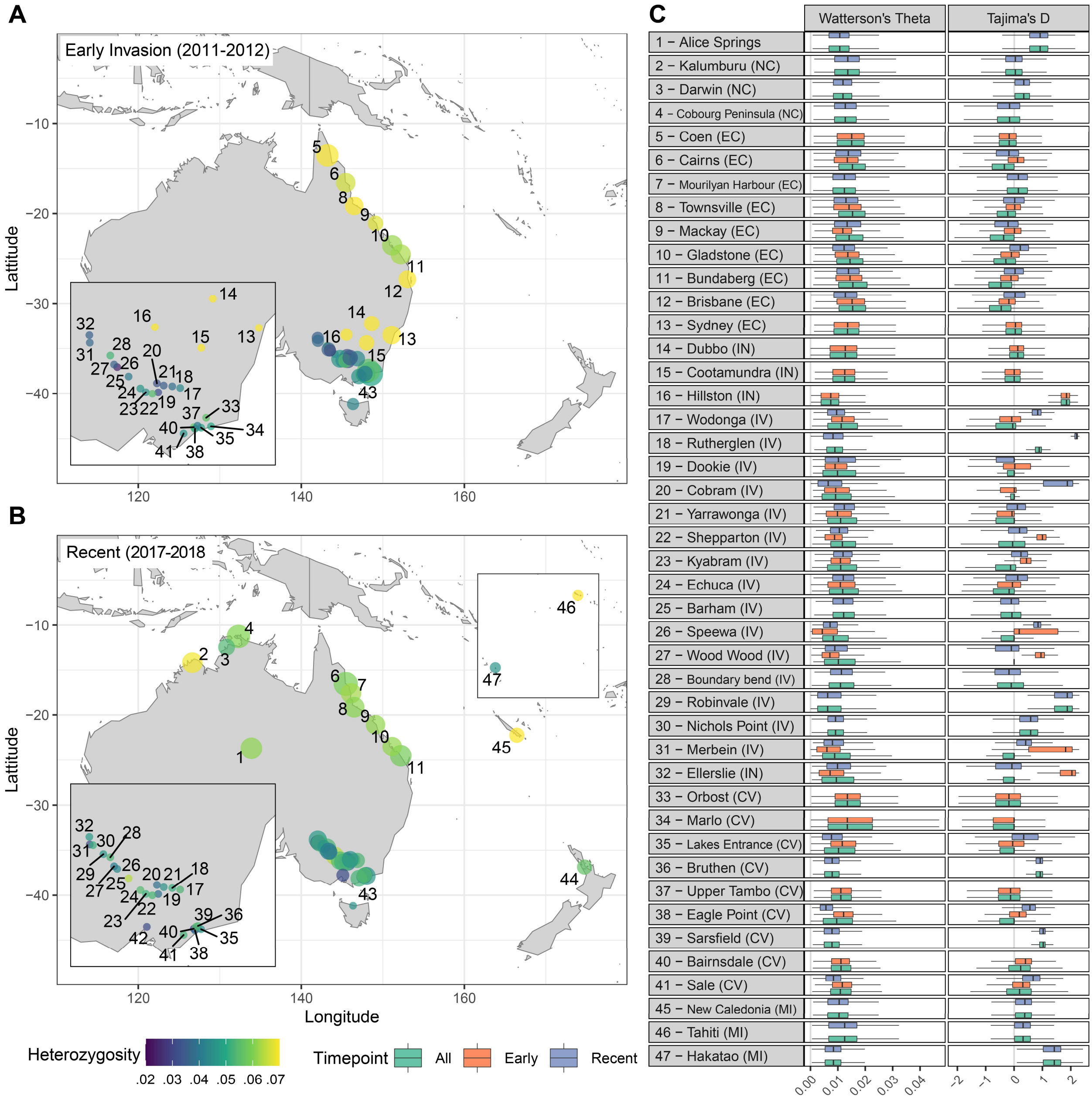


Figure 2: Average genome-wide Heterozygosity (H_E) for each collection location during **A)** early Victorian invasion (2011-2012) and **B)** more recently (2017-2018). **C)** Distribution of Watterson's Theta (θ_W) and Tajima's D across 15 kb genomic windows per population for early, recent, and combined samples. Population ID numbers correspond to Table 1. Abbreviations: NC; North Coast, EC; East Coast, IN; Inland NSW, IV; Inland Victoria, CV; Coastal Victoria, MI; Melanesian islands.

from the East Coast (F_{ST} 0.09-0.14). Across all the recent samples, Mantel tests found significant correlation between geographic distance and distance in PCA latent space (r = 0.74, p < .001), as well as pairwise F_{ST} (r = 0.31, p < .001). When the highly differentiated isolated populations were removed from the analysis the correlation between geographic and PCA distance was reduced but still significant (r = 0.65, p < .001), while the correlation with F_{ST} was no longer significant (r = -0.2, p > .05). When just the recent samples from the NSW and Victorian invasive range were considered, Mantel tests found a significant correlation between geographic distance and PCA distance ($r = 0.42$, $p < .001$), but not with F_{ST} ($r = 0.09$, $p = .018$).

Genetic diversity

Heterozygosity (H_E) within the early samples was a uniformly high 0.06-0.07 along the East Coast native range into NSW but saw a rapid reduction to only 0.02-0.04 in the Victorian invasive range (Fig. 2A). While H_E has broadly increased across Victoria in the more recent samples, it remains below that of the East Coast (Fig. 2B). H_E within the early Victorian invasive range was highest in coastal Gippsland populations ($H_E = 0.058$), and decreased with both distance from Gippsland (linear regression; $R^2 = .2$, $p < .001$), and the major southern population of Sydney ($R^2 = .18$, $p < .001$). This was not, however, reflected in the recent samples where the more inland population of Barham showed the highest H_E, and no relationship was found between distance from Gippsland ($R^2 = .012$, $p > .05$) or Sydney ($R^2 = .001$, $p > .05$). Despite the isolated populations showing similar H_E to the East Coast populations (Fig. 2B), all showed a strong positive skew in the genome wide distribution of Tajima's D, with Alice Springs and Hakatao being particularly high (Fig. 2C). Positive genome-wide values of *D* arise from a lack of rare alleles, characteristic of a sudden population contraction, and similar patterns were observed across many of the southern invasive range populations (Fig. 2C). Locations with a strongly positive skew of *D* at the early timepoint, but less so in the recent samples included Shepparton, Kyabram, Wood Wood, Merbein, and Ellerslie, suggesting that early population bottlenecks have been partially erased through continued immigration. On the other hand, the sites of Wodonga and Cobram showed increased *D* in 2017 compared to 2011, potentially indicating die-off and recent recolonisation by a limited number of founding individuals. For the major East- and North-Coast populations θ_w was consistently high and *D* either neutral or slightly negative across both timepoints, indicating that population sizes have remained stable or slightly expanded in the native range since the early collections in 2011/12 (Fig. 2C).

Admixture

A four ancestral population (K=4) model was chosen as most parsimonious from the MAP test, and under this model the East Coast and North Coast populations were determined

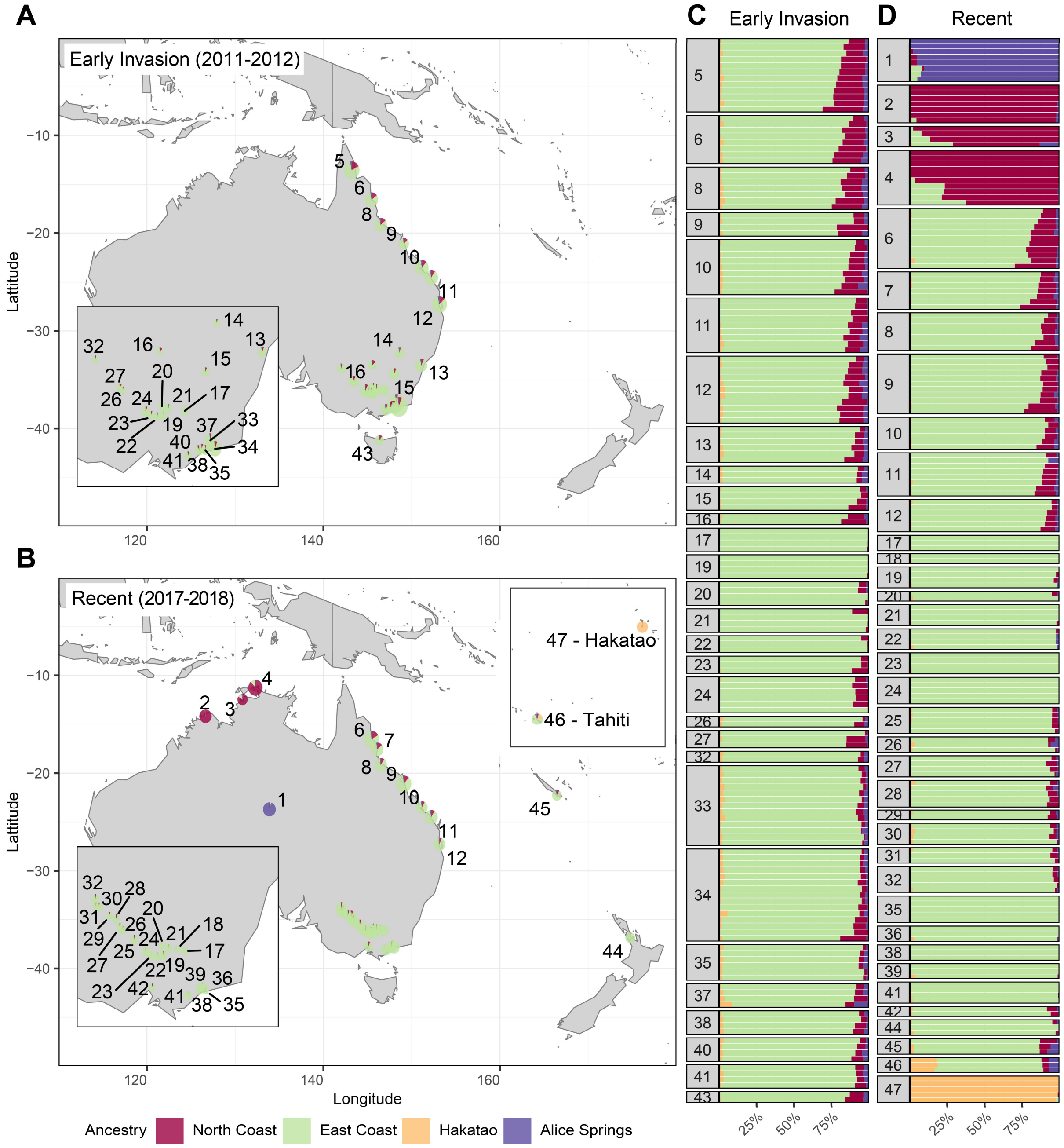


Figure 3: Population admixture proportions for each collection location from **A)** early Victorian invasion (2011-2012), and **B)** more recently (2017-2018). Insets show greater detail of the Victorian invasive range. **C)** Individual admixture proportions for each specimen collected during the early and **D)** recent samples. Population ID numbers correspond to Table 1.

to be the two main ancestry groups, with Alice Springs and Hakatao representing minor ancestry components (Fig. 3). For the North Coast populations, the WA location of Kalumburu was the most ancestral, while individuals collected from the more eastward NT locations of Darwin and Cobourg Peninsula showed up to 35% East Coast ancestry. Similarly, the more northern populations within the East Coast endemic range showed between 20 and 35% North Coast ancestry, indicating bi-directional gene flow between the NT and the East Coast. While the North Coast populations were only sampled at the recent timepoint, the presence of their ancestry within specimens collected along the East Coast was consistent across both timepoints, and steadily reduced with latitude until reaching complete absence at ~35°S in the specimens from Wodonga, Rutherglen and Dookie at the NSW/VIC border (Fig. 3A, B). While specimens collected from inland VIC showed almost no North Coast ancestry at both timepoints, those from the coastal Gippsland region saw a reduction from 10% North Coast ancestry in the 2011 to none in 2018 (Fig. 3C, D). Ancestral Alice Springs and Hakatao alleles on the other hand were practically absent from all other populations, with exception of Tahiti which contained 20% Hakatao ancestry (Fig. 3D). Despite both Alice Springs and Hakatao being identified as 'ancestral', the very high value of *D* for these isolated populations suggests that the admixture analysis is instead capturing the effects of a strong population bottleneck and subsequent genetic drift. Therefore, rather than the ~20% Hakatao ancestry seen in the Tahiti population being a real sign of admixture, this may instead reflect the extreme bottleneck undergone by the Hakatao populations where only ~20% of the allelic diversity present within the Tahiti population successfully colonised the more distant island of Hakatao.

