A transcriptome-based analytical workflow for identifying loci for species diagnosis: A case study with Bactrocera fruit flies (Diptera: Tephritidae)

Article *in* Austral Entomology · November 2017

DOI: 10.1111/aen.12321

CITATIONS
14

READS
449

5 authors, including:

Matt N. Krosch
Queensland Police Service
71 PUBLICATIONS **1,501** CITATIONS
SEE PROFILE

Francesca Strutt
Queensland University of Technology
9 PUBLICATIONS **72** CITATIONS
SEE PROFILE

Stephen L Cameron
Purdue University
189 PUBLICATIONS **9,415** CITATIONS
SEE PROFILE

Austral Entomology

Austral Entomology (2017) ••, ••–••

A transcriptome-based analytical workflow for identifying loci for species diagnosis: a case study with *Bactrocera* fruit flies (Diptera: Tephritidae)

Matt N Krosch,[1,2]* Mark K Schutze,[1,2,3] Francesca Strutt,[1,2] Anthony R Clarke[1,2] and Stephen L Cameron[1,2,4]

[1]*School of Earth, Environmental and Biological Sciences, Queensland University of Technology, PO Box 2434, Brisbane, QLD 4001, Australia.*
[2]*Plant Biosecurity Cooperative Research Centre, LPO Box 5012, Bruce, ACT 2617, Australia.*
[3]*Department of Agriculture and Fisheries, Queensland Primary Industries Insect Collection (QDPC), 41 Boggo Road, Dutton Park, QLD 4102, Australia.*
[4]*Department of Entomology, Purdue University, 610 Purdue Mall, West Lafayette, IN 47907, USA.*

Abstract Development of novel molecular methods for accurate and economical identification of species has become critical both for pure biological research and for a wide range of applied areas. The most widely used current molecular diagnostic tool, the mitochondrial cytochrome c oxidase subunit 1 gene (COI), the so-called DNA barcode, has been highly criticised and is known to be ineffective at distinguishing species in many groups. Alternative markers are needed to circumvent these issues and provide diagnosticians with a greater range of tools for making accurate identifications. To address this, we describe here a novel analytical workflow for diagnostic marker development that utilises near-genomic-scale data to search for potential informative loci. The workflow takes advantage of orthologous gene databases, in combination with tests of phylogenetic resolution, and benchmarking of nucleotide variation against COI, to determine putative loci that might outperform COI. We use transcriptomes of 14 tephritid fruit flies and especially the taxonomically complex genus *Bactrocera*, as a case study. Of 1646 orthologues searched, our workflow retained a total of five loci following our conservative filtering strategy. One locus, POP4, had strong potential as a novel diagnostic marker for *Bactrocera* fruit flies. POP4 discriminates most species in the training set of taxa, but like COI fails to separate the sibling species *Bactrocera tryoni* and *Bactrocera neohumeralis*. Further validation of this potential new marker against a broader taxonomic sample is ongoing. We advocate that this simple and efficient workflow is, with minor modification, customisable for diagnostic development in almost any taxonomic group.

Key words COI, molecular diagnostics, OrthoGraph, POP4, RNA-Seq.

INTRODUCTION

The need to reliably and efficiently identify species is fundamental to biological research and to most applied and/or commercial uses of biological data, including areas as disparate as quarantine, food security, invasive species monitoring, bioprospecting, conservation, infectious disease management and forensic science. While the majority of species diagnostic methods developed over the last two centuries rely on morphological features, the limitations of purely morphological diagnoses are now well documented (De Salle *et al*. 2005). As alternatives, DNA-based diagnostic protocols have increased in availability in line with improvements in molecular technologies in general and cost reductions for DNA sequencing in particular. The use of molecular diagnostic protocols for most species, however, faces an immediate hurdle in that species delimitation, for all but a few species was not initially based on differences in molecular-level characters (Schlick-Steiner *et al*. 2010). Instead, most species delimitations are based on discriminating morphological differences. Although such morphological characters may differ greatly in how easily they are observed, whether they are present in all life stages and/or genders and how easily they can be translated into a diagnostic protocol (such as a dichotomous key), they commonly remain the ultimate arbiters of species identification (Clarke & Schutze 2014). Molecular characters, in contrast, rarely form part of the set of discriminating characters for the majority of species descriptions, and so, the development of molecular diagnostic protocols centres on identifying novel characters that correspond to the species limits already defined by other, usually morphological, criteria (but refer to Renner 2016).

The methods by which diagnostic molecular variation is assessed to assign species identifications means that available diagnostic technologies thus fall into two broad groups: categorical tests and tree-based diagnostics. Categorical tests are those that represent underlying molecular variation without direct inference of the molecular sequence and include a wide variety of techniques based on diagnostic variation in the size of DNA fragments as inferred through gel electrophoresis (e.g. RFLP, AFLP, DGGE, RAPD) (Mattock *et al*. 2010). In contrast, tree-based diagnostics use sequence variation to infer phylogenetic relationships and interpret membership of particular clades in the tree

*m.krosch@qut.edu.au

doi: 10.1111/aen.12321

as indicative of species identity (e.g. Collins *et al.* 2012). Categorical tests have advantages of speed and economy, but are disadvantaged by how previously uncharacterised species are dealt with; i.e. if they match a characterised species, the test will return a false positive, and entirely novel molecular profiles cannot be objectively designated as either uncharacterised species or uncharacterised variation within a characterised species. Tree-based diagnoses are generally slower and more expensive but are more flexible with respect to how novel molecular variation can be interpreted. Specifically, placing an 'unknown' barcode specimen in a phylogenetic context allows a more accurate identification to be made. Fundamentally, this aligns with the phylogenetic species concept (Rosen 1979; Mishler & Donoghue 1982; Donoghue 1985; Mishler 1985; Nixon & Wheeler 1990); however, the nature of speciation is such that this approach also complies with most other species concepts. Additionally, because there is now a well-developed theoretical framework for phylogenetic species delimitation (e.g. Sites & Marshall 2004; Pons *et al.* 2006; but also refer to Carstens *et al.* 2013), tree-based diagnostic methods can be based on the same markers used for species delimitation.

Although a large number of categorical molecular species diagnostics have been developed, the majority of research in this field has relied on tree-based diagnoses, of which the DNA barcoding paradigm (Hebert *et al.* 2003) has been far and away the most widely adopted. DNA barcoding, as originally proposed, was intended as a universal molecular identification tool for animals by sequencing a standardised gene, the mitochondrial cytochrome c oxidase subunit 1 (COI) gene, and analysed using genetic distance phylogenetic inference (NJ tree building) (Hebert *et al.* 2003; Collins & Cruickshank 2013). While variants of this standard approach have included using different standard genes for some taxa (e.g. plants, Hollingsworth *et al.* 2011), multi-locus approaches (Dupuis *et al.* 2012) and coalescent tree building (Dowton *et al.* 2014), 'classical' DNA barcoding has been used extensively, with, as of February 2016, over 4.7 million barcodes sequenced representing over 230 000 species (Barcode of Life Database, www.boldsystems.org). Despite its widespread application, DNA barcoding has been highly criticised (e.g. Will & Rubinoff 2004; Meier *et al.* 2006; Rubinoff *et al.* 2006a; Taylor & Harris 2012; Collins & Cruickshank 2013), both from a traditionalist perspective favouring morphological identifications and also from molecular evolutionary perspectives.