Spatial assignment

During Locator cross validation, a random 10% (32-38) of individuals with known origins were iteratively dropped out of the training set and their geographic locations re-predicted, returning a median error of 298 km between the expected and predicted locations (95% quantiles: 85-1777 km, Fig. 4B). Prediction accuracy was relatively consistent across each separate 10k SNP window with no significant outlier windows observed (Supplementary Figure 4), however, the confidence contours derived from these windowed predictions only marginally captured the true uncertainty of inferences: in only 20% of predictions was the true location contained within the 95% confidence contour, 44.4% within the 50% contour and 55.3% within the 10% contour (Fig. 4C). Hakatao and Tahiti showed the highest CV error, reflecting their geographic isolation from the rest of the dataset, however the locations of Darwin and Cobourg peninsula on the north coast and Hillston in inland NSW also showed relatively high error (Fig. 4A). In contrast, specimens collected within the intensively sampled Victorian invasive range all showed the least error (Fig. 4A), with predictions deviating 50-400 km from known

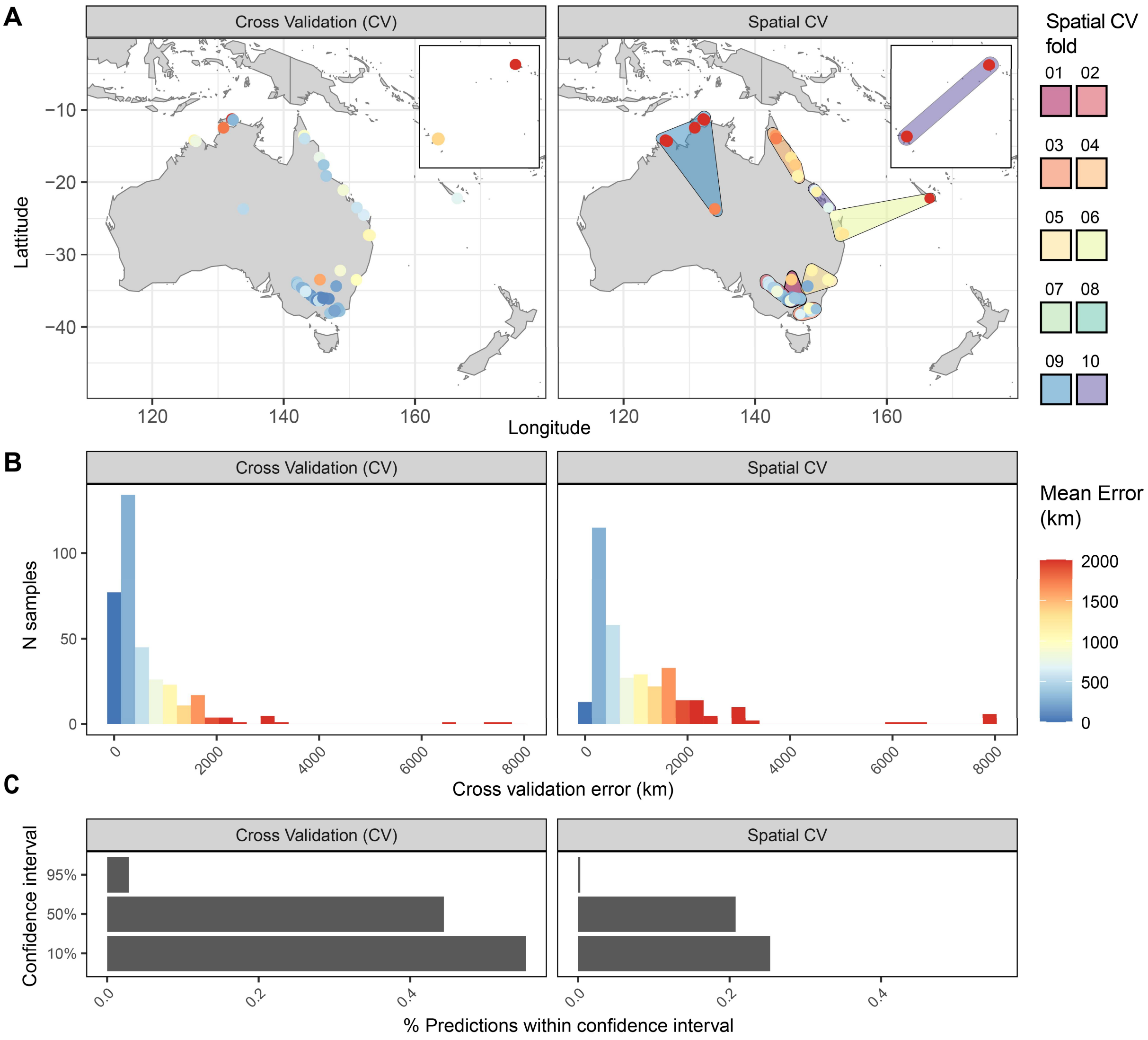


Figure 4: A) Mean cross validation error for Locator predictions summarised by collection location for both the conventional cross-validation (left panels) and spatial CV folds (right panels). **B)** Histogram of cross validation error by individual sample, and **C)** number of times the true location was within the 95%, 50% and 10% confidence interval for each CV procedure.

collection locations. Predictive accuracy across the 10 spatial CV folds was substantially lower than the conventional CV, returning a median 625 km-error (95% quantiles: 157-3011 km), with only 0.2% of predictions having the true locations within the 95% confidence interval, 20.8% within the 50% confidence interval and 25.4% within the 10% confidence interval (Fig. 4). The North Coast and Melanesian island populations showed the highest spatial CV error, with predictions for these samples being off by more than 2,000 km (Fig. 4A).

Following cross-validation, the model was re-trained using all samples then used to predict the most probable geographic origin of outbreak specimens. All four successfully sequenced individuals from the 2015 Auckland outbreak were assigned to the northern Victorian invasive range, with point estimates around north-eastern Victoria (Fig. 5). However, the 50% confidence contour for three of these samples encompassed most of Northern Victoria and NSW, and 10% confidence contour almost the entire east coast. This was not the case for the remaining sample, where the 95% and 50% confidence contours encompassed only a small set of towns in northern Victoria, but the 10% contour also included coastal Gippsland (Figure 5). The two samples collected from the Tasmanian outbreak in 2011 (VAITC2086 & VAITC2087) were also placed towards central Victoria, but the confidence intervals around these predictions were more geographically constrained (Fig. 5). The single successfully sequenced sample (VAITC7710) from the Tasmanian outbreak in 2018 was placed to a similar area, with the 50% confidence contour covering North-Central Victoria, and the 10% contour further including populations in coastal Gippsland. Finally, the 2 successfully samples from the 2018 outbreak in the Yarra Valley, Victoria were also placed to the southern invasive range, with point estimates and 95% confidence contours around North-Central Victoria, and 50% and 10% contours covering most of the inland Victorian invasive range (Fig. 5).

Discussion

The demographic history of Queensland fruit fly can be characterised by two large ancestral populations residing on the East and Northern coasts of Australia, of which the former has been the major source for both the southern range expansion and isolated island populations in Melanesia. Despite the massive geographic scale of the East Coast population, there is little differentiation seen between geographically distant locations and no evidence for isolation-by-distance. This indicates that despite the distances involved, migration and gene flow must occur at a rate sufficient to overcome any regional differentiation and local adaptation that would otherwise occur (Wright, 1931, 1943). This could be due to a combination of natural and anthropogenic factors; for instance temperatures within the endemic range remain high enough for breeding to occur year-round (Clarke et al., 2011; Meats & Fay, 2000), and regular dispersal from breeding sites is likely to play a role in maintaining genetic homogeneity on a local scale (Fletcher, 1973, 1974). At a regional scale, this natural dispersal would act in concert with long-distance human mediated movement to produce the pattern of apparent panmixia

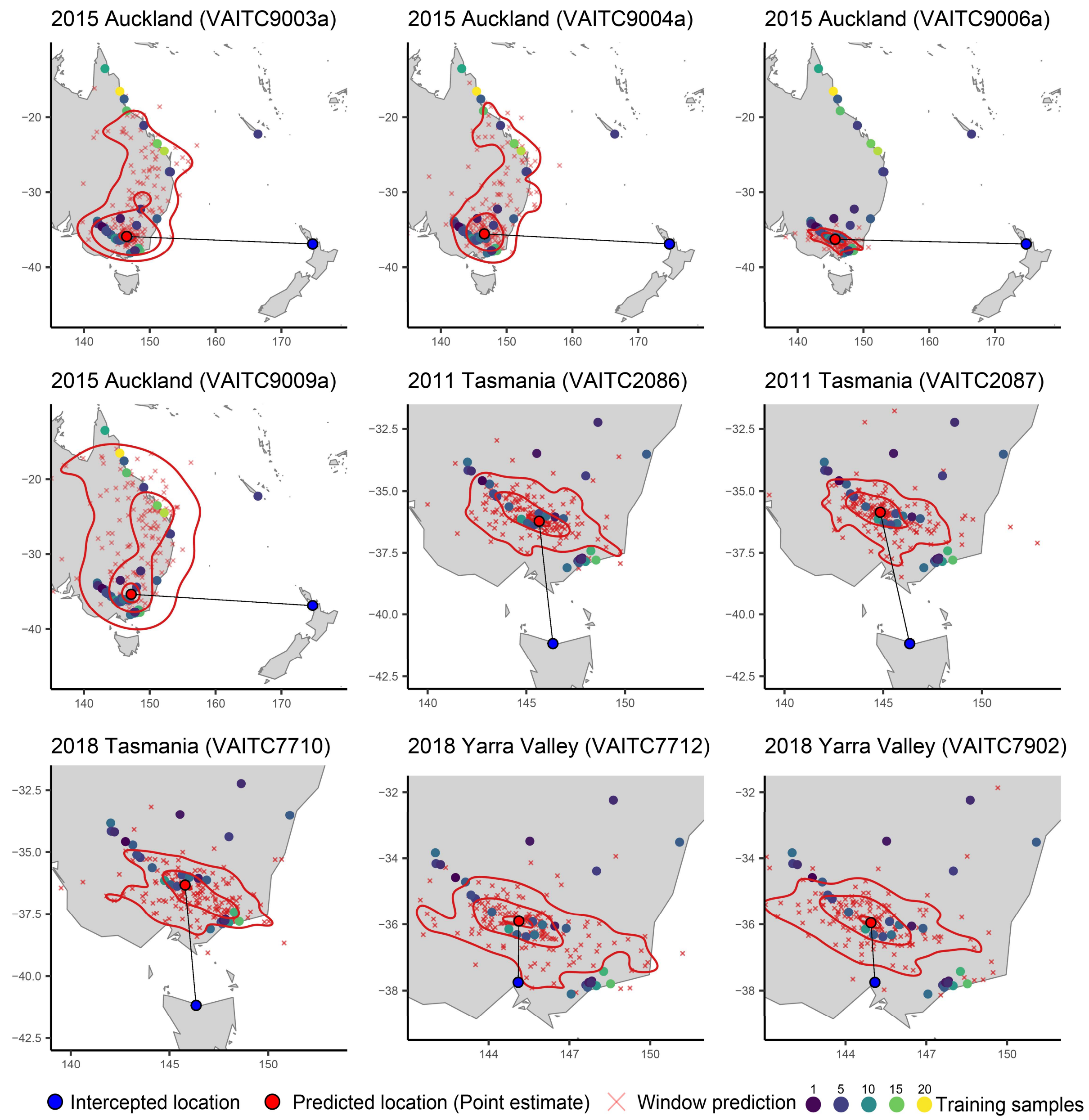

2015 Auckland (VAITC9003a)
2015 Auckland (VAITC9004a)
2015 Auckland (VAITC9006a)
2015 Auckland (VAITC9009a)
2011 Tasmania (VAITC2086)
2011 Tasmania (VAITC2087)
2018 Tasmania (VAITC7710)
2018 Yarra Valley (VAITC7712)
2018 Yarra Valley (VAITC7902)
−20
−30
−40
140
150
160
170
−32.5
−35.0
−37.5
−40.0
−42.5
145
−32
−34
−36
−38
144
147
Intercepted location
Predicted location (Point estimate)
Window prediction
1
5
10
15
20
Training samples

Figure 5: Predicted geographic origins of specimens intercepted from recent outbreaks, with their associated 95%, 50% and 10% confidence contours.

seen southwards until Sydney. On the Northern coast of Australia, a distinct ancestral population extends across both NT and northern WA, geographically separated from the East Coast population by an arid zone between NT and QLD (Drew, 1989). Whether this North Coast population represents the sibling species *B. aquilonis*, a hybrid *B. aquilonis-tryoni*, or whether *B. aquilonis* ever existed has been debated since the late 1980s, when populations around Darwin underwent a 10-fold increase in host range (Cameron et al., 2010; Drew & Lambert, 1986; Morrow et al., 2000; Smith et al., 1988). While our study does not attempt to resolve this taxonomic question, we provide evidence for both substantial genetic diversity in the North Coast population and a large degree of differentiation from the East Coast, indicating that it considerably predates the first collection record of 1953 (Cameron, 2006). On the other hand, our admixture analyses found bi-direction gene flow between the North Coast locations of Darwin and Cobourg Peninsula and the northernly populations on East Coast, leaving recent hybridisation between the two as a plausible cause for the increased pestiferousness of North Coast populations (Morrow et al., 2000; Osborne et al., 1997; Yu et al., 2001). Future studies should examine the genomic context around these introgressed East Coast alleles to determine if they could play an adaptive role in host selection, or whether the recent change in pest status was simply a behavioural shift due to increased host availability. Integrating historical specimens of North Coast B. *tryoni* into this dataset using 'museum genomics' approaches (Mikheyev et al., 2017) may further assist in resolving the taxonomic validity of B. *aquilonis*, an issue that continues to impact trade and research (Clarke et al., 2011) .

Under a stepping-stone model of range expansion, consecutive founder effects should lead to a steady decrease of genetic diversity into the invasive range (Austerlitz et al., 1997). While populations within the early Victorian invasive range showed low H_E and high values of D characteristic of founder effects, there was little evidence for isolation-by-distance and instead genetic differentiation between adjacent sites was often as great as those at opposite ends of the invasive range. This patchy distribution of allele frequencies suggests early invasive populations were separately founded by long distance dispersal of small numbers of specimens, which were then subjected to strong genetic drift causing localised random fixation of different alleles (Ibrahim et al., 1996; Nichols & Hewitt, 1994). This pattern matches that seen in the mitochondrial haplotype data from Blacket et al., (2017) who separately analysed many of the specimens used in our study, yet the larger genomic and geographic context presented here reveals that despite their distinct mitochondrial haplotypes, all these outbreaks arise from the same East Coast range expansion. Since these early samples were collected, however, H_E has broadly increased across inland Victoria, suggesting an ongoing consolidation of genetic diversity through further immigration from the primary range expansion. The main exceptions to this pattern were the populations of Cobram, Rutherglen, and Wodonga in North-central Victoria, as well as Bruthen and Sarsfield in coastal Gippsland, which all showed recent evidence of population bottlenecks. This may indicate that rather than having an established resident population, these areas may still be undergoing regular local extinction and re-colonisation as recently as 2018. This is despite *B. tryoni* recently being declared endemic within Victoria, and the fruit fly exclusion zone which used to cover the north east of the state being repealed in favour of area wide management practices (Dominiak & Mapson, 2017). The potential recent replacement of Gippsland populations is of particular interest, as *B. tryoni* has been present in this area since the 1960s, the earliest recorded in Victoria and thus most likely to have undergone local adaptation (O'Loughlin, 1964; O'Loughlin et al 1984). Furthermore, the relative proportion of North Coast ancestry within many Victorian populations, and Gippsland in particular, has reduced since the early timepoint, suggesting the presence of a more ancestrally East Coast population on the NSW side of the border, unsampled in our study, which may have acted as a source for more recent incursions into Victoria. Taken together, these patterns indicate that the mode of colonisation in Victoria combines aspects of both mainland/island and stepping-stone dynamics, where satellite colonies generated through long-distance dispersal remain in isolation only for a short period before either dying off, or coalescing with their slowly expanding parent population (Shigesada & Kawasaki, 2002).