One of the key molecular biology-based criticisms of DNA barcoding relates to choice of the standardised marker gene COI. From the original DNA barcode proposal by Hebert *et al.* (2003), and in most subsequent empirical DNA barcoding studies of various faunal and taxonomic groups, no attempt was made to justify the choice of COI over other potential marker loci. The performance of DNA barcode studies has typically been assessed by a verification approach whereby the proportion of species correctly assigned by COI is the measure of diagnostic success. This is similar to most methods for developing categorical diagnostic tests that assess if a given marker can successfully discriminate species, rather than assessing multiple markers to determine which has the greatest discriminating power. Attempts to identify which of a range of markers work best for barcoding particular taxonomic groups have been attempted (e.g. Vences *et al.* 2005; Luo *et al.* 2011; Nelson *et al.* 2012), with the most notable being the search for an optimal barcode locus for plants (Kress *et al.* 2005; Rubinoff *et al.* 2006b; Hollingsworth *et al.* 2011). Alternatively, Coissac *et al.* (2016) and others have argued for an 'extended barcode', the use of the whole organelle genome as the diagnostic loci collected by genome skimming or other low-pass next-generation sequencing approaches; however, such methods remain far too expensive for routine diagnostic applications (>$200/sample). Nonetheless, both classical COI barcoding and studies that have attempted to identify alternative and/or complementary markers still rely almost exclusively on organelle genomes (mitochondria for animals, chloroplasts for plants) as the source of marker loci. This reliance on organelle-derived markers has also been heavily criticised due to the unique aspects of organelle inheritance (e.g. Rubinoff *et al.* 2006a) and the existence of nuclear pseudogenes (Song *et al.* 2008). Attempts to utilise nuclear genes as DNA barcode-style diagnostic markers have yet to be significantly used for any group other than fungi, where ribosomal ITS has been applied (Seifert 2009).

The ideal design for a DNA-based, species diagnostic test would thus address the concerns outlined in the previous texts. First, it would preferably use a tree-based approach as this is more flexible with respect to uncharacterised species and/or populations than categorical tests, and tree-based approaches are in line with modern phylogenetic methods of species delimitation. Second, the marker/markers to be used would be chosen from the largest potential pool of markers that could be practically assessed. Experimentally determining which markers have the highest discriminating power is vastly superior to simply verifying if an arbitrarily chosen marker can discriminate a test set of species. Finally, the markers chosen would have diagnostic utility as individual genes or as a small number of loci to allow broad and economical use of the test for applied diagnostics. A diagnostic protocol based on complete genome sequencing for all unknown samples is unlikely to be routinely used (cf. Coissac *et al.* 2016). An ideal diagnostic would thus resemble a DNA barcode (sensu Hebert *et al.* 2003) but with marker loci chosen from a broad survey of genome-level data for their ability to discriminate between species within the focal taxon. Genomic and near-genomic datasets (e.g. EST libraries) have been used to screen for single-nucleotide polymorphism (SNP) and simple sequence repeat (SSR) markers (e.g. Duran *et al.* 2009; Davey *et al.* 2011), but these marker types are not suitable for tree-based analysis and cannot be effectively down-scaled for use by sequencing one or even a small number of loci. The object of the present study was to determine if RNA-Seq data, as a form of near-genomic data, could be mined for such potentially 'ideal DNA barcodes' in a group with a high need for efficient molecular diagnostics, the tephritid fruit fly genus *Bactrocera* Macquart.

The genus *Bactrocera* (Insecta: Diptera: Tephritidae) is exemplary of the taxonomic groups most in need of molecular diagnostic methods and the challenges in developing them. *Bactrocera* is a large genus, over 600 species, which includes

the most significant pests of horticulture in Asia, Africa and Australasia (White & Elson-Harris 1992). A comparatively small number of species are pests (only 45 economically important species are listed in Drew & Romig 2013, or ~7% of the genus), but they are often very difficult to distinguish morphologically from non-pest species, with a series of morphologically nearly homogenous 'species complexes' formally recognised within the genus (Schutze *et al*. 2017). Additionally, the destructive phase of the insect is the maggot, which develops within ripening fruit and lacks nearly all species-level diagnostic morphological features. This life stage is thus the target of quarantine efforts to prevent the international spread of pest fruit flies in infested fruit (Dohino *et al*. 2017). Identifications of suspect intercepts have historically required rearing maggots to adulthood, greatly slowing trade and biosecurity responses. Several molecular diagnostic approaches have been developed for the genus to speed the identification process; however, they all possess considerable limitations.

Earlier methods, including allozyme electrophoresis and RFLP, were developed for key pest groups such as species in the oriental fruit fly (*Bactrocera dorsalis*) complex (Yong 1995; Armstrong *et al*. 1997; Armstrong & Cameron 1998; Muraji & Nakahara 2002), and these remain part of current diagnostic protocols (Plant Health Australia 2016). However, these approaches were developed for only a small number of species, have an especially high risk of detecting false positives, do not distinguish some very closely related species and thus have extremely limited capacity across the genus (Armstrong & Cameron 1998; Plant Health Australia 2016). COI barcoding has been used extensively to improve *Bactrocera* diagnostics, due to its high-throughput capacity and ability to make comparisons against broader barcoding studies and available databases (e.g. BOLD). In arguably the largest single tephritid barcoding initiative to date, Armstrong and Ball (2005) analysed COI barcodes for 60 tephritid species, of which nearly 40 were from the genus *Bactrocera*. Akin to other techniques, the COI barcode does not resolve among closely related members of some species complexes (e.g. *Bactrocera tryoni* cannot be discriminated from *Bactrocera neohumeralis*); levels of intraspecific vs. interspecific sequence divergence overlap for some taxa; NJ analyses nest some unambiguous species within others, and preferentially amplified pseudogenes confound interpretation of sequencing results (Armstrong & Ball 2005; Blacket *et al*. 2012; Morrow *et al*. 2015). Barcoding in fruit flies also suffers from the requirement for an extensive, accurately identified, reference database for species identification (Barr *et al*. 2012; Frey *et al*. 2013). This is particularly problematic when datasets are drawn from publicly available databases, such as NCBI Genbank or BOLD (e.g. Jiang *et al*. 2016), for which taxon assignment may be erroneous and is especially challenging for groups where species boundaries remain unresolved. The highly diverse *B. dorsalis* species complex perhaps best exemplifies this (Boykin *et al*. 2014; Schutze *et al*. 2015), but it is not unique (Schutze *et al*. 2017).

It has become evident that 'standard' molecular diagnostic approaches have reached their limits of capacity for *Bactrocera* (Blacket *et al*. 2012; Jiang *et al*. 2014), and while some tools are useful under specific circumstances, there is a need to develop novel markers that are genus-specific and based on robust phylogenetic associations. Such new diagnostic tools would be more likely to discriminate among previously 'un-diagnosable' taxa should they be supported as real biological entities. In the present study, therefore, we test the utility of a novel approach to genome mining for identifying candidate markers in an economically important group with a demonstrated need for new and improved molecular diagnostic tools. We use a variety of criteria to filter new loci and benchmark against COI, including testing specific hypotheses about the diagnostic capability of potential novel loci by sequencing a selection of species from the *B. dorsalis* and *B. tryoni* complexes. Specifically, COI does not separate *B. tryoni* and *B. neohumeralis*, whereas it does discriminate *B. dorsalis*, *Bactrocera carambolae* and *Bactrocera kandiensis*. This specific comparison of the diagnostic informativeness of novel loci allows identification of loci that perform as well or better than COI and which are of greatest priority for further testing.

MATERIALS AND METHODS

Locus discovery workflow

The conceptual workflow for locus discovery used is outlined in Figure 1. Conceptually, the workflow consists of finding sets of 1:1 orthologous nuclear protein coding genes shared across the taxon of interest (in this case, *Bactrocera* fruit flies). Each 1:1 orthologue is then tested for the resolution of species-level divergences within this group by inferring its corresponding gene tree. Individual loci whose gene trees recover a set of relationships 'expected' on the basis of previous phylogenetic studies are retained: Those that lack them are excluded. This step selects for loci whose gene tree resolution matches that of the species tree for the taxon of interest, ensuring that, for any locus selected as a novel barcode, new unknown species can be placed accurately in a phylogenetic context, thereby aiding species identification. Retained loci were then assessed for exon–intron boundaries and useful size by mapping against available genome data, to exclude those which could not be reliably amplified in a single PCR. Finally, nucleotide variability within each locus was compared with that of COI sequences from the same set of taxa; candidate loci that had higher variability than the standard barcode gene were retained for experimental verification. Exact methodologies used for each step in this workflow are outlined in the succeeding texts.

Taxon coverage, data collection and de novo transcriptome assembly

Taxon choice aimed to maximise coverage of *Bactrocera* and included members of the closely related genus *Zeugodacus* (*Zeugodacus cucumis* (French) and *Zeugodacus cucurbitae* (Coquillett)) and two other tephritid taxa (*Ceratitis capitata* (Weidemann) and *Rhagoletis pomonella* (Walsh)). We downloaded transcriptome gene sets for all *Bactrocera* species that were available on GenBank at 13 July 2015, including *B. dorsalis*

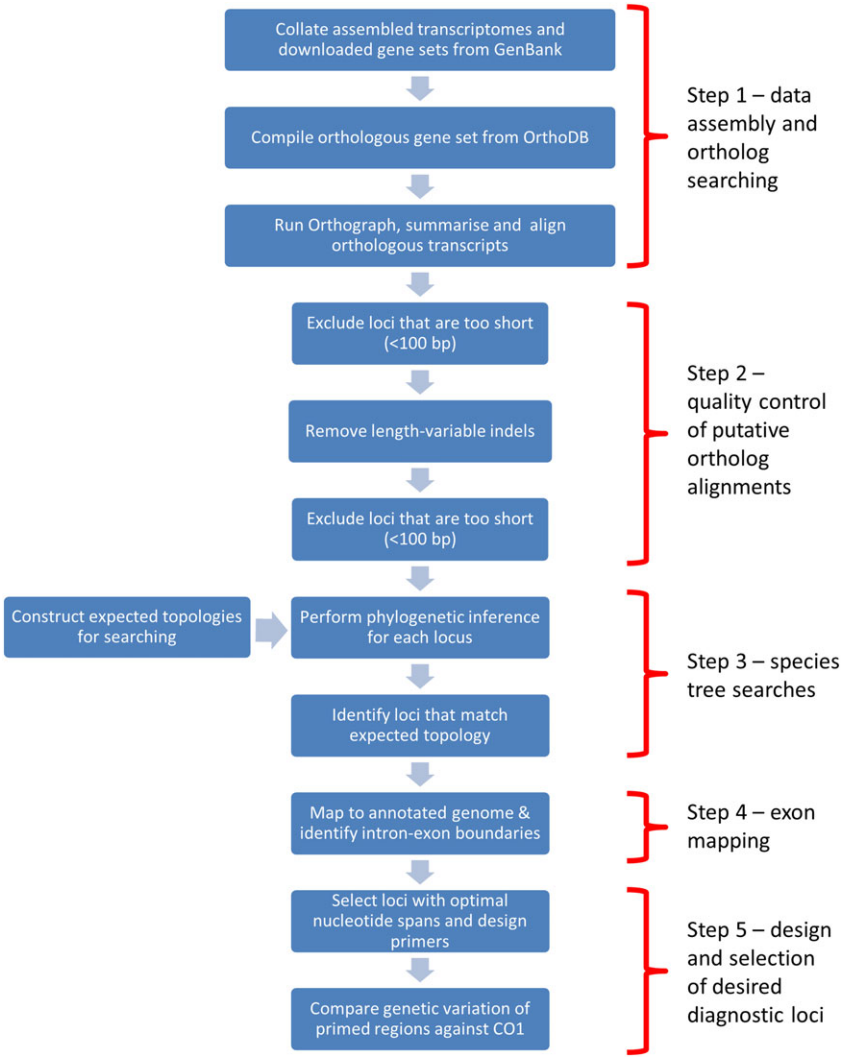

Fig. 1. Schematic flow chart that illustrates the major steps in the analytical workflow for diagnostic locus discovery and development described in this paper. [Colour figure can be viewed at wileyonlinelibrary.com]

(Hendel), *B. minax* (Enderlein) and *B. oleae* Gmelin, along with *Z. cucurbitae*, *C. capitata* and *R. pomonella* (Table 1). We used an unpublished transcriptome assembly for *B. tryoni* (Froggatt) from the study of Kumaran *et al*. (2014). Short read archives (SRAs) for *B. latifrons* (Hendel) were also downloaded and assembled in-house (refer to the succeeding texts).

To complement the existing GenBank data, we *de novo* sequenced adult transcriptomes for six additional dacine species (*B. bryoniae* (Tryon), *B. frauenfeldi* (Schiner), *B. jarvisi* (Tryon), *B. kraussi* (Hardy), *B. musae* (Tryon) and *Z. cucumis*). Again, this was to maximise taxonomic diversity and breadth across the tree in the initial set, but was limited by availability of fresh material and project scope. Specimens were collected from laboratory cultures maintained by Queensland Department of Agriculture and Fisheries, Cairns, Australia. Cultures were kept in separate rooms for each species, and adult males and females were collected between 3 and 5 days after emergence, killed by freezing and transferred immediately to RNA*later* for transport to the Molecular Genetics Research Facility (MGRF), QUT, Brisbane, Australia. One individual of each sex per species was pooled in a single replicated RNA extraction, conducted using a modified Trizol/Trisure RNA isolation as described in Krosch and Bryant (2015). RNA quality and quantity were assessed on an Agilent 2100 Bioanalyser (Agilent Technologies, USA) and an RNA 6000 Nano kit for total RNA. Total RNA from replicates was pooled for cDNA library preparation, and paired-end (150 bp) sequencing was conducted according to manufacturer protocols and performed on an Illumina NextSeq 500 at the MGRF. All bioinformatic analyses were conducted on QUT's High Performance Computing Facility. Reads were quality assessed using FastQC (Andrews 2011), before trimming and assembly using default commands incorporating inbuilt Trimmomatic (Bolger *et al*. 2014) functionality in Trinity Version 20140413p1 (Haas *et al*. 2013). All new assemblies are available on the Transcriptome Shotgun Assembly (TSA) database associated with BioProject PRJNA385731. Assembly quality was assessed using the TrinityStats.pl script built into the Trinity distribution to attain basic summary statistics. *De novo* assemblies and downloaded gene sets were further compared for quality using two strategies: CEGMA analysis (Parra *et al*.