Continued die-off and recolonisation within the Victorian invasive range could be due to a combination of bioclimatic stressors and localised pest control efforts undertaken by government agencies and individual growers. Tolerance of extreme low temperatures and desiccation stress are considered key factors restricting the distribution and abundance of *B. tryoni* (Meats, 1981; O'Loughlin et al., 1984; Yonow & Sutherst, 1998), and in northern Victoria where winter frosts are common, considerable selection pressure would be expected to increase cold tolerance and capacity for adults and pupae to overcome the critical winter 'breeding gap' (Clarke et al., 2019; Gilchrist & Meats, 2010). However, continued immigration from larger subtropical populations which have not been subjected to the same selective pressures may oppose local adaptation, particularly if seasonal extinction of invasive populations continues to occur. Our windowed F_{ST} comparisons found a substantial portion of the genome to be highly differentiated between populations, and future work should explore the genomic context of these outlier regions to determine if they may play a role in climatic adaptation, or are simply driven to high frequencies through phenomena such as allele surfing (Klopfstein et al., 2006). Nevertheless, colonisation of the temperate Victorian environment may not require an adaptive hypothesis and instead could be explained by a combination of microclimatic features and adult behavioural traits. While *B. tryoni* pupae do not diapause, in a recent review Clarke et al. (2019) highlighted the ability for adult flies to survive cold winters by sheltering in dense bushland, or by taking advantage of the urban heat island effect (Dominiak et al., 2006; Yonow & Sutherst, 1998). Fletcher (1974) also proposed that populations in drier temperate regions can persist through short distance movements between orchards and nearby water sources, and uptake of irrigation systems within Victoria has dramatically increased since these early observations (Millar & Roots, 2012). Therefore, while the recent range expansion may well be an evolutionary novel situation for *B. tryoni*, it does not necessarily mean the species should be considered maladapted to temperate regions. In fact, *B. tryoni* is now considered rare in much of its native rainforest compared to more artificial habitats in peripheral and suburban areas (Ero, 2009; Raghu et al., 2000; Zalucki et al., 1984), and future studies should investigate fine-scale patterns of gene flow along the urban-rural gradient to clarify whether urban overwintering could be taking place followed by seasonal dispersal into crops.

In addition to mainland Australia, *B. tryoni* is considered invasive to the Melanesian islands of New Caledonia, French Polynesia, Pitcairn, and Cook (Clarke et al., 2011). Popa-Baez and colleagues (2020) found that colonisation of Melanesian islands followed a stepping stone pattern, where flies were introduced from the east coast of Australia into New Caledonia (~1600 km), which in turn became the source of migrants to the more distant Tahiti (~4700 km). Our study additionally included the island population of Hakatao (1,359 km north west of Tahiti), where we found similar patterns of reduced genetic diversity and substantial differentiation from mainland Australian ancestors. However, while our data supported Hakatao populations arising from the nearby Tahiti, we found no evidence to support Tahiti populations being introduced from New Caledonia and this may instead represent a separate introduction from the east coast of Australia. Similar to the Melanesian islands, the isolated Central Australian population of Alice springs also showed reduced genetic diversity and a high value of *D*, fitting with the theory that this population was founded by a small number of flies, potentially around 1987 (Cameron, 2006). The geographic origin of this population remains elusive, however, as it showed a high degree of differentiation from all major East- and North Coast endemic populations. For all of these isolated populations, further investigation using alternative statistical approaches such as Approximate Bayesian Computation may be required to resolve the conflicting introduction scenarios (Estoup & Guillemaud, 2010), but these methods will need further validation to ensure they remain fit-for-purpose within the constraints of low-coverage datasets. From a management perspective, given many of these isolated populations show no evidence of recent gene flow with major endemic populations they may be prime targets for future eradication efforts (Suckling et al., 2016).

The Locator model predicted the inland Victorian invasive range to be the source for not only a local outbreak in the Yarra Valley, but also interstate and international outbreaks in Tasmania and Auckland, NZ. Many of these predictions should be interpreted with caution, however, due to their wide confidence intervals and the large errors seen between model predictions and true locations during CV. The use of genomic windows based upon physical distances, or preferably recombination-based distances, to measure prediction confidence was recommended by the original Locator paper (Battey et al., 2020), but the highly fragmented *B. tryoni* reference genome instead required the use of SNP-based windows in our study. While this meant the number of variants remained consistent between windows, the physical length of each window varied substantially, with some covering multiple unplaced contigs and potentially, multiple chromosomes. Much like an admixed individual, Locator is thus attempting to model a patchwork of different evolutionary and geographic histories within each window, which may have impacted prediction accuracy (Battey et al., 2020). Furthermore, while our study covered almost the entirety of the known *B. tryoni* range (with exception of the distant Pitcairn and Cook Islands), the sampling was heavily biased towards the Victorian invasive range, at the expense of other key locations such as NSW. While Locator has been shown to interpolate unsampled locations reasonably well if allele frequencies change smoothly over the landscape (Battey et al., 2020), the patchiness of genetic differentiation across Victoria indicates that this may not hold for *B. tryoni*. Therefore, if these outbreaks actually originated from NSW or another unsampled location, it is possible that Locator may erroneously project them towards the nearest and most densely sampled geographic area in the training set, in this case Northern Victoria. Due to these confounding factors, it is worth comparing our results to the study of Popa-Báez et al. (2021) who analysed separate specimens from the same 2015 Auckland and 2018 northern Tasmanian outbreaks, assigning both to the 'East Coast of Australia'. While it is possible that the specimens analysed here represent distinct sub-groups within mixed introduction events, the discrepancy between our studies is more likely related to differences in geographic coverage of training samples as well as the loci used for assignment. The reference panel used in Popa-Báez et al. (2021) comprehensively sampled the endemic East and North coast populations but only 9 individuals from Shepparton were included from the Victorian invasive range, while our reference panel instead contained a densely sampled Victorian invasive range and used substantially more SNPs for assignment (607,012 vs 2,361 to 2,428). Considering the somewhat conflicting results of our two studies, and the geographic limitations of the reference panels used in each, it may be best to consider the outbreaks from both Tasmania and Auckland as arising from somewhere along the East Coast of Australia, inclusive of the southern invasive range. While this conservative assignment greatly reduces the resolution of predictions, it still rules out alternative introduction scenarios, such as these outbreaks arising from North Coast, Alice springs, or Melanesian island populations.

Understanding the pathways and processes underlying colonisation and establishment by B. *tryoni* will become increasingly important as much of the southern invasive range transitions to area-wide management, and new incipient populations appear along the South Australian border (Florec et al., 2013; Jessup et al., 2007). The use of genomic sequencing to identify populations with limited genetic connectivity—and thus low recolonisation risk—will likely prove important for future control efforts such as with the Sterile Insect Technique (Raphael et al., 2014). While improvements in predictive accuracy will no doubt be required before geographic assignment techniques can be used in a management context, the importance of pathway tracing for increasing the likelihood of future pest exclusion, as well as the substantial costs incurred during quarantine and outbreak control should promote further investment in these methods. In particular, the development of international collaborative working groups to generate larger and more diverse reference collection, and the assembly of a chromosomal scale reference genome for B. *tryoni* would increase inferential power for both population structure and geographic placement of incursion samples. Ultimately, the complex dispersal patterns of B. *tryoni* and limited spatial-genetic structure means an integrated approach that takes advantage of multiple data sources including shipping and transport records, host information, genotypes, and trapping records may be required to effectively trace and respond to new outbreaks of this destructive horticultural pest.

Acknowledgements

We thank Richard Mapson (Agriculture Victoria) for providing B. *tryoni* specimens from Victoria, Bernie Dominiak for NSW specimens, Brian Thistleton (Northern Territory Department of Primary Industries and Resources) for NT and WA specimens, Jane Royer (Queensland Department of Agriculture and Fisheries) for Queensland specimens, and Jamie Davies (Department of Primary Industries, Parks, Water and Environment) for Tasmanian specimens. We extend this gratitude to our international colleagues Dongmei Lei (Ministry for Primary Industries) for providing NZ and Hakatao specimens, Christian Mille (Institut Agronomique néo-Calédonien) for New Caledonia specimens, and Julie Grandgirard (Direction de l'Agriculture) Tahiti specimens. We further thank Michelle Drayton, Preeti Thakur, Rebecca Bailie, Michelle Malmberg, and Tracie Webster (Agriculture Victoria Research) for their advice and assistance with genomic library preparation and sequencing. Funding for this study was provided by Agriculture Victoria's Improved Market Access for Horticulture programme (CMI105584) and A.M.P. is further supported by an Australian Government Research Training Program Scholarship.

Author contributions

A.M.P., N.O.I.C and M.J.B. conceptualised the study, A.M.P performed all laboratory procedures, bioinformatic and statistical analyses, and wrote the first draft of the manuscript with input and supervision from J.P.C, N.O.I.C, and M.J.B. All authors read and approved the final manuscript.

Data Archiving Statement:

Demultiplexed sequence reads are available at NCBI SRA acc no: (XXXXX, to be assigned later) while aligned BAM files can be obtained from Dryad acc no: (XXXXX, to be assigned later). All code required to reproduce the bioinformatic and statistical analyses presented in this manuscript is available at the following GitHub repository: https://github.com/alexpiper/genomeskim

References

Alex Buerkle, C., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: How low should we go? *Molecular Ecology*, 22(11), 3028–3035. https://doi.org/10.1111/mec.12105

Arora, K., Shah, M., Johnson, M., Sanghvi, R., Shelton, J., Nagulapalli, K., Oschwald, D. M., Zody, M. C., Germer, S., Jobanputra, V., Carter, J., & Robine, N. (2019). Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms. *Scientific Reports*, 9, 19123. https://doi.org/10.1038/s41598-019-55636-3

Austerlitz, F. De, Jung-muller, B., Godelle, B., & Gouyon, P. (1997). Evolution of Coalescence Times, Genetic Diversity and Structure during Colonization. *Theoretical Population Biology*, 51, 148–164.

Barr, N., Ruiz-Arce, R., & Armstrong, K. (2014). Using molecules to identify the source of fruit fly invasions. In *Trapping And The Detection, Control, and Regulation of Tephritid Fruit Flies: Lures, Area-Wide Programs, and Trade Implications*. https://doi.org/10.1007/978-94-017-9193-9_10

Battey, C., Ralph, P. L., & Kern, A. D. (2020). Predicting geographic location from genetic variation with deep neural networks. *ELife*, 9, e54507. https://doi.org/10.7554/elife.54507

Blacket, M. J., Malipatil, M. B., Semeraro, L., Gillespie, P. S., & Dominiak, B. C. (2017). Screening mitochondrial DNA sequence variation as an alternative method for tracking established and outbreak populations of Queensland fruit fly at the species southern range limit. *Ecology and Evolution*, 7(8), 2604–2616. https://doi.org/10.1002/ece3.2783

Blacket, M. J., Semeraro, L., & Malipatil, M. B. (2012). Barcoding Queensland Fruit Flies (Bactrocera tryoni): Impediments and improvements. *Molecular Ecology Resources*, 12(3), 428–436. https://doi.org/10.1111/j.1755-0998.2012.03124.x

Bradburd, G. S., & Ralph, P. L. (2019). Spatial Population Genetics: It's About Time. *Annual Review of Ecology, Evolution, and Systematics*, 50(1), 427–449. https://doi.org/10.1146/annurev-ecolsys-110316-022659

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. *International Geoscience and Remote Sensing Symposium* (IGARSS), 5372–5375. https://doi.org/10.1109/IGARSS.2012.6352393

Cameron, E. C. (2006). *Fruit fly pests of Northwestern Australia* (*Ph.D. Thesis*). University of Sydney.