Table 1 Summary data for 14 transcriptomes analysed in this study

Species	GenBank BioProject accession	Reference	Life stage/gender	No. of raw reads (PE) used for assembly in Trinity	No. of transcript	N50	Mean transcript length (bp)	No. of full-length transcripts	No. of full-length hits to CEGMA's 248 core gene set	No. of reciprocal best hits to COG's in OrthoGraph
Bactrocera bryoniae	PRJNA385731	This study	Adult/♂, ♀	29 546 152	62 366	1983	1019	2940	241 (97%)	1559
Bactrocera dorsalis	PRJNA167923	Geib *et al.* (2014)	Mixed/♂, ♀	n/a	23 539	3460	2637	2702	224 (90%)	1524
Bactrocera frauenfeldi	PRJNA385731	This study	Adult/♂, ♀	21 164 218	62 990	1499	848	2751	238 (96%)	1538
Bactrocera jarvisi	PRJNA385731	This study	Adult/♂, ♀	29 167 616	87 840	1271	762	2717	232 (94%)	1543
Bactrocera kraussi	PRJNA385731	This study	Adult/♂, ♀	26 266 375	79 680	1384	811	2759	236 (95%)	1543
Bactrocera latifrons	PRJNA281765‡	USDA	Adult/♀	30 362 086	66 917	2144	1076	3037	244 (98%)	1579
Bactrocera minax	PRJNA244614	Dong *et al.* (2014)	Mixed/♂, ♀	n/a	47 189	1423	815	2602	218 (88%)	1519
Bactrocera musae	PRJNA385731	This study	Adult/♂, ♀	37 709 330	64 461	1813	956	3018	244 (98%)	1560
Bactrocera oleae	PRJNA195424	Pavlidi *et al.* (2013)	Mixed/♂, ♀	n/a	11 836	1016	827	1210	153 (62%)	849
Bactrocera tryoni	PRJNA227398†	Kumaran *et al.* (2014)	Adult/♂	n/a	37 097	1103	722	1657	159 (64%)	1404
Zeugodacus cucumis	PRJNA385731	This study	Adult/♂, ♀	22 797 947	56 850	1711	930	2950	247 (99%)	1555
Zeugodacus cucurbitae	PRJNA259566	Sim *et al.* (2015)	Mixed/♂, ♀	n/a	17 654	3477	2735	2885	234 (94%)	1504
Ceratitis capitata	PRJNA208956	Calla *et al.* (2014)	Mixed/♂, ♀	n/a	21 748	3914	3065	2857	230 (93%)	1547
Rhagoletis pomonella	PRJNA39555	Schwarz *et al.* (2009)	Mixed/♂, ♀	n/a	24 371	425	345	472	93 (37.5%)	846

Downloaded transcriptome assemblies were sequenced either from adults-only or mixed life stages (larvae, pupae and adults).

†We used a ***B. tryoni*** assembly held in-house at QUT, but associated with the listed BioProject and citation.

‡We used reads for ***B. latifrons*** from the listed BioProject and assembled them in-house.

2007) that searches for hits to 248 core eukaryote genes and determining the number of transcripts that span the full length of a gene according to BLASTX searches against the SwissProt database.

Candidate locus discovery

We developed a set of clusters of orthologous gene (COGs) by searching the OrthoDB7 database (http://cegg.unige.ch/orthodb7) for all single-copy orthologous genes found in all of the following reference species: *Drosophila ananassae*, *D. erecta*, *D. grimshawi*, *D. melanogaster*, *D. mojavensis*, *D. persimilis*, *D. pseudoobscura*, *D. sechellia*, *D. simulans*, *D. virilis*, *D. willistoni*, *D. yakuba*, *Glossina morsitans*, *Lutzomyia longipalpis* and *Phlebotomus papatasi*. *Bactrocera* transcriptomes were then searched for hits to reference COGs using OrthoGraph (https://github.com/mptrsen/Orthograph), following the recommended protocol. OrthoGraph uses a reciprocal BLAST approach to compare transcripts against reference COGs and reports only the 'best reciprocal hit' (BRH) for each COG. BRHs for each COG were summarised into separate amino acid and nucleotide FASTA files using built-in scripts, such that each file contained all BRH transcripts from each target species for a given COG. COGs that were deemed too short (<100 bp) to be useful diagnostic markers were excluded from further analysis.

Transcripts were aligned using MUSCLE Version 3.8.31 (nucleotide, Edgar 2004) and MAFFT Version 7.221 (amino acid, Katoh & Standley 2013) under default conditions, and length variable sections of alignments removed using GBLOCKS Version 0.91 (Castresana 2000) to minimise any effects of assembly error (manifested in length variable insertion–deletions, indel, regions, possibly resulting from chimeric transcripts) in estimating interspecific genetic diversity and phylogenetic relationships. We filtered COGs again to remove those that were deemed too short (<100 bp).

We assessed phylogenetic signal at each remaining locus as a proxy for diagnostic utility, on the basis that loci that produce topologies that match expectations from independent datasets should be highly accurate for species diagnosis. Thus, gene trees were inferred for all amino acid and nucleotide alignments in FastTree Version 2.1.8 (Price *et al.* 2009) and rerooted with either *C. capitata* or *R. pomonella* (*C. capitata* was used if both were present in a given alignment) using NewickUtils (Junier & Zdobnov 2010). Branch lengths were removed from resulting Newick tree strings, and all trees were compared against a set of expected topologies using simple text-based searches. Expected topologies were constructed based on accepted phylogenetic analyses for the dacines (Fig. 2), especially the recent densely sampled multi-locus works of Krosch *et al.* (2012) and Virgilio *et al.* (2015). Only loci that passed this filter for topology at both the amino acid and nucleotide levels were retained.

Nucleotide alignments for selected putative diagnostic loci were manually mapped to genomic sequences for *B. dorsalis* and *B. oleae* in BioEdit (Hall 1999) to identify exon–intron boundaries and assess intronic variation. This mapping process was used as a further filter to identify loci that possessed

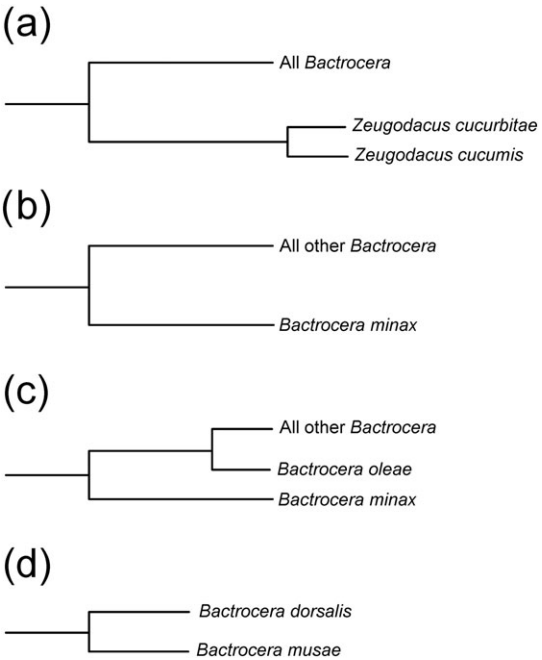

Fig. 2. Critical expected relationships among sampled species that were used to compare against inferred gene trees for putative diagnostic loci. These relationships are relative only to the taxa represented in the dataset and are based on the much more densely sampled phylogenies of Krosch *et al.* (2012) and Virgilio *et al.* (2015).

nucleotide spans that lacked substantial indels and possessed sufficient length (>400 bp) and variation to be developed as a PCR-based assay. Primers were designed using the online tool Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/) for such loci to determine the expected fragment length for each putative diagnostic locus. Average percent difference (as the inverse of average percent similarity) among the sampled tephritids was estimated as a proxy for average variation for each locus using T-Coffee (Notredame *et al.* 2000). We compiled a dataset of representative COI barcodes from GenBank for the same target species to provide a benchmark to compare the diagnostic utility of our newly developed markers against (Table S1). Loci that possessed greater average percent difference than COI (Table 2, Table S2) were selected for validation by PCR and Sanger sequencing among a small number of carefully chosen sibling species.

Experimental verification of diagnostic loci

As tests of the diagnostic capacity of candidate loci to resolve closely related and/or sibling species, we selected individuals of *B. dorsalis*, *B. carambolae* and *B. kandiensis*, and *B. tryoni* and *B. neohumeralis*, members of the *B. dorsalis* and *B. tryoni* species complexes, respectively (Table S3). Within each of the two species complexes, data from one of the test species (*B. dorsalis* and *B. tryoni*, respectively) formed part of the locus discovery workflow outlined in the previous texts, but the other three species had not been previously

Table 2 Alignment details for the five loci retained post-filtering

COG name (OrthoDB7)	Transcript alignment length (aa)	Transcript alignment length (nt)	Length of primed region (bp)	Average percent difference (nt)	*B. dorsalis* gene annotation	*B. dorsalis* genome accession
EOG7XDNSQ	177	534	500	20.65	Ribonuclease P protein subunit p29 (POP4)	NW_011876313
EOG735F6M	597	1794	650	16.75	Nodal modulator 1 (NOMO1)	NW_011876390
EOG71GN4X	592	1775	640	15.87	Carnitine *O*-palmitoyltransferase 2, mitochondrial (CPT2)	NW_011873997
EOG7F5BCP	186	567	500	15.29	Low-quality protein: replication protein A 32 kDa subunit (RPA2)	NW_011876190
EOG7V1S1K	607	1820	670	14.84	Dolichyl-diphosphooligosaccharide–protein glycosyltransferase subunit 2 isoform X2 (DDOSTs2)	NW_011876390

Alignment length is given in amino acids for protein alignments and base pairs for nucleotide alignments. Lengths of primed regions correspond to the region spanned by primers designed in Primer3 on the nucleotide alignment and are given in base pairs. Average percent difference was calculated on the nucleotide alignment. Gene function was assigned by BLAST searches to the *B. dorsalis* genome: Annotations and accessions are provided here.

assessed for these genes and thus represent a real test of the candidate gene's resolving power. Additionally, multiple individuals of each of the five species were included to test intraspecific variation. Representatives from sympatric populations of *B. dorsalis* and *B. carambolae,* and *B. tryoni* and *B. neohumeralis*, were included to minimise any influence of geographic differentiation. COI barcodes were available already for some of these specimens (GenBank Accessions in Table S1) and newly sequenced for the remainder.

All specimens were wild-caught adult males from locations that either spanned the known range of the species or were confirmed invasive locations (*B. carambolae* from Suriname). Total genomic DNA was extracted from three legs using an ISOLATE II DNA® kit (Bioline, Australia) according to manufacturer's protocol, and loci were amplified in an Eppendorf Mastercycler Pro (Eppendorf, Australia). Reaction recipes and thermocycling protocols for each locus are given in Tables S4 and 5. COI barcodes for new specimens were produced according to Schutze *et al*. (2015). Amplicons were purified using an ISOLATE PCR and Gel Kit® (Bioline) and direct sequenced using ABI Big Dye® Terminator 3.1 chemistry on an ABI 3500 Capillary Electrophoresis Genetic Analyser at the Molecular Genetics Research Facility at QUT. All sequences were checked and aligned by eye in BioEdit, average percent difference calculated in MEGA4 (Tamura *et al*. 2007). Phylogenies were reconstructed for each test locus incorporating all taxa (including those with transcriptome sequences) under both Bayesian inference (MrBayes Version 3.2.6, Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003) and maximum likelihood (RAxML Version 8.2.9, Stamatakis 2006), implemented on the CIPRES portal (Miller *et al*. 2010). All new sequences derived from target taxa were deposited in GenBank (accessions provided in Table S3).

RESULTS

Transcriptome sequencing undertaken in this study resulted in between 21 (21 164 218) and 38 (37 709 330) million high-quality reads per species, which were assembled into between 62 990 and 87 840 transcripts (Table 1). All indices of assembly quality suggested that *de novo* assemblies were quite rigorous, and comparisons with downloaded gene sets from *B. oleae*, *B. tryoni* and *R. pomonella* suggested that these published datasets possessed fewer core eukaryote genes with fewer full-length transcripts and were thus perhaps of lesser quality (Table 1); however, we considered all assemblies to be sufficient for downstream analyses. Average transcript length ranged from 345 (*R. pomonella*) to 3065 bp (*C. capitata*), and N50 statistics (a weighted median statistic that describes the transcript length that half of all transcripts in an assembly are equal to or larger than) ranged from 425 (*R. pomonella*) to 3914 (*C. capitata*).

Our set of target loci developed using OrthoDB7 contained 1646 COGs, of which a total of 1634 had a BRH to at least one target tephritid species, and 547 COGs had BRHs to transcripts from all 14 target species. Initial filtering by alignment length retained 1279 and 1005 loci at the nucleotide and amino acid levels, respectively. Gene trees from 619 of these loci matched the expected topology when analysed at the nucleotide level, while 134 loci matched in amino acid analyses, with only 69 loci inferring gene tree topologies that match the expected topology for both nucleotide and amino acid datasets. As a conservative approach to locus discovery and to limit the number of loci that needed exon mapping (refer to the succeeding texts), only loci for which the target topology was inferred from both nucleotide and amino acid datasets (69 total loci) were retained for subsequent analysis.

The 69 potential diagnostic loci were mapped against the assembled *B. dorsalis* and *B. oleae* genomes to determine exon–intron boundaries and gauge intron length differences among species. Aligned regions demonstrated that 57 of 69 loci were multi-exonic, and the intervening introns often varied in length between the two genomic references, suggesting a high probability of intron length variation across *Bactrocera* species. Primer design was therefore restricted to single exons (either loci that comprised only a single exon, or within a single selected exon for multi-exonic loci), to avoid developing targets that may require PCR amplification across large intronic indels. Further restricting our targets to loci for which the primed exonic region was greater than 400 bp resulted in 45 loci. Our final filter, that potential loci have average pairwise divergences greater than that of the COI barcode region for the corresponding species

(14.57%), resulted in just five loci (CPT2, NOMO1, RPA2, DDOSTs2, POP4, Table 2). These genes have functional roles in DNA replication, cell signalling, fatty acid oxidation, RNA processing and protein glycosylation, respectively.