Cameron, E. C., Sved, J. A., & Gilchrist, A. S. (2010). Pest fruit fly (Diptera: Tephritidae) in northwestern Australia: One species or two? *Bulletin of Entomological Research*, 100(2), 197–206. https://doi.org/10.1017/S0007485309990150

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Clarke, A. R., Merkel, K., Hulthen, A. D., & Schwarzmueller, F. (2019). Bactrocera tryoni (Froggatt) (Diptera: Tephritidae) overwintering: an overview. *Austral Entomology*, 58(1), 3–8. https://doi.org/10.1111/aen.12369

Clarke, A. R., Powell, K. S., Weldon, C. W., & Taylor, P. W. (2011). The ecology of Bactrocera tryoni (Diptera: Tephritidae): What do we know to assist pest management? *Annals of Applied Biology*, 158(1), 26–54. https://doi.org/10.1111/j.1744-7348.2010.00448.x

Cristescu, M. E. (2015). Genetic reconstructions of invasion history. *Molecular Ecology*, 24(9), 2212–2225. https://doi.org/10.1111/mec.13117

De-Kayne, R., Frei, D., Greenway, R., Mendes, S. L., Retel, C., & Feulner, P. G. D. (2021). Sequencing platform shifts provide opportunities but pose challenges for combining genomic data sets. *Molecular Ecology Resources*, 21, 653–660. https://doi.org/10.1111/1755-0998.13309

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491–501. https://doi.org/10.1038/ng.806

Diniz-Filho, J. A. F., Soares, T. N., Lima, J. S., Dobrovolski, R., Landeiro, V. L., Telles, M. P. de C., Rangel, T. F., & Bini, L. M. (2013). Mantel test in population genetics. *Genetics and Molecular Biology*, 36(4), 475–485. https://doi.org/10.1590/S1415-47572013000400002

Dominiak, B. C. (2012). Review of Dispersal, Survival, and Establishment of Bactrocera tryoni (Diptera: Tephritidae) for Quarantine Purposes. *Annals of the Entomological Society of America*, 105(3), 434–446. https://doi.org/10.1603/an11153

Dominiak, B. C., & Mapson, R. (2017). Revised Distribution of Bactrocera tryoni in Eastern Australia and Effect on Possible Incursions of Mediterranean Fruit Fly: Development of Australia's Eastern Trading Block. *Journal of Economic Entomology*, 110(6), 2459–2465. https://doi.org/10.1093/jee/tox237

Dominiak, B. C., Mavi, H. S., & Nicol, H. I. (2006). Effect of town microclimate on the Queensland fruit fly Bactrocera tryoni. *Australian Journal of Experimental Agriculture*, 46(9), 1239–1249. https://doi.org/10.1071/EA04217

Dominiak, B. C., Wiseman, B., Anderson, C., Walsh, B., McMahon, M., & Duthie, R. (2015). Definition of and management strategies for areas of low pest prevalence for Queensland fruit fly Bactrocera tryoni Froggatt. *Crop Protection*, 72, 41–46. https://doi.org/10.1016/j.cropro.2015.02.022

Drew, R. A. I. (1989). The tropical fruit flies (Diptera: Tephritidae: Dacinae) of the Australasian and Oceanian regions. *Memoirs of the Queensland Museum*, 26.

Drew, R. A. I., & Lambert, D. M. (1986). On the Specific Status of Dacus (Bactrocera) aquilonis and D. (Bactrocera) tryoni (Diptera: Tephritidae). *Annals of the Entomological Society of America*, 79(3), 870–878. https://doi.org/10.1093/aesa/79.6.870

Elton, C. S. (1958). *The Ecology of Invasions by Animals and Plants*. https://doi.org/10.1007/978-1-4899-7214-9

Ero, M. M. (2009). *Host Searching Behaviour of Diachasmimorpha kraussii (Fullaway) (Hymenoptera: Braconidae: Opiinae), a Polyphagous Parasitoid of Dacinae Fruit Flies (Diptera: Tephritidae) (PhD Thesis)*. Queensland University of Technology.

Estoup, A., & Guillemaud, T. (2010). Reconstructing routes of invasion using genetic data: Why, how and so what? *Molecular Ecology*, 19(19), 4113–4130. https://doi.org/10.1111/j.1365-294X.2010.04773.x

Fitzpatrick, B. M., Fordyce, J. A., Niemiller, M. L., & Reynolds, R. G. (2012). What can DNA tell us about biological invasions? *Biological Invasions*, 14(2), 245–253. https://doi.org/10.1007/s10530-011-0064-1

Fletcher, B. S. (1973). The Ecology of a natural population of the Queensland Fruit Fly, Dacus tryoni IV. The immigration and Emigration of Adults. *Australian Journal of Zoology*, 21(4), 437–475. https://doi.org/10.1071/ZO9730541

Fletcher, B. S. (1974). The ecology of a natural population of the Queensland fruit fly, dacus tryoni V. the dispersal of adults. *Australian Journal of Zoology*, 22(2), 117–129. https://doi.org/10.1071/ZO9740189

Florec, V., Sadler, R. J., White, B., & Dominiak, B. C. (2013). Choosing the battles: The economics of area wide pest management for Queensland fruit fly. *Food Policy*, 38(1), 203–213. https://doi.org/10.1016/j.foodpol.2012.11.007

Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). NgsLD: Evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*, 35(19), 3855–3856. https://doi.org/10.1093/bioinformatics/btz200

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS ONE*, 8(11), e79667. https://doi.org/10.1371/journal.pone.0079667

Gilchrist, A. S., Dominiak, B., Gillespie, P. S., & Sved, J. A. (2006). Variation in population structure across the ecological range of the Queensland fruit fly, Bactrocera tryoni. *Australian Journal of Zoology*, 54(2), 87–95. https://doi.org/10.1071/ZO05020

Gilchrist, A. S., & Meats, A. W. (2010). The genetic structure of populations of an invading pest fruit fly, Bactrocera tryoni, at the species climatic range limit. *Heredity*, 105(2), 165–172. https://doi.org/10.1038/hdy.2009.163

Gilchrist, A. S., Shearman, D. C. A., Frommer, M., Raphael, K. A., Deshpande, N. P., Wilkins, M. R., Sherwin, W. B., & Sved, J. A. (2014). The draft genome of the pest tephritid fruit fly Bactrocera tryoni: Resources for the genomic analysis of hybridising species. *BMC Genomics*, 15, 1153. https://doi.org/10.1186/1471-2164-15-1153

Gilchrist, A. S., Sved, J. A., & Meats, A. (2004). Genetic relations between outbreaks of the Queensland fruit fly, Bactrocera tryoni (Froggatt) (Diptera: Tephritidae), in Adelaide in 2000 and 2002. *Australian Journal of Entomology*, 43(2), 157–163. https://doi.org/10.1111/j.1440-6055.2003.00389.x

Han, E., Sinsheimer, J. S., & Novembre, J. (2014). Characterizing bias in population genetic inferences from low-coverage sequencing data. *Molecular Biology and Evolution*, 31(3), 723–735. https://doi.org/10.1093/molbev/mst229

Hanghøj, K., Moltke, I., Andersen, P. A., Manica, A., & Korneliussen, T. S. (2019). Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience*, 8(5), giz034. https://doi.org/10.1093/gigascience/giz034

Holt, R. D. (2003). On the evolutionary ecology of species' ranges. *Evolutionary Ecology Research*, 5(2), 159–178.

Ibrahim, K. M., Nichols, R. A., & Hewitt, G. M. (1996). Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity*, 77, 282–291.

Jessup, A. J., Dominiak, B., Woods, B., De Lima, C. P. F., Tomkins, A., & Smallridge, C. J. (2007). Area-Wide Management of Fruit Flies in Australia. In *Area-Wide Control of Insect Pests* (pp. 685–697). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6059-5_63

Klopfstein, S., Currat, M., & Excoffier, L. (2006). The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, 23(3), 482–490. https://doi.org/10.1093/molbev/msj057

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, 15, 356. https://doi.org/10.1186/s12859-014-0356-4

Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* https://www.tidymodels.org

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Preprint*, 1303.3997. http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, 19(3), 639–647. https://doi.org/10.1111/1755-0998.12995

Lodge, D. M. (2003). Biological invasions: Lessons for ecology. *Trends in Ecology & Evolution*, 8(4), 133–136.

Lou, R. N., Jacobs, A., Wilder, A., & Therkildsen, N. O. (2020). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Authorea Preprints*. https://doi.org/10.22541/au.160689616.68843086/v2

Matz, M. V. (2018). Fantastic Beasts and How To Sequence Them: Ecological Genomics for Obscure Model Organisms. *Trends in Genetics*, 34(2), 121–132. https://doi.org/10.1016/j.tig.2017.11.002

May, A. W. S. (1962). The Fruit Fly Problem in Eastern Australia. *Australian Journal of Entomology*, 1(1), 1–4. https://doi.org/10.1111/j.1440-6055.1962.tb00160.x

McInnis, D., Hendrichs, J., Shelly, T., Barr, N., Hoffman, K., Rodriguez, R., Lance, D., Bloem, K., Suckling, D., Enkerlin, W., Gomes, P., & TAN, K. (2016). Can Polyphagous Invasive Tephritid Pest Populations Escape Detection for Years Under Favorable Climatic and Host Conditions? *American Entomologist*, 63(2), 89–99.

Meats, A. (1981). The bioclimatic potential of the Queensland fruit fly, Dacus tryoni, in Australia. *Proceedings of the Ecological Society of Australia*, 11, 151–161.

Meats, A., & Fay, H. A. C. (2000). Distribution of mating frequency among males of the Queensland fruit fly, Bactrocera tryoni (Froggatt), in relation to temperature, acclimation and chance. *General and Applied Entomology*, 29, 27–30.

Meats, A., & Hartland, C. L. (1999). Upwind anemotaxis in response to cue-lure by the Queensland fruit fly, Bactrocera tryoni. *Physiological Entomology*, 24(1), 90–97. https://doi.org/10.1046/j.1365-3032.1999.00118.x

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, *210*(2), 719–731. https://doi.org/10.1534/genetics.118.301336

Meisner, J., & Albrechtsen, A. (2019). Testing for Hardy–Weinberg equilibrium in structured populations using genotype or low-depth next generation sequencing data. *Molecular Ecology Resources*, 19(5), 1144–1152. https://doi.org/10.1111/1755-0998.13019

Mikheyev, A. S., Zwick, A., Magrath, M. J. L., Grau, M. L., Qiu, L., Su, Y. N., & Yeates, D. (2017). Museum Genomics Confirms that the Lord Howe Island Stick Insect Survived Extinction. *Current Biology*, 27(20), 3157–3161. https://doi.org/10.1016/j.cub.2017.08.058

Millar, J., & Roots, J. (2012). Changes in Australian agriculture and land use: Implications for future food security. *International Journal of Agricultural Sustainability*, 10(1), 25–39. https://doi.org/10.1080/14735903.2012.646731

Morrow, J., Scott, L., Congdon, B., Yeates, D., Frommer, M., & Sved, J. (2000). Close genetic similarity between two sympatric species of tephritid fruit fly reproductively isolated by mating time. *Evolution*, 54(3), 899–910. https://doi.org/10.1111/j.0014-3820.2000.tb00090.x

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76(10), 5269–5273. https://doi.org/10.1073/pnas.76.10.5269

Nichols, R. A., & Hewitt, G. M. (1994). The genetic consequences of long distance dispersal during colonization. *Heredity*, 72(3), 312–317. https://doi.org/10.1038/hdy.1994.41

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, 7(7), e37558. https://doi.org/10.1371/journal.pone.0037558

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6), 443–451. https://doi.org/10.1038/nrg2986

North, H., Mcgaughran, A., & Jiggins, C. (2021). *The population genomics of invasive species*. 1–25. https://doi.org/10.22541/au.160968166.65928724/v1

O'Loughlin, G. T. (1964). The Queensland fruit fly in Victoria. *The Journal of Agriculture*, 62, 391–402.

O'Loughlin, G. T., East, R. A., & Meats, A. (1984). Survival, development rates and generation times of the Queensland fruit fly, Dacus tryoni, in a marginally favourable climate: Experiments in Victoria. *Australian Journal of Zoology*, 32(3), 311–318. https://doi.org/10.1071/ZO9840353

Osborne, R., Meats, A., Frommer, M., Sved, J. A., Drew, R. A. I., & Robson, M. K. (1997). Australian Distribution of 17 Species of Fruit Flies (Diptera: Tephritidae) Caught in Cue Lure Traps in February 1994. *Australian Journal of Entomology*, 36(1), 45–50. https://doi.org/10.1111/j.1440-6055.1997.tb01430.x

Plant Health Australia. (2018). *The Australian Handbook for the Identification of Fruit Flies*. Plant Health Australia.

Popa-Báez, Á. D., Catullo, R., Lee, S. F., Yeap, H. L., Mourant, R., Frommer, M., Sved, J. A., Cameron, E. C., Edwards, O. R., Taylor, P. W., & Oakeshott, J. G. (2020). Genome-wide patterns of differentiation over space and time in the Queensland fruit fly. *Scientific Reports*, 10, 10788. https://doi.org/10.1038/s41598-020-67397-5

Popa-Báez, Á. D., Lee, S. F., Yeap, H. L., Westmore, G., Crisp, P., Li, D., Catullo, R., Frommer, M., Sved, J. A., Cameron, E. C., Edwards, O. R., Taylor, P. W., & Oakeshott, J. G. (2021). Tracing the origins of multiple Queensland fruit fly incursions into South Australia, Tasmania and New Zealand. *Biological Invasions*, 23, 1117–1130. https://doi.org/10.1007/s10530-020-02422-2

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Raghu, S., Clarke, A. R., Drew, R. A. I., & Hulsman, K. (2000). Impact of habitat modification on the distribution and abundance of fruit flies (Diptera: Tephritidae) in Southeast Queensland. *Population Ecology*, 42(2), 153–160. https://doi.org/10.1007/PL00011994

Raphael, K. A., Shearman, D. C. A., Gilchrist, A. S., Sved, J. A., Morrow, J. L., Sherwin, W. B., Riegler, M., & Frommer, M. (2014). Australian endemic pest tephritids: genetic, molecular and microbial tools for improved Sterile Insect Technique. *BMC Genetics*, 15(Suppl 2), S9. https://doi.org/10.1186/1471-2156-15-S2-S9

Reynolds, J., Weir, B. S., & C., C. C. (1983). Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics*, 105(3), 767–779.

Sadler, R. J., Florec, V., White, B., & Dominiak, B. C. (2011). Calibrating a jump-diffusion model of an endemic invasive: metamodels, statistics and Qfly. *19th International Congress on Modelling and Simulation MODSIM2011*, 2549–2555.