The five candidate loci identified by bioinformatic means were then tested for their reliability in diagnostic settings by PCR amplification and Sanger sequencing: Primers are provided in Table 3. PCR amplification produced single bands for four of the five loci (RPA2, POP4, DDOSTs2, NOMO1); however, CPT2 produced several sub-bands, likely resulting from non-specific binding of primers. This locus was excluded from further consideration. POP4 was notably more difficult to amplify from *B. dorsalis* than other species; however, this could be associated with specimen condition and storage. Direct sequencing of PCR products resulted in final nucleotide alignments as follows: RPA2: 477 bp, POP4: 460 bp, DDOSTs2: 651 bp and NOMO1: 532 bp. Of these, both DDOSTs2 and NOMO1 showed less sequence differentiation than COI among selected species pairs (Table 4). However, RPA2 and POP4 show greater differentiation than COI between the close sibling species *B. dorsalis* and *B. carambolae*, but less than COI in all other pairwise comparisons. Phylogenies for POP4, DDOSTs2 and NOMO1 supported both the *B. dorsalis* and *B. tryoni* species complex members as monophyletic clades (Fig. 3, Figs S1 and S2); however, RPA2 could not resolve deeper relationships among species and did not support the *B. dorsalis* complex species as monophyletic (Fig. S3). By comparison, COI did not fit our species tree expectations (Fig. 2) with respect to the placement of *B. minax* and *B. oleae*: These were moderately supported as sister taxa, with these being sister to the two *Zeugodacus* species (Fig. 3). COI also does not resolve *B. dorsalis* as a well-supported monophyletic group, nor does it support *B. tryoni* or *B. neohumeralis* as monophyletic. Branch lengths for DDOSTs2 and NOMO1 were very shallow within each species complex, and generally, neither locus could resolve differences between members of each species complex (Figs S1 and S2). The locus POP4, on the other hand, resolves *B. dorsalis*, *B. carambolae* and *B. kandiensis* in strongly supported monophyletic clades, with *B. carambolae* and *B. kandiensis* as sister taxa and as a clade sister to *B. dorsalis*. As with COI, POP4 does not separate *B. tryoni* and *B. neohumeralis* as reciprocally monophyletic. While not 100% discriminatory at the species level, taken together, POP4 is supported by our analyses as meeting or exceeding the diagnostic capacity of COI and therefore most likely to hold potential as a new diagnostic marker for *Bactrocera* species.

Table 4 Average percent pairwise genetic distance between members of selected sibling pairs

	COI	POP4	RPA2	NOMO1	DDOSTs2
B. dorsalis/B. carambolae	1.97%	3.49%	1.41%	0.63%	0.31%
B. dorsalis/B. kandiensis	7.96%	5.71%	0.66%	0.64%	0.70%
B. carambolae/B. kandiensis	8.34%	4.13%	1.01%	0.79%	0.97%
B. tryoni/B. neohumeralis	4.14%	0.92%	0.32%	0.11%	0.04%

DISCUSSION

We present a novel analytical workflow for identifying diagnostic marker loci from the comparative analysis of transcriptome data. Comparative genomic approaches have been previously applied to marker discovery for phylogenetic and systematic studies, but attention to diagnostic applications has lagged behind. For instance, both anchored hybrid enrichment (AHE, Lemmon *et al.* 2012) and ultraconserved element methods (UCE, Faircloth *et al.* 2012) rely on identifying conserved genomic regions as targets for enrichment baits. Automated pipelines for identifying these regions have been developed (e.g. Baitfisher, Mayer *et al.* 2016; DISCOMARK, Rutschmann *et al.* in press) and indeed share considerable similarities with what was done here. The difference between such pipelines and the present study is that these studies did not assess suitability of individual loci. Marker discovery for systematic or phylogenetic purposes at the present is invariably undertaken with a view to downstream use in multi-locus analyses, with target loci obtained for sequencing either by PCR or genome reduction methods (e.g. AHE, UCE). Accordingly, the information content of individual loci is less important than the total information content of the combined suite of loci. In this way, marker discovery methods mirror contemporary phylogenetic practise where comparisons of individual gene tree compatibility, which were common 20 years ago, have given way to default concatenation or species tree coalescent methods.

The widespread adoption of any molecular diagnostic protocol will be subject to additional practical limitations beyond

Table 3 Primer details for five selected novel diagnostic loci

COG name (OrthoDB7)	Locus name	Primer name	Primer sequence (5′–3′)	T_m (°C)	Length of fragment (bp)
EOG7XDNSQ	POP4	POP4-f	ACATTACAATGTTGGAAGGGGG	55	520
		POP4-r	CTTYAYCTTYTTGACGCTGCG	55	
EOG7F5BCP	RPA2	RPA2-f	ACAAATCTTATATTCGCBTGAGGG	54	525
		RPA2-r	AATTTTTDTTGCAAYTCTTTGCGG	53	
EOG735F6M	NOMO1	NOMO1-f	TCATTTTCGATGAAGGYTCAAATT	52	570
		NOMO1-r	CGATATGATACTCACTTGCAAC	51	
EOG7V1S1K	DDOSTs2	DDOSTs2-f	GTGGCAGATCGTGTTGAAGA	53	695
		DDOSTs2-r	GGAACTTTAAAGGCCGATAATACTC	55	
EOG71GN4X	CPT2	CPT2-f	GAAGTGCTRATGATRTTGTGATTGA	53	645
		CPT2-r	ACGGARTYGCCGTACTAAGAT	55	