Schmidt, T. L., Swan, T., Chung, J., Karl, S., Demok, S., Yang, Q., Field, M. A., Odwell Muzari, M., Ehlers, G., Brugh, M., Bellwood, R., Horne, P., Burkot, T. R., Ritchie, S., & Hoffmann, A. A. (2021). Spatial population genomics of a recent mosquito invasion. *Molecular Ecology*, 30, 1174–1189. https://doi.org/10.1111/mec.15792

Shigesada, N., & Kawasaki, K. (2002). Invasion and the range expansion of species: effects of long-distance dispersal. *Dispersal Ecology*, 350–373.

Shinozuka, H., Cogan, N. O. I., Shinozuka, M., Marshall, A., Kay, P., Lin, Y.-H., Spangenberg, G. C., & Forster, J. W. (2015). A simple method for semi-random DNA amplicon fragmentation using the methylation-dependent restriction enzyme MspJI. *BMC Biotechnology*, 15, 25. https://doi.org/10.1186/s12896-015-0139-7

Shriner, D. (2011). Investigating population stratification and admixture using eigenanalysis of dense genotypes. *Heredity*, 107(5), 413–420. https://doi.org/10.1038/hdy.2011.26

Smith, E. S. C., Chin, D., Allwood, A. J., & Collins, S. G. (1988). A revised host list of fruit flies (Diptera: Tephritidae) from the Northern Territory of Australia. *Queensland Journal of Agricultural and Animal Sciences*, 45(1), 19–28.

Suckling, D. M., Kean, J. M., Stringer, L. D., Cáceres-Barrios, C., Hendrichs, J., Reyes-Flores, J., & Dominiak, B. C. (2016). Eradication of tephritid fruit fly pest populations: Outcomes and prospects. *Pest Management Science*, 72(3), 456–465. https://doi.org/10.1002/ps.3905

Sved, J. A., Yu, H., Dominiak, B., & Gilchrist, A. S. (2003). Inferring modes of colonization for pest species using heterozygosity comparisons and a shared-allele test. *Genetics*, 163(2), 823–831. https://doi.org/10.1093/genetics/163.2.823

Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, 123(3), 585–595.

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, 17(2), 194–208. https://doi.org/10.1111/1755-0998.12593

Waples, R. K., Albrechtsen, A., & Moltke, I. (2019). Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Molecular Ecology*, 28(1), 35–48. https://doi.org/10.1111/mec.14954

Waples, R. S., & Allendorf, F. (2015). Testing for hardy-weinberg proportions: Have we lost the plot? *Journal of Heredity*, 106(1), 1–19. https://doi.org/10.1093/jhered/esu062

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 276(7), 256–276.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2), 97–158. https://doi.org/10.1007/BF02459575

Wright, S. (1943). Isolation by Distance. *Genetics*, 28(2), 114.

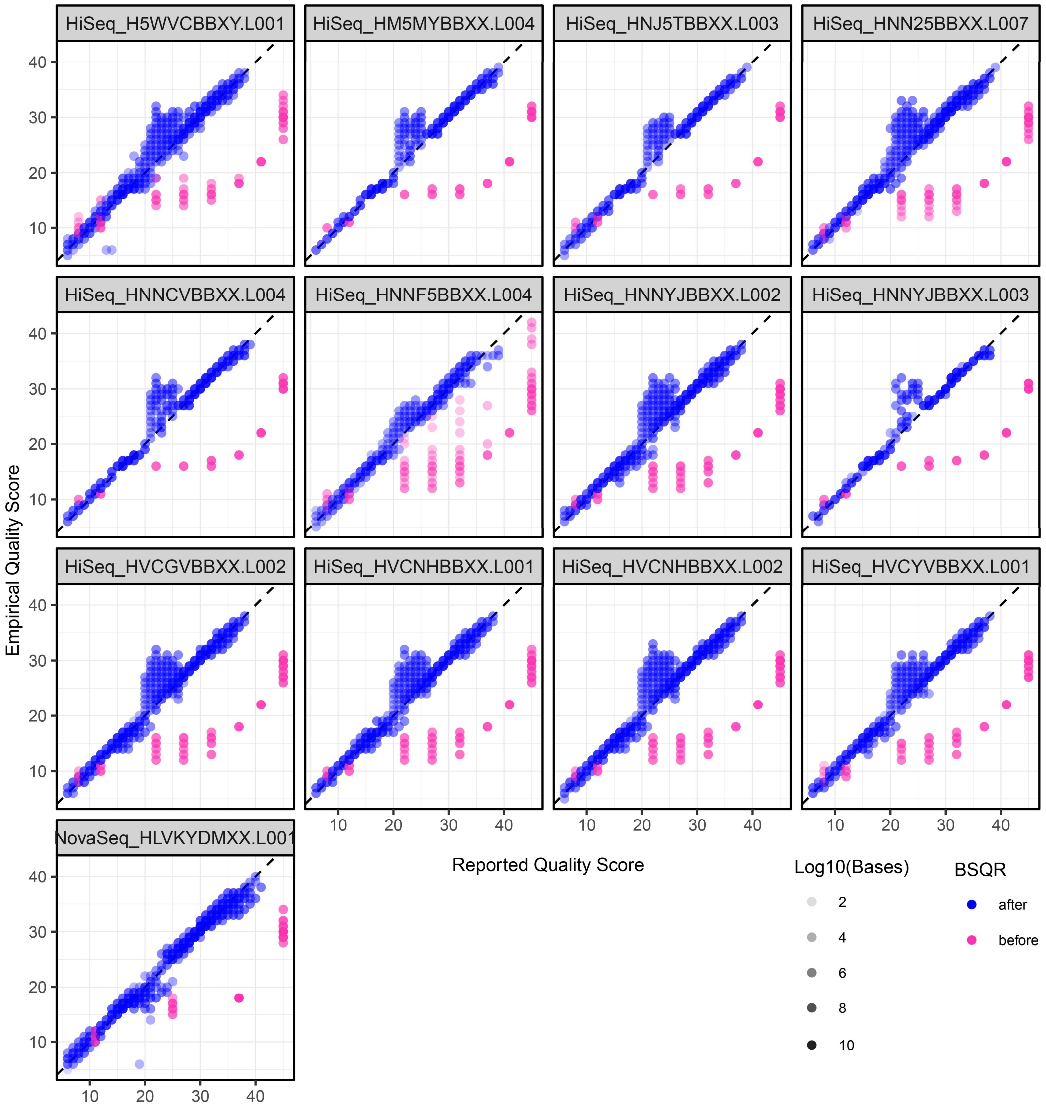
Wright, S. (1946). Isolation by distance under diverse systems of mating. *Genetics*, 31(1), 39–59.

Yonow, T., & Sutherst, R. W. (1998). The geographical distribution of the Queensland fruit fly, Bactrocera (Dacus) tryoni, in relation to climate. *Australian Journal of Agricultural Research*, 49(6), 934–954.

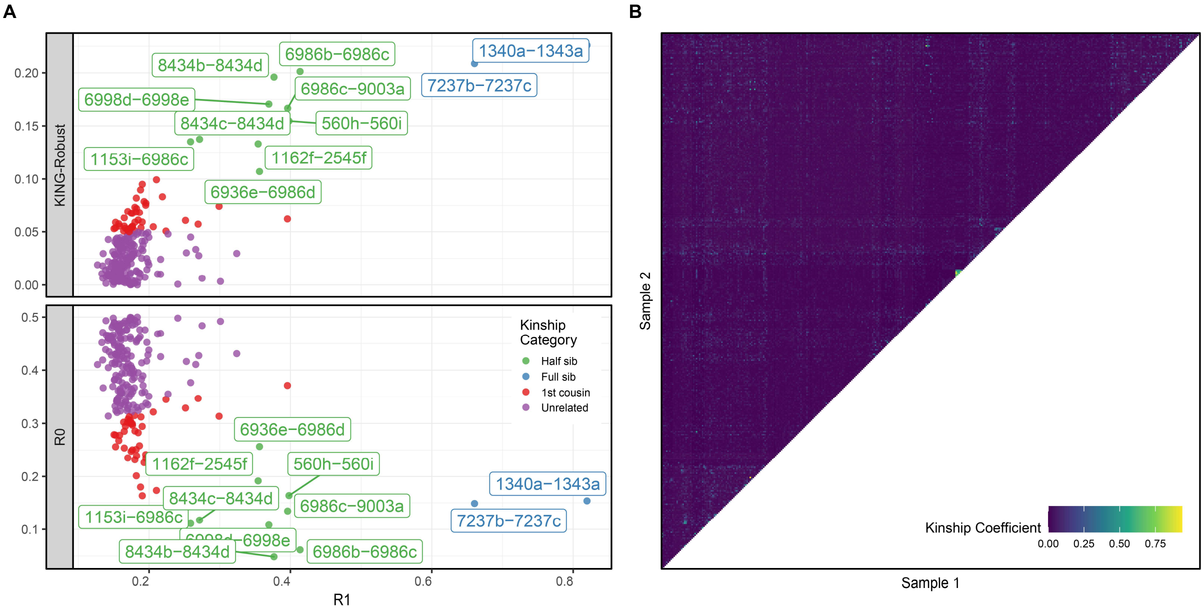
Yu, H., Frommer, M., Robson, M. K., Meats, a W., Shearman, D. C., & Sved, J. a. (2001). Microsatellite analysis of the Queensland fruit fly Bactrocera tryoni (Diptera: Tephritidae) indicates spatial structuring: implications for population control. *Bulletin of Entomological Research*, 91(2), 139–147. https://doi.org/10.1079/BER200075

Zalucki, M. P., Drew, R. A. I., & Hooper, G. H. S. (1984). Ecological studies of Eastern Australian fruit flies (Diptera: Tephritidae) in their endemic habitat - II. The spatial pattern of abundance. *Oecologia*, 64(2), 273–279. https://doi.org/10.1007/BF00376882

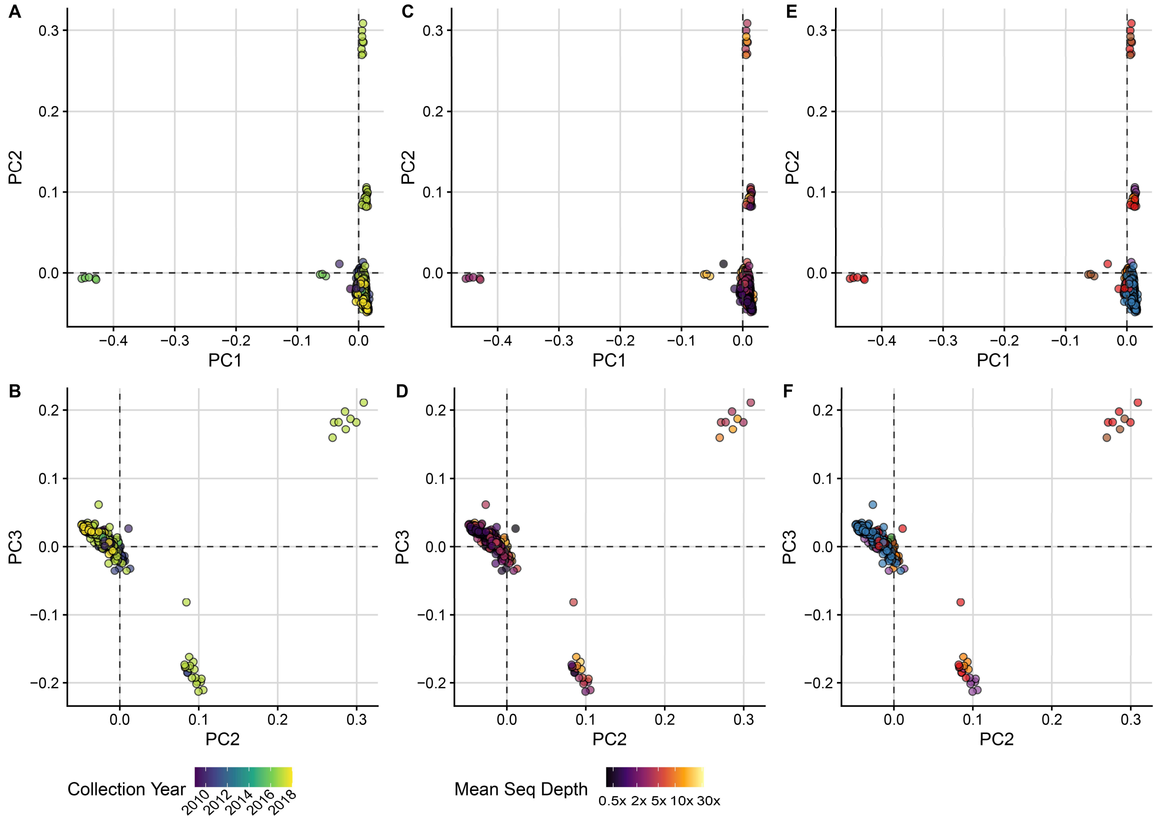
6.5 Supplementary Information:



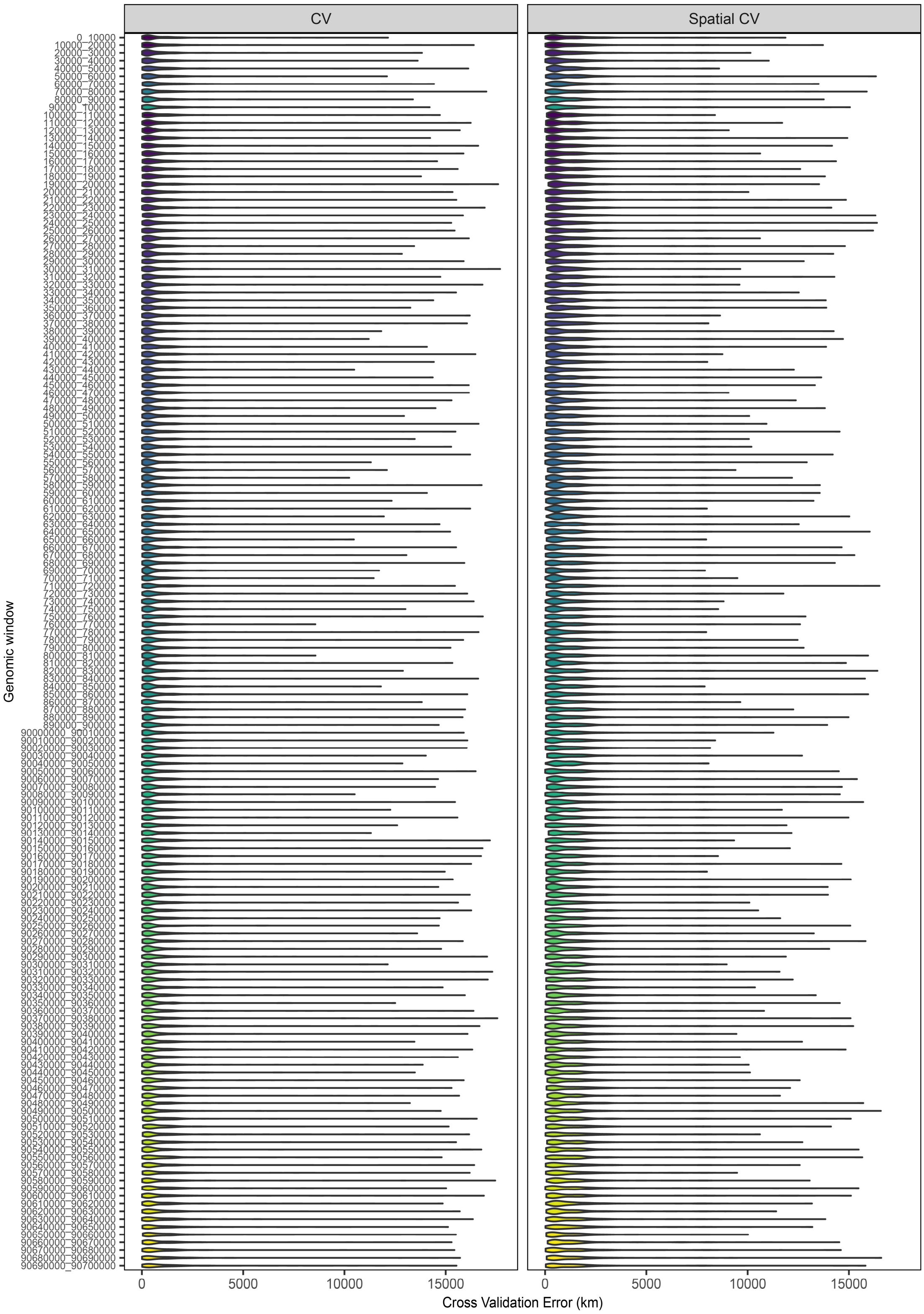
Supplementary Figure 1: Empirical vs reported quality scores for each flow cell lane pre- and post- base quality score recalibration (BQSR)



Supplementary Figure 2: A) Identification of close-kin dyads using the combination of KING-Robust, R0, and R1 statistics. **B)** Genome-wide kinship coefficient for all samples.



Supplementary Figure 3: Principal component analysis of genetic distance between individuals, coloured by **A, B)** collection year, **C, D)** Mean sequencing depth, and **E, F)** Flow-cell each sample was sequenced across. Top panels display principal component axes 1 & 2, while bottom panels display axes 2 & 3.



Supplementary Figure 4: Cross-validation error across each separate 10k SNP window used for Locator training and prediction.

7
General Discussion

The research presented within this thesis demonstrates how genomic biosurveillance can be applied to efficiently detect, quantify abundance, and trace the origin of insect pest outbreaks. As an outcome of this research, two practical diagnostic tools were developed, each designed to be suitable for use across the broad scope of taxa targeted by biosecurity surveillance programmes. The first tool is a high-throughput metabarcoding assay that enables detection of both target and unanticipated non-target invasive insects within large unsorted trap samples. The second tool is a low-coverage whole genome sequencing (lcWGS) assay that provides dense genome-wide SNP markers from single specimens, which can be used to trace the geographic origin of new outbreaks and explore the genetic structure of established populations. These tools were developed and evaluated across four experimental chapters, combining meta-analysis of public datasets, laboratory and field evaluation, statistical modelling, and population genetic analyses. Within each chapter, a separate exotic or established insect pest was used as a model system on which each approach was validated, serving to demonstrate a flexible genomic biosurveillance pipeline that could be readily expanded to any emerging threat.

Universal metabarcoding diagnostics

A primary aim of this thesis was to develop a molecular technique that could overcome the ‘needle in a haystack’ problem of detecting low abundance pests within the diverse mixed specimens collected by surveillance traps. Chapter 2 explored the use of DNA metabarcoding for this purpose, highlighting its eminent suitability, but also the technical and regulatory challenges that must be overcome before adoption within routine diagnostic operations. The first two experimental chapters (Chapters 3 and 4) then set out to address several of these issues: Chapter 3 established the taxonomic breadth across which short subregions of COI can achieve species-level resolution, demonstrating suitable diagnostic performance for the majority of insect taxa registered on global invasive species lists. Chapter 4 then applied this mini-barcode in a non-destructive metabarcoding assay, showing it could successfully detect invasive

Drosophila suzukii within unsorted trap catches, whilst retaining intact voucher specimens for confirmation of any detected exotic taxa using morphological methods. This chapter highlighted the need for increased replication to enable reliable detection with non-destructive assays and compared methods for resolving real detections from index-switching, a pervasive source of false positives in metabarcoding assays. This study demonstrates the practical feasibility of metabarcoding-based diagnostics and provides laboratory and bioinformatic protocols to facilitate uptake within early detection surveillance programmes for *D. suzukii* currently being launched in Australia. When considering the massive capacity of contemporary HTS platforms (Piper et al., 2019), adopting this metabarcoding assay could enable a substantial increase in the geographic scale and intensity of planned trapping, and thus likelihood of detecting a new incursion (Epanchin-Niell et al., 2012).

Metabarcoding is further unique among diagnostic assays in that, in addition to the target species, it also provides the identities of diverse taxa caught as bycatch within surveillance traps: in Chapter 4, 34 non-drosophilid species were detected within field deployed traps, and in Chapter 5 an additional 26 species were recorded by metabarcoding compared to morphological sorting. While in both chapters all identified bycatch had been previously recorded in Australia, in other studies, metabarcoding has revealed the presence of unanticipated exotic species that were not actively being searched for (Batovska et al., 2020; Brown et al., 2016; Hardulak et al., 2020; Young et al., 2021). Considering the high initial cost of implementing a pest survey, further examining trap bycatch for other potential new introductions presents a cost effective decision (Looney et al., 2016), yet this only rarely occurs due to the considerable taxonomic effort and expertise required by traditional identification methods (Spears & Ramirez, 2015). This ability of metabarcoding to screen all trapped specimens is an exciting step toward comprehensive surveillance programmes that aim to detect and evaluate all newly introduced species, not just those regulated by national quarantine agencies. For instance, those which may be minor or non-pests in other countries and thus overlooked by risk assessment but could emerge as damaging pests within unique Australian ecosystems (Lott & Rose, 2016). Fully realising this goal will, however, require more than just a universal diagnostic assay: before an insect can be identified it must first be caught in a trap, and both trap designs and surveillance grid layouts, which have traditionally been targeted to single species, may need to be reconsidered to ensure a broader taxonomic diversity is captured for metabarcoding screening. This could involve supplementing highly-specific pheromone lures with more generic semiochemical attractants (Gandhi et al., 2009), or use of completely passive malaise traps commonly used in biodiversity surveys (Hardulak et al., 2020). This does present a trade-off, however, as more broadly tuned lures generally lack the sensitivity of pheromone lures over long distances (Byers et al., 1989; Larsson, 2016). Comprehensive species detection will also require baseline knowledge of endemic biodiversity against which potential new introductions can be assessed (Bishop & Hutchings, 2011), information which is both geographically limited and taxonomically biased in Australia and abroad (Cranston, 2010; Rocha-Oretga et al., 2021). As conducting the required baseline biodiversity surveys is outside the traditional scope of biosecurity agencies, this would benefit from increased engagement with biodiversity researchers, natural resource managers, and international biosurveillance efforts such as the BIOSCAN and Genomic Observatories initiatives (Arribas et al., 2021; Hobern, 2020). If systematically implemented across global ports of entry, comprehensive and ongoing metabarcoding surveys could not only improve detection of newly introduced species, but also reveal the pool of potentially invasive species within source locations (Chown et al., 2015), and provide fundamental insights into the dynamics and distribution of global insect biodiversity (Arribas et al., 2021).

Towards quantitative metabarcoding

For biosurveillance applications such as population monitoring to support pest eradication or suppression efforts, obtaining presence or absence information alone may be insufficient for effective decision making. Chapter 3 predicted substantial differences in taxonomic bias across published COI primers, but also showed this could be largely alleviated by incorporating 4-5 degenerate nucleotide bases during primer design. Despite this, species-specific variation in detection efficiency was still seen in the metabarcoding assay of Chapter 4, reinforcing the diverse molecular and morphological contributors (Liu et al., 2020; Piper et al., 2019). Chapter 5 subsequently evaluated the use of statistical models to account for taxonomic bias during analysis, demonstrating that bias-corrected metabarcoding assays can provide insect abundance measurements comparable to those obtained through conventional morphological sorting. A major limitation remains, however, in that these bias correction models must be trained on pre-identified communities and are therefore only applicable to taxa that be acquired in advance. While of little importance for the small cohort of pheromone-trapped *Carpophilus* species analysed in the case study, this restriction becomes constraining when considering the diverse assemblages that can be collected through less targeted sampling methods such as wind-based trapping (Hardulak et al., 2020; Watts et al., 2019). While the use of phylogenetic imputation methods to generalise correction factors to closely related taxa provides a potential workaround (Goberna & Verdú, 2016; McLaren et al., 2019), ultimately, further mechanistic investigation into the processes contributing to taxonomic bias will be required. The issues of primer-template mismatch and specimen biomass have seen considerable attention in the literature (Clarke et al., 2014; Deagle et al., 2014; Elbrecht et al., 2017; Elbrecht & Leese, 2017; Piñol et al., 2019), yet this has largely been at the expense of other important contributors such as exoskeleton hardness, mitochondrial copy number variation, and differential degradation within field deployed insect traps (Krehenwinkel et al., 2017, 2018; Marquina et al., 2019). Rather than considering taxonomic bias as one single process, future studies should partition the total protocol bias into its constituent laboratory, field, and bioinformatic components, then separately optimise protocols for each (Brooks et al., 2015).

Genomic sequencing & Pathway tracing

Following detection of a newly invasive population, determining the pathway and timeframe for its introduction can inform eradication efforts and allow targeting of regulatory and extension activities to reduce the likelihood of future pest introduction along that pathway (Barr et al., 2014; Liebhold et al., 2016). While the metabarcoding assay developed in chapters 3-5 could detect and measure the abundance of invasive insect pests, the limited nucleotide variation contained within the COI mini-barcode provides insufficient resolution for this kind of pathway analysis. Chapter 6 saw the development of a complementary lcWGS assay that enables high-resolution investigation of genetic diversity within invasive populations. Applied to the recent range expansion of the Queensland fruit fly, *Bactrocera tryoni*, the resulting genome-wide SNP data revealed endemic populations to be genetically homogenous over large distances, while both incipient populations in the invasive range and disjoint island populations showed genetic bottlenecks and limited gene flow. Using this genomic dataset as a reference panel, specimens from recent outbreaks were assigned to a probable geographic origin, however limited concordance between genetic and spatial structure resulted in confidence intervals that covered a large geographic area. Rather than being an inadequacy of the lcWGS assay itself, this uncertainty instead reflects the complex patterns of genetic diversity that can arise for species such as *B. tryoni*, where human mediated long-distance dispersal events are common (Nichols & Hewitt, 1994; Shigesada & Kawasaki, 2002). Despite the limited success of outbreak tracing in this study, the identification of genetically bottlenecked populations that show limited connectivity, and therefore low recolonisation risk, could help define eradicable units suitable for future control efforts (Liebhold et al., 2016; Robertson & Gemmell, 2004). Further comparing the strength of these genetic bottlenecks across incipient populations and relating this to population reduction control measures and future population persistence may enable outbreak thresholds for *B. tryoni* to be refined through an alternative genomic lens (Dominiak et al., 2011; Suckling et al., 2016).

Similar to the metabarcoding assay, the lcWGS approach developed here requires no prior ascertainment of target loci (Lou et al., 2020; Therkildsen & Palumbi, 2017) and could therefore be applied to other invasive insects where outbreak tracing may prove more successful. A challenge still remains, however, in the lack of published statistical methods that account for genotype uncertainty inherent to low-coverage datasets, meaning the geographic assignments made within Chapter 6 may have been biased by the requirement to hard-call genotypes before assignment (Nielsen et al., 2011, 2012). While a wider suite of methods for analysing low-coverage sequencing data is desirable, as HTS costs continue to decrease the sequencing depth applied to samples could be raised without any required change in protocol (Malmberg et al., 2018). This would eventually allow the lcWGS assay to transition into conventional whole genome sequencing, opening up a broader suite of statistical techniques for examining invasion processes (North et al., 2021). For instance, approximate Bayesian computation can leverage whole genome data to model complex demographic histories (C. C. R. Smith & Flaxman, 2020; van Boheemen et al., 2017), making it possible to calculate the relative probabilities of competing introduction scenarios such as those seen for disjoint Alice Springs and Melanesian populations in Chapter 6. Furthermore, while Chapter 6 explored a continental-scale range expansion, genomic datasets are also informative for studying dispersal processes at much finer spatial scales (Bradburd & Ralph, 2019). Genome-wide SNP data allows robust identification of familial relationships (Waples et al., 2019), and recent statistical advances can leverage the spatial distribution of close-kin pairs to infer individual dispersal distances and estimate the number of breeding individuals within an area (Filipović et al., 2020; Jasper et al., 2019; Ruzzante et al., 2019). This fine-scale information has clear application to outbreak eradication, and similar kinship-based methods may provide a more tractable approach for understanding population structure for species such as B. *tryoni*, where high levels of shared genetic variation from regional co-ancestry makes application of conventional methods challenging (Schmidt et al., 2021).