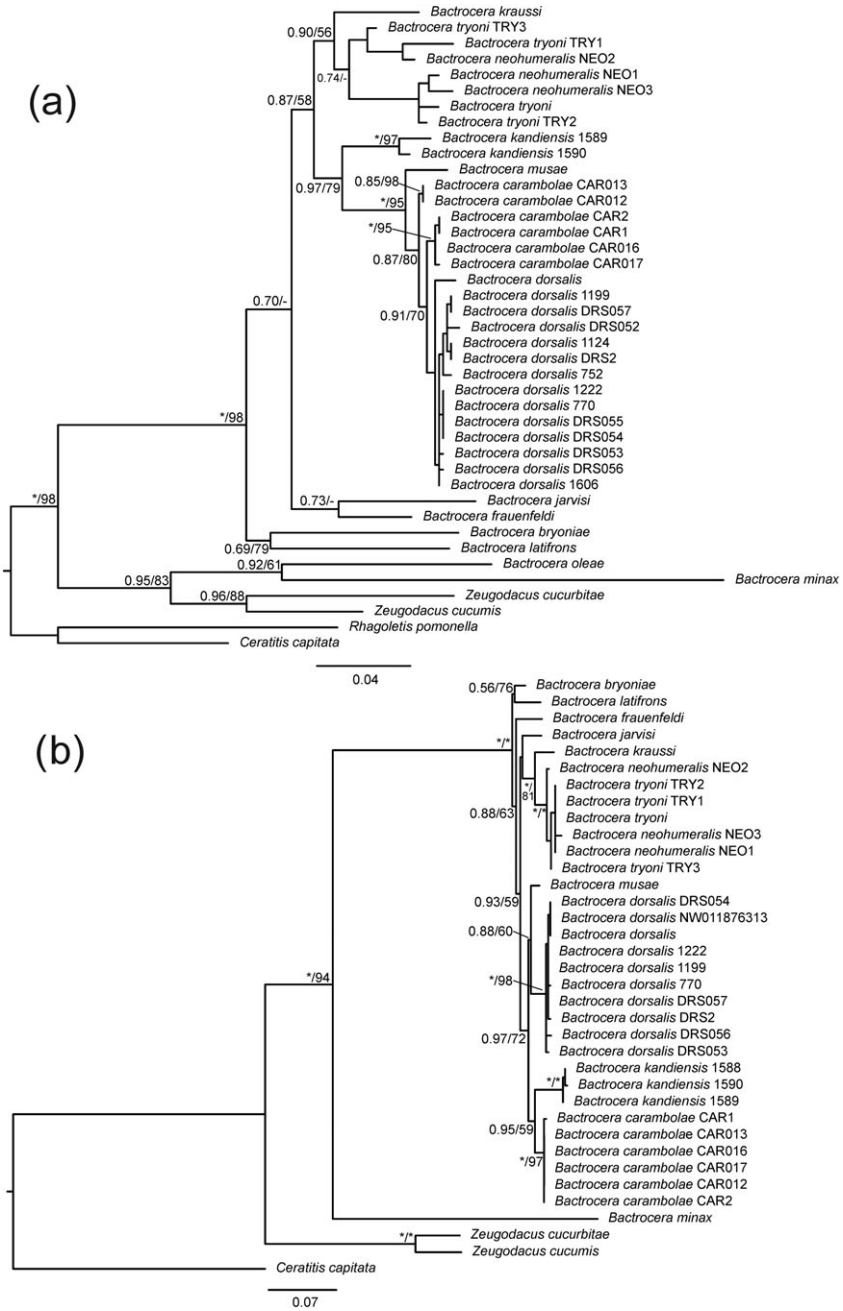

Fig. 3. Maximum likelihood topologies for (a) COI and (b) POP4. Node support values represent Bayesian posterior probabilities (left) and maximum likelihood bootstrap scores (right); posterior probabilities of 1.0 or bootstrap scores of 100 were relabeled as '*', and any scores <50 were coded as '-'. Sample codes and Genbank accession numbers correspond to Table S1 and S3. The *B. dorsalis* genomic scaffold used for mapping exon boundaries in POP4 is indicated by its Genbank accession.

those affecting research applications in systematics. In particular, methods that require specialised collecting techniques (e.g. RNA-Seq) or equipment (e.g. AHE, UCE) or are expensive on a per specimen basis (e.g. any method involving high-throughput sequencing) are unlikely to be widely adopted, especially for monitoring programs such as quarantine screening or invasive species detection. *Bactrocera* fruit flies are an example of taxa that are the subject of both identifications in routine interceptions from international trade and landscape-scale monitoring within horticultural districts. Practical diagnostic protocols for *Bactrocera* therefore need to be cheap, methodologically simple and accommodate the large species diversity in the genus. COI barcoding fulfils most of these criteria except for its demonstrated failure to discriminate between some *Bactrocera* species and known problems with nuclear-encoded copies or pseudogenes (Blacket *et al.* 2012). We therefore set out to find better barcodes by combining comparative genomic approaches known to identify phylogenetic markers with additional criteria to ensure that individual loci have diagnostic utility.

The main advantages of this marker discovery workflow are that it produces large numbers of candidate diagnostic loci relatively quickly and economically. We initially developed a set

of 1646 single-copy COGs from reference species available on OrthoDB7. Although more COGs may have been resolved had we developed a tephritid-specific set of orthologues, the general lack of detailed genomic data for the group resulted in a trade-off between maximising the number of loci and ensuring that loci were single copy. We opted for a conservative approach: 1646 well-characterised loci were considered a sufficiently large dataset to search against. Of these 1646 single-copy orthologues identified by comparative genomics, 69 loci had gene tree topologies that matched the phylogeny of the genus when analysed as both DNA and amino acid alignments. This initial test set of 69 loci is over 10 times the size of the largest previous study of *Bactrocera* phylogenetics based on arbitrarily chosen markers (Boykin *et al.* 2014) and over five times the size of the mitochondrial protein-coding gene set, which has also been the subject of marker discovery efforts (cf. Nelson *et al.* 2012). The remaining two filters – presence of a single exon longer than 400 bp and average nucleotide variability greater than the COI barcode – greatly reduced the number of candidate loci to just five markers. However, alternative measures could have been applied at each of these filtering steps to either open up or restrict the pool of candidate loci. For example, assessing gene tree topological congruence with the species tree under both nucleotide and amino acid analyses was much more conservative than assessing just nucleotides (69 vs. 619 candidate loci). Further, for other taxa where diagnostic failure of COI is due to causes other than simply low variation (e.g. retention of ancestral polymorphisms or mitochondrial introgression), raw variation may not be an imperative, and any locus with a gene tree congruent with the species tree could be experimentally assessed as a diagnostic marker.

As a new protocol for developing potential diagnostic loci, some comment on practicality is necessary. We estimate that to conduct the initial wet lab work for RNA sequencing to the identification of the five candidate loci for verification testing against a broad taxon set would take approximately 40 days of full-time work. Wet lab work represents only a small component of the total time, whereas the bulk of the time involves the post-sequencing bioinformatics. Given that the workflow largely utilises existing software, bioinformatic time is determined mostly by user proficiency and computational resources. Furthermore, the number of samples analysed will impact on the time frame of several stages of this workflow. Transcriptome assembly and OrthoGraph searches can be run in parallel for each sample (given sufficient computing resources) and are essentially scalable. However, more hands-on procedures, such as gene tree searches against a true tree and exon–intron mapping, will significantly increase the time frame as loci and samples increase.

The investment in initial comparative data for this method is quite modest relative to the number of loci that can be assessed; however, sampling needs to effectively capture the evolutionary diversity of the target group, as recommended for traditional phylogenetic studies (Nabhan & Sarkar 2012). Methods for the multiplex sequencing of RNA samples are now well developed, resulting in significant economies of scale. The six newly sequenced transcriptomes used in this study were pooled in a single Illumina NextSeq flow lane using the Mid-Output chemistry, and even larger numbers of specimens can be readily pooled, depending on available barcodes and desired sequencing depth, using the High-Output chemistry or on the Illumina HiSeq system. While sequencing costs vary between providers, the basic RNA-Seq data necessary for marker discovery through this workflow can be obtained at US$250–US$500 per sample. Mitochondrial genomes can be obtained extremely cheaply (US $20–US$50 per sample) when sequenced 90–100 at a time using pooled genomic DNA extracts (e.g. Gillett *et al.* 2014); however, this approach will at best yield 15 candidate loci prior to any filtering. In contrast, even low-pass genomic sequencing is expensive for all but the smallest animal genomes, and the sequencing of multiple genomes is probably not economically viable for most diagnostic discovery programs. RNA-Seq is thus a very economical method of collecting the input data for marker discovery.

Although the efficient *de novo* collection of genomic-scale data is critical for diagnostic marker discovery using this analytical workflow, the benefit of pre-existing genomic resources should also be noted. The mapping of exon boundaries (step 4 of the workflow) was only possible due to the previous sequencing of full genomes for two *Bactrocera* species. While the number of available draft genome sequences is rapidly expanding, they still represent a miniscule proportion of total species diversity, and it is likely that draft genomic sequences are not available for each taxon of diagnostic interest. In the absence of draft genomes, exons can be mapped against the nearest relative for which a genome has been sequenced, although the low conservation of exon–intron boundaries across larger taxonomic scales makes this approach problematic (e.g. Lohse *et al.* 2011). Alternatively, exon mapping could be omitted entirely. While the majority of candidate genes identified by step 3 (gene tree congruence) were multi-exonic (57 of 69), more than half (45 of 69) had single exons >400 bp long and were thus reasonable targets for a single PCR amplification. While exon–intron patterns vary considerably between taxa, designing multiple PCR priming sites for each candidate gene would allow the experimental verification of conserved exons of usable length. This approach is certainly less efficient than bioinformatic determination of exon lengths by genome mapping; however, the time spent on such a work-around could be less expensive than completing a draft genome sequence.