Technological access & diagnostic turnaround

Early detection surveillance for new introductions and population monitoring of established pests are both time-critical activities, as remedial action must be taken before breeding populations can establish or widespread crop damage occurs (Pluess et al., 2012; Reaser et al., 2020). Considering this, diagnostic turnaround time presents a major remaining limitation for application of genomic approaches within insect diagnostics, with complex molecular and bioinformatic protocols extending for multiple days and sequencing itself taking between 40-84 hours (Rossen et al., 2018). The logistical challenge of regularly drawing together sufficient samples to fill the massive throughput of contemporary HTS platforms may further constrain diagnostic turnaround, as running sequencers below capacity substantially increases costs (Piper et al., 2019). For practical implementation of genomic biosurveillance, a central diagnostic hub model may prove the most effective, where the samples collected through various surveillance programmes, each targeting different taxonomic groups, are all identified in the same location using universal genomic assays. Achieving this will, however, require increased coordination between the various national, state, and industry organisations which conduct insect surveillance (Lott & Rose, 2016). Alternatively, recent nanopore HTS platforms offer more flexible input requirements, substantially lower purchase price, and real time data production, which may allow for decentralised adoption of sequencers within separate state or regional laboratories (Jain et al., 2016). To date, uptake of nanopore sequencing for species identification has been limited by considerably higher per-base error rates (Benítez-Páez et al., 2016; Krehenwinkel et al., 2019), but recent chemistry advancements and innovative molecular protocols have now reduced this to a level acceptable for most diagnostic applications (Baloğlu et al., 2021; Karst et al., 2021). The extremely low ($1000 USD) purchase price of the Oxford Nanopore Technologies MinION is particularly noteworthy when considering countries most affected by emerging insect pests might be those less likely to have biosecurity agencies sufficiently funded to invest in other sequencing platforms (Bebber et al., 2014; Early et al., 2016). This platform has been successfully applied to both metabarcoding (Baloğlu et al., 2021) and lcWGS (Malmberg et al., 2019), with its small physical size and real-time data production enabling use for on-site diagnostics under challenging field conditions (Boykin et al., 2019; Pomerantz et al., 2018). Ultimately, the goal of genomic biosurveillance is not to replace the role of diagnosticians, but to augment diagnostic decision making with more scalable tools and higher resolution datasets. Nanopore sequencing may therefore offer a promising route towards placing genomic biosurveillance tools directly into the hands of diagnosticians in the form of flexible and portable assays that are as readily accessible as a microscope.

Integrated biosurveillance pipeline

To fully realise the potential of genomic biosurveillance for insect pest management, the high-throughput screening provided by metabarcoding (many specimens, single loci) and the fine-scale resolution of whole genome sequencing (single specimens, many loci) need to be integrated. Chapters 3 and 4 demonstrated the use of non-destructive DNA extraction methods to retain specimens following metabarcoding analysis, which can be used as specimen vouchers for morphological confirmation or have their DNA re-extracted and used for conventional DNA barcoding (Batovska et al., 2020). Following on from this, an integrated genomic biosurveillance pipeline may involve any species of concern detected through non-destructive metabarcoding being confirmed via morphological analysis, and subsequently sequenced using an lcWGS assay to determine its likely introduction pathway. Even closer integration may be possible through alternative 'metagenomic' sequencing approaches discussed in Chapter 2. While metagenomics remains significantly more expensive than metabarcoding, recent advances in meta-haplotyping algorithms may soon allow population genomic analyses such as those conducted in Chapter 6 to be applied directly to mixed samples (Nicholls et al., 2020). Developing this approach to the point of practical implementation will require a large-scale effort to increase the availability of whole genomes within reference databases, a process that will face similar curation challenges as current DNA barcode databases (Piper et al., 2019). Nevertheless, as costs of sequencing continue to fall it is conceivable that in the not too distant future a single metagenomic assay may be able to identify, estimate the abundance, and trace the origin of all species within a mixed trap sample, providing a powerful tool for detection and control of invasive insect pest populations.

An integrated genomic biosurveillance pipeline promises more than just improving detection and understanding of dispersal, as high-density genomic SNP data can shed further light on the evolutionary adaptations facilitating or resulting from colonisation of a new environment (Prentis et al., 2008). Identification of facilitatory 'invasion genes', and an increased understanding of the genomic architecture underlying climatic adaptation would enable refined forecasts of the non-native range of potentially invasive species (Kearney et al., 2009). Further comparing these patterns across a broad range of invasive taxa, facilitated by the species-independent nature of these tools, would have important consequences for both practical biosecurity and fundamental biology. For instance, if similar genetic pathways evolve across species during invasion, it may help predict when colonisation is likely to be successful (Chown et al., 2015), as well as provide new insight into longstanding paradoxes such as how invasive species can successfully adapt and establish despite extreme population bottlenecks (Estoup et al., 2016). Importantly, genomic approaches are not restricted to the invasive species themselves, and can additionally be used to measure changes in community composition following invasion (Chown et al., 2015), determine whether invasive species occur in the diets of potential natural enemies (Cohen et al., 2020; Sow et al., 2019), and evaluate the evolutionary response of native species to novel community members (Strauss et al., 2006). Furthermore, fine scale mapping of genotype-phenotype associations can identify genetic regions conferring insect resistance within affected plants (H. M. Smith et al., 2018), which can then be used as targets for genomic-selection in breeding programmes (Poland & Rutkoski, 2016). Adoption of genomic approaches will therefore prove important not only for invasive insect biosurveillance, but also ensuring continued resilience of agricultural and natural ecosystems to the increasing burden of biological invasion.

7.1 Concluding remarks:

This thesis provides a practical demonstration of how a genomic biosurveillance pipeline can be applied to invasive and established insect pests, focussing on the development of flexible and universal diagnostic tools that provide high-resolution data to inform biosecurity and pest management decision making. Further work will be required to formally validate these tools and ensure reported outputs are clearly defined and interpretable to end users, while widespread adoption will require significant international investment into the infrastructure, human capacity, and taxonomic frameworks underlying insect identification. Genomic tools are not a 'silver bullet', however, and end users will need to be aware of their limitations, in particular the dependence on accurate and well sampled reference databases. This becomes especially important when considering the legal dimensions of biosecurity, as questions of whether a DNA-based detection alone is sufficient to support prosecution of a suspected perpetrator, or restricting trade with a certain country, will no doubt arise in the near future (Bilodeau et al., 2019). While non-destructive assays show promise for circumventing many regulatory challenges, development of new approaches for communicating detection uncertainty, as well as harmonisation of laboratory and bioinformatic approaches across diagnostic laboratories would be desirable. A promising first step would be the creation of a set of national and international guidelines for selecting, developing, validating, and ongoing quality assurance of HTS diagnostics in order to align emerging genomic approaches with the global plant pest regulatory framework. Ultimately, genomic tools form only a component of a larger biosecurity toolkit that integrates rapid, high-resolution diagnostics along with improved risk forecasting, effective trap designs, robust taxonomy, and an overarching decision support system. In an increasingly globalised world, the continued effectiveness of biosecurity surveillance will depend upon close collaboration between academic scientists, diagnosticians, and the many stakeholders who rely on effective surveillance data to manage the spread of invasive pests and pathogens.

7.2 References

Arribas, P., Andújar, C., Bidartondo, M. I., Bohmann, K., Coissac, É., Creer, S., DeWaard, J. R., Elbrecht, V., Ficetola, G. F., Goberna, M., Kennedy, S., Krehenwinkel, H., Leese, F., Novotny, V., Ronquist, F., Yu, D. W., Zinger, L., Creedy, T. J., Meramveliotakis, E., … Emerson, B. C. (2021). Connecting high-throughput biodiversity inventories: Opportunities for a site-based genomic framework for global integration and synthesis. *Molecular Ecology*, 30, 1120–1135. https://doi.org/10.1111/mec.15797

Baloğlu, B., Chen, Z., Elbrecht, V., Braukmann, T., MacDonald, S., & Steinke, D. (2021). A workflow for accurate metabarcoding using nanopore MinION sequencing. *Methods in Ecology and Evolution*, 00, 1–11. https://doi.org/10.1111/2041-210x.13561

Barr, N., Ruiz-Arce, R., & Armstrong, K. (2014). Using molecules to identify the source of fruit fly invasions. In T. E. Shelly, N. Epsky, E. B. Jang, J. Reyes-Flores, & R. I. Vargas (Eds.), *Trapping And The Detection, Control, and Regulation of Tephritid Fruit Flies: Lures, Area-Wide Programs, and Trade Implications*. https://doi.org/10.1007/978-94-017-9193-9_10

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

Bebber, D. P., Holmes, T., Smith, D., & Gurr, S. J. (2014). Economic and physical determinants of the global distributions of crop pests and pathogens. *New Phytologist*, 202, 901–910. https://doi.org/10.1111/geb.12214

Benítez-Páez, A., Portune, K. J., & Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *GigaScience*, 5(1), 4. https://doi.org/10.1186/s13742-016-0111-z

Bilodeau, P., Roe, A. D., Bilodeau, G., Blackburn, G. S., Cui, M., Cusson, M., Doucet, D., Griess, V. C., Lafond, V. M. A., Nilausen, C., Paradis, G., Porth, I., Prunier, J., Srivastava, V., Stewart, D., Torson, A. S., Tremblay, E., Uzunovic, A., Yemshanov, D., & Hamelin, R. C. (2019). Biosurveillance of forest insects: part II—adoption of genomic tools by end user communities and barriers to integration. *Journal of Pest Science*, 92(1), 71–82. https://doi.org/10.1007/s10340-018-1001-1

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

Boykin, L. M., Sseruwagi, P., Alicai, T., Ateka, E., Mohammed, I. U., Stanton, J. A. L., Kayuki, C., Mark, D., Fute, T., Erasto, J., Bachwenkizi, H., Muga, B., Mumo, N., Mwangi, J., Abidrabo, P., Okao-Okuja, G., Omuut, G., Akol, J., Apio, H. B., … Ndunguru, J. (2019). Tree Lab: Portable genomics for early detection of plant viruses and pests in Sub-Saharan Africa. *Genes*, 10, 632. https://doi.org/10.3390/genes10090632

Bradburd, G. S., & Ralph, P. L. (2019). Spatial Population Genetics: It's About Time. *Annual Review of Ecology, Evolution, and Systematics*, 50(1), 427–449. https://doi.org/10.1146/annurev-ecolsys-110316-022659

Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., Reris, R. A., Sheth, N. U., Huang, B., Girerd, P., Strauss, J. F., Jefferson, K. K., & Buck, G. A. (2015). The truth about metagenomics: Quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology*, 15, 66. https://doi.org/10.1186/s12866-015-0351-6

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. https://doi.org/10.1111/ddi.12465

Byers, J. A., Anderbrant, O., & Löqvist, J. (1989). Effective attraction radius - A method for comparing species attractants and determining densities of flying insects. *Journal of Chemical Ecology*, 15(2), 749–765. https://doi.org/10.1007/BF01014716

Chown, S. L., Hodgins, K. A., Griffin, P. C., Oakeshott, J. G., Byrne, M., & Hoffmann, A. A. (2015). Biological invasions, climate change and genomics. *Evolutionary Applications*, 8(1), 23–46. https://doi.org/10.1111/eva.12234

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. https://doi.org/10.1111/1755-0998.12265

Cohen, Y., Bar-David, S., Nielsen, M., Bohmann, K., & Korine, C. (2020). An appetite for pests: Synanthropic insectivorous bats exploit cotton pest irruptions and consume various deleterious arthropods. *Molecular Ecology*, 29(6), 1185–1198. https://doi.org/10.1111/mec.15393

Cranston, P. S. (2010). Insect biodiversity and conservation in Australasia. *Annual Review of Entomology*, 55, 55–75. https://doi.org/10.1146/annurev-ento-112408-085348

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biology Letters*, 10(9), 20140562. https://doi.org/10.1098/rsbl.2014.0562

Dominiak, B. C., Daniels, D., & Mapson, R. (2011). Review of the outbreak threshold for Queensland fruit fly (Bactrocera Tryoni Froggatt). *Plant Protection Quarterly*, 26(4), 141–147.

Early, R., Bradley, B. A., Dukes, J. S., Lawler, J. J., Olden, J. D., Blumenthal, D. M., Gonzalez, P., Grosholz, E. D., Ibañez, I., Miller, L. P., Sorte, C. J. B., & Tatem, A. J. (2016). Global threats from invasive alien species in the twenty-first century and national response capacities. *Nature Communications*, 7, 12485. https://doi.org/10.1038/ncomms12485

Elbrecht, V., & Leese, F. (2017). Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. *Frontiers in Environmental Science*, 5, 11. https://doi.org/10.3389/fenvs.2017.00011

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. https://doi.org/10.1002/ece3.3192

Epanchin-Niell, R. S., Haight, R. G., Berec, L., Kean, J. M., & Liebhold, A. M. (2012). Optimal surveillance and eradication of invasive species in heterogeneous landscapes. *Ecology Letters*, 15(8), 803–812. https://doi.org/10.1111/j.1461-0248.2012.01800.x

Estoup, A., Ravign, V., Hufbauer, R., Vitalis, R., Gautier, M., & Facon, B. (2016). Is There a Genetic Paradox of Biological Invasion ? *Annu Rev Entomol*, 47, 51–72. https://doi.org/10.1146/annurev-ecolsys-121415

Filipović, I., Hapuarachchi, H. C., Tien, W. P., Razak, M. A. B. A., Lee, C., Tan, C. H., Devine, G. J., & Rašić, G. (2020). Using spatial genetics to quantify mosquito dispersal for control programs. *BMC Biology*, 18(1), 104. https://doi.org/10.1186/s12915-020-00841-0

Gandhi, K. J. K., Gilmore, D. W., Haack, R. A., Katovich, S. A., Krauth, S. J., Mattson, W. J., Zasada, J. C., & Seybold, S. J. (2009). Application of semiochemicals to assess the biodiversity of subcortical insects following an ecosystem disturbance in a sub-boreal forest. *Journal of Chemical Ecology*, 35(12), 1384–1410. https://doi.org/10.1007/s10886-009-9724-3

Goberna, M., & Verdú, M. (2016). Predicting microbial traits with phylogenies. *ISME Journal*, 10(4), 959–967. https://doi.org/10.1038/ismej.2015.171

Hardulak, L. A., Morinière, J., Hausmann, A., Hendrich, L., Schmidt, S., Doczkal, D., Müller, J., Hebert, P. D. N., & Haszprunar, G. (2020). DNA metabarcoding for biodiversity monitoring in a national park: screening for invasive and pest species. *Molecular Ecology Resources*, 20(6), 1542–1557. https://doi.org/10.1111/1755-0998.13212

Hobern, D. (2020). BIOSCAN: DNA barcoding to accelerate taxonomy and biogeography for conservation and sustainability. *Genome*, 64(3), 161–164. https://doi.org/10.1139/gen-2020-0009

Hulme, P. E., Bacher, S., Kenis, M., Klotz, S., Kühn, I., Minchin, D., Nentwig, W., Olenin, S., Panov, V., Pergl, J., Pyšek, P., Roques, A., Sol, D., Solarz, W., & Vilà, M. (2008). Grasping at the routes of biological invasions: A framework for integrating pathways into policy. *Journal of Applied Ecology*, 45(2), 403–414. https://doi.org/10.1111/j.1365-2664.2007.01442.x

Jain, M., Olsen, H. E., Paten, B., Akeson, M., Branton, D., Daniel, B., Deamer, D., Andre, M., Hagan, B., Benner, S., Deamer, D., Akeson, M., Branton, D., Kasianowicz, J., Brandin, E., Branton, D., Deamer, D., Cherf, G., Lieberman, K., … Koren, S. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1), 239. https://doi.org/10.1186/s13059-016-1103-0

Jasper, M., Schmidt, T. L., Ahmad, N. W., Sinkins, S. P., & Hoffmann, A. A. (2019). A genomic approach to inferring kinship reveals limited intergenerational dispersal in the yellow fever mosquito. *Molecular Ecology Resources*, 19(5), 1254–1264. https://doi.org/10.1111/1755-0998.13043

Karst, S. M., Ziels, R. M., Kirkegaard, R. H., Sørensen, E. A., McDonald, D., Zhu, Q., Knight, R., & Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. N*ature Methods*, 18(2), 165–169. https://doi.org/10.1038/s41592-020-01041-y

Kearney, M., Porter, W. P., Williams, C., Ritchie, S., & Hoffmann, A. A. (2009). Integrating biophysical models and evolutionary theory to predict climatic impacts on species' ranges: The dengue mosquito Aedes aegypti in Australia. *Functional Ecology*, 23(3), 528–538. https://doi.org/10.1111/j.1365-2435.2008.01538.x

Krehenwinkel, H., Fong, M., Kennedy, S., Huang, E. G., Noriyuki, S., Cayetano, L., & Gillespie, R. (2018). The effect of DNA degradation bias in passive sampling devices on metabarcoding studies of arthropod communities and their associated microbiota. *PLoS ONE*, 13(1), e0189188. https://doi.org/10.1371/journal.pone.0189188

Krehenwinkel, H., Pomerantz, A., Henderson, J. B., Kennedy, S. R., Lim, Y., Swamy, V., Shoobridge, J. D., Patel, N. H., Rosemary, G., Prost, S., Lim, J. Y., Swamy, V., Shoobridge, J. D., Graham, N., Patel, N. H., Gillespie, R. G., & Prost, S. (2019). Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *GigaScience*, 8(6), giz006. https://doi.org/10.1093/gigascience/giz006

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. https://doi.org/10.1038/s41598-017-17333-x

Larsson, M. C. (2016). Pheromones and Other Semiochemicals for Monitoring Rare and Endangered Species. *Journal of Chemical Ecology*, 42(9), 853–868. https://doi.org/10.1007/s10886-016-0753-4

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

Liu, M., Clarke, L. J., Baker, S. C., Jordan, G. J., & Burridge, C. P. (2020). A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology*, 45(3), 373–385. https://doi.org/10.1111/een.12831

Looney, C., Murray, T., Lagasa, E., Hellman, W. E., & Passoa, S. C. (2016). Shadow surveys: How non-target identifications and citizen outreach enhance exotic pest detection. *American Entomologist*, 62(4), 247–254. https://doi.org/10.1093/ae/tmw063

Lott, M. J., & Rose, K. (2016). Emerging threats to biosecurity in Australasia: The need for an integrated management strategy. *Pacific Conservation Biology*, 22(2), 182–188. https://doi.org/10.1071/PC15040

Lou, R. N., Jacobs, A., Wilder, A., & Therkildsen, N. O. (2020). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Authorea Preprints*. https://doi.org/10.22541/au.160689616.68843086/v2

Malmberg, M. M., Barbulescu, D. M., Drayton, M. C., Shinozuka, M., Thakur, P., Ogaji, Y. O., Spangenberg, G. C., Daetwyler, H. D., & Cogan, N. O. I. (2018). Evaluation and Recommendations for Routine Genotyping Using Skim Whole Genome Resequencing in Canola. *Frontiers in Plant Science*, 9, 1809. https://doi.org/10.3389/fpls.2018.01809

Malmberg, M. M., Spangenberg, G. C., Daetwyler, H. D., & Cogan, N. O. I. (2019). Assessment of low-coverage nanopore long read sequencing for SNP genotyping in doubled haploid canola (Brassica napus L.). *Scientific Reports*, 9, 8688. https://doi.org/10.1038/s41598-019-45131-0

Marquina, D., Esparza-Salas, R., Roslin, T., & Ronquist, F. (2019). Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. *Molecular Ecology Resources*, 19(6), 1516–1530. https://doi.org/10.1111/1755-0998.13071

McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing measurements. *ELife*, 8, e46923. https://doi.org/10.7554/eLife.46923

Nicholls, S. M., Aubrey, W., De Grave, K., Schietgat, L., Creevey, C. J., & Clare, A. (2020). On the complexity of haplotyping a microbial community. *Bioinformatics*, btaa977. https://doi.org/10.1093/bioinformatics/btaa977

Nichols, R. A., & Hewitt, G. M. (1994). The genetic consequences of long distance dispersal during colonization. *Heredity*, 72(3), 312–317. https://doi.org/10.1038/hdy.1994.41

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, 7(7), e37558. https://doi.org/10.1371/journal.pone.0037558

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6), 443–451. https://doi.org/10.1038/nrg2986

North, H., Mcgaughran, A., & Jiggins, C. (2021). The population genomics of invasive species. *Authorea Preprints*. https://doi.org/10.22541/au.160968166.65928724/v1

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(2), 407– 419. https://doi.org/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), giz092. https://doi.org/10.1093/gigascience/giz092

Pluess, T., Jarošík, V., Pyšek, P., Cannon, R., Pergl, J., Breukers, A., & Bacher, S. (2012). Which Factors Affect the Success or Failure of Eradication Campaigns against Alien Species? PLoS ONE, 7(10). https://doi.org/10.1371/journal.pone.0048157

Poland, J., & Rutkoski, J. (2016). Advances and Challenges in Genomic Selection for Disease Resistance. *Annual Review of Phytopathology*, 54, 79–98. https://doi.org/10.1146/annurev-phyto-080615-100056

Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L. A., Barrio-Amorós, C. L., Salazar-Valenzuela, D., & Prost, S. (2018). Real-time DNA barcoding in a rainforest using nanopore sequencing: Opportunities for rapid biodiversity assessments and local capacity building. *GigaScience*, 7(4), giy033. https://doi.org/10.1093/gigascience/giy033

Prentis, P. J., Wilson, J. R. U., Dormontt, E. E., Richardson, D. M., & Lowe, A. J. (2008). Adaptive evolution in invasive species. *Trends in Plant Science*, 13(6), 288–294. https://doi.org/10.1016/j.tplants.2008.03.004

Reaser, J. K., Burgiel, S. W., Kirkey, J., Brantley, K. A., Veatch, S. D., & Burgos-Rodríguez, J. (2020). The early detection of and rapid response (EDRR) to invasive species: a conceptual framework and federal capacities assessment. *Biological Invasions*, 22, 1–19. https://doi.org/10.1007/s10530-019-02156-w

Reaser, J. K., Meyerson, L. A., & von Holle, B. (2008). Saving camels from straws: How propagule pressure-based prevention policies can reduce the risk of biological invasion. *Biological Invasions*, 10(7), 1085–1098. https://doi.org/10.1007/s10530-007-9186-x

Robertson, B. C., & Gemmell, N. J. (2004). Defining eradication units to control invasive pests. *Journal of Applied Ecology*, 41(6), 1042–1048. https://doi.org/10.1111/j.0021-8901.2004.00984.x

Rocha-Oretga, M., Rodríguez, P., & Córdoba-Aguilar, A. (2021). Geographical , temporal and taxonomic biases in insect GBIF data on biodiversity and extinction. *Ecological Entomology*. https://doi.org/10.1111/een.13027

Rossen, J. W. A., Friedrich, A. W., & Moran-Gilad, J. (2018). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clinical Microbiology and Infection*, 24(4), 355–360. https://doi.org/10.1016/j.cmi.2017.11.001

Ruzzante, D. E., McCracken, G. R., Førland, B., MacMillan, J., Notte, D., Buhariwalla, C., Mills Flemming, J., & Skaug, H. (2019). Validation of close-kin mark–recapture (CKMR) methods for estimating population abundance. *Methods in Ecology and Evolution*, 10(9), 1445–1453. https://doi.org/10.1111/2041-210X.13243

Schmidt, T. L., Swan, T., Chung, J., Karl, S., Demok, S., Yang, Q., Field, M. A., Odwell Muzari, M., Ehlers, G., Brugh, M., Bellwood, R., Horne, P., Burkot, T. R., Ritchie, S., & Hoffmann, A. A. (2021). Spatial population genomics of a recent mosquito invasion. *Molecular Ecology*, 30, 1174–1189. https://doi.org/10.1111/mec.15792

Shigesada, N., & Kawasaki, K. (2002). Invasion and the range expansion of species: effects of long-distance dispersal. *Dispersal Ecology*, 350–373.

Smith, C. C. R., & Flaxman, S. M. (2020). Leveraging whole genome sequencing data for demographic inference with approximate Bayesian computation. *Molecular Ecology Resources*, 20(1), 125–139. https://doi.org/10.1111/1755-0998.13092

Smith, H. M., Clarke, C. W., Smith, B. P., Carmody, B. M., Thomas, M. R., Clingeleffer, P. R., & Powell, K. S. (2018). Genetic identification of SNP markers linked to a new grape phylloxera resistant locus in Vitis cinerea for marker-assisted selection. *BMC Plant Biology*, 18, 360. https://doi.org/10.1186/s12870-018-1590-0

Sow, A., Brévault, T., Benoit, L., Chapuis, M. P., Galan, M., Coeur d'acier, A., Delvare, G., Sembène, M., & Haran, J. (2019). Deciphering host-parasitoid interactions and parasitism rates of crop pests using DNA metabarcoding. *Scientific Reports*, 9, 3646. https://doi.org/10.1038/s41598-019-40243-z

Spears, L. R., & Ramirez, R. A. (2015). Learning to love Leftovers: Using By-Catch to Expand Our Knowledge in Entomology. *American Entomologist*, 61(3), 168–173. https://doi.org/https://doi.org/10.1093/ae/tmv046

Strauss, S. Y., Lau, J. A., & Carroll, S. P. (2006). Evolutionary responses of natives to introduced species: What do introductions tell us about natural communities? *Ecology Letters*, 9(3), 357–374. https://doi.org/10.1111/j.1461-0248.2005.00874.x

Suckling, D. M., Kean, J. M., Stringer, L. D., Cáceres-Barrios, C., Hendrichs, J., Reyes-Flores, J., & Dominiak, B. C. (2016). Eradication of tephritid fruit fly pest populations: Outcomes and prospects. *Pest Management Science*, 72(3), 456–465. https://doi.org/10.1002/ps.3905

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, 17(2), 194–208. https://doi.org/10.1111/1755-0998.12593

van Boheemen, L. A., Lombaert, E., Nurkowski, K. A., Gauffre, B., Rieseberg, L. H., & Hodgins, K. A. (2017). Multiple introductions, admixture and bridgehead invasion characterize the introduction history of Ambrosia artemisiifolia in Europe and Australia. *Molecular Ecology*, *26*(20), 5421–5434. https://doi.org/10.1111/mec.14293

Waples, R. K., Albrechtsen, A., & Moltke, I. (2019). Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Molecular Ecology*, 28(1), 35–48. https://doi.org/10.1111/mec.14954

Watts, C., Dopheide, A., Holdaway, R., Davis, C., Wood, J., Thornburrow, D., & Dickie, I. A. (2019). DNA metabarcoding as a tool for invertebrate community monitoring: a case study comparison with conventional techniques. *Austral Entomology*, 58(3), 675–686. https://doi.org/10.1111/aen.12384

Young, R. G., Milián-García, Y., Yu, J., Bullas-Appleton, E., & Hanner, R. H. (2021). Biosurveillance for invasive insect pest species using an environmental DNA metabarcoding approach and a high salt trap collection fluid. *Ecology and Evolution*, *11*(4), 1558–1569. https://doi.org/10.1002/ece3.7113