The second potential artefact in our analysis was the use of RNA-Seq data previously deposited on GenBank which had been generated by multiple labs, using different NGS platforms. It is well known that RNA-Seq studies vary considerably with the number and length of raw reads generated and that this has flow-on effects on the length, read depth and quality of the resulting RNA assemblies. In the present study, both *Rhagoletis pomonella* and *Bactrocera oleae* were represented by considerably smaller and seemingly poorer quality RNA-Seq datasets than the remaining species. Correspondingly, these two species returned the lowest numbers of reciprocal best hits in the OrthoGraph analysis, 846 and 849, respectively, only slightly more than half those found for the other 12 species (Table 1). Although it could be expected that the inclusion of a poor-quality transcriptome may exclude loci that are present in the species but not sequenced in that transcriptome, the approach used here

was conservative in this regard. Excluding *R. pomonella* and *B. oleae* from the OrthoGraph analysis results, only a single extra COG present in at least one species; however, the number of COGs with hits in all species rises considerably (1232 vs. 547). These differences did not have an effect in the current study, as gene tree analyses were conducted on all loci after filtering by alignment length, and the gene tree–species tree topology comparisons were conducted such that missing data for a given species did not result in a locus being scored as failing to return the expected topology. Removing candidate loci that are not present in all study species is a conservative and potentially more accurate approach, but would be dependent on high-quality input transcriptomes to avoid needless removal of loci. This issue should be addressed on a case-by-case basis; nevertheless, we show that variable quality data can be included and gene absences accounted for without impacting the downstream pool of candidate loci.

Although this analytical workflow was developed for the identification of nuclear protein-coding genes suitable as diagnostic markers, with limited modification, it can also be applied to discovering other gene types with diagnostic utility. A version of this workflow could be used to develop exon-primed intron-crossing (EPIC) markers. While existing EPIC marker discovery pipelines (e.g. Li *et al.* 2010) depend on genomic data to identify exon–intron boundaries, RNA-Seq data is a means of economically expanding the range of taxa assessed. RNA-Seq data can contribute to EPIC discovery in three ways: first, by identifying loci that are genuine 1:1 homologues across the target taxon (step 1 of our workflow). Second, alignments of coding region variation provide clues about how conserved exon–intron boundaries are within the target taxon (step 4). Finally, RNA-Seq data assists primer design by identifying highly conserved regions within exonic sequences (step 5). Modifying this workflow for EPIC discovery simply involves reordering step 3 (infer gene trees and compare to target phylogeny) and step 4 (exon mapping). Most previous studies that developed EPIC markers within insects have utilised pre-defined target genes (e.g. Lohse *et al.* 2011; White *et al.* 2015) rather than assessing locus variability as we advocate here. Additionally, most rely exclusively on genome data (e.g. Lardeux *et al.* 2012; White *et al.* 2015) and do not use RNA-Seq data to expand taxon coverage, although Lohse *et al.* (2011) successfully did so over broad taxonomic scales (multiple hymenopteran families). Given that the intron-like ribosomal RNA internal transcribed spacer (rRNA-ITS) has proven to be an effective species-level marker within *Bactrocera* (Boykin *et al.* 2014), EPICs may represent a useful additional source of diagnostic markers for this genus and will be the topic of future research.

We have emphasised here that novel barcode-like diagnostic markers would be analysed using tree-building methods, but ultimately, these markers can be analysed using the same variety of non-sequencing technologies as COI barcodes. Diagnostic sequence variation with the COI barcode region has been used to design restriction fragment length polymorphism (RFLP) tests to discriminate between species, particularly for industrial applications where routine screening far exceeds the economic use of sequencing-based tests (e.g. food identification Haider *et al.* 2012; Mueller *et al.* 2015). Similarly, species-specific real-time PCR systems have been designed for tephritid fruit flies based on COI barcode libraries (Dhami *et al.* 2016; Jiang *et al.* 2016). While all these systems have the same conceptual and/or practical weakness outlined in the previous texts for regular COI barcoding (as demonstrated in Jiang *et al.* 2016), these alternative technologies can be implemented for any genetic locus. Utilising the workflow detailed here to identify promising diagnostic markers and experimentally verify their utility within a target taxon will also improve their utility in downstream technologies such as RFLP and qPCR tests.

In conclusion, we outline an analytical workflow for utilising the most economical near-genomic-scale sequencing technology currently available (RNA-Seq) to identify the most promising species diagnostic markers. This workflow enables the assessment of a large number of potential markers relatively quickly and cheaply, while the filters built into it can be applied flexibly to expand or reduce the number of markers to be experimentally verified. Taxonomic groups for which molecular diagnostics are most needed frequently also have features that confound the use of arbitrarily chosen diagnostic markers, such as poorly defined species limits, introgression and/or retained ancestral polymorphisms (Rubinoff *et al.* 2006a). Searching a larger proportion of the genome for markers via RNA-Seq greatly improves the chances of identifying optimal diagnostic markers over arbitrary gene choice and greatly reduces the susceptibility of downstream diagnostic applications to failure due to choosing genes which cannot discriminate between species. We do not advocate that the loci identified here will accurately resolve species in taxonomic groups outside *Bactrocera*, and we do not yet recommend any of these loci as replacements for COI. We stress that the major import of this workflow is to provide a mechanism by which taxon-specific alternatives to COI can be developed. Validation of identified potential loci against a broader taxonomic sampling to determine their diagnostic utility is ongoing.

ACKNOWLEDGEMENTS

The authors would like to thank Jacinta McMahon (QUT School of Earth, Environmental and Biological Sciences) for sorting and sourcing flies from in-house collections for marker validation; Sahana Manoli, Kevin Dudley and Vincent Chand (QUT Central Analytical Research Facility) for sequencing assistance; Shorash Amin, Joachim Surm and Peter Prentis (QUT School of Earth, Environmental and Biological Sciences) for assistance with the analyses; Justin Lee (QUT High Performance Computing) for help with cluster scripting; Sybilla Oczkowicz and Peter Leach (Queensland Dept. of Agriculture, Fisheries) for providing *Bactrocera* samples; and Malte Petersen (Zoologisches Forschungsmuseum Alexander Koenig) for trouble shooting his program OrthoGraph. The data reported in this paper were partly obtained at the Central Analytical Research Facility, operated by the Institute for Future Environments (QUT) and QUT's High Performance Computing (HPC) Facility. Access to CARF and the HPC is supported by generous funding from the Science and Engineering Faculty (QUT). The authors acknowledge the

support of the Australian Government's Cooperative Research Centres Program. SLC was also supported by the Australian Research Council Future Fellowship Scheme (FT120100746).

REFERENCES

Andrews S. 2011. *FastQC: a Quality Control Tool for High Throughput Sequence Data*. Babraham Institute, Cambridge, UK.

Armstrong KF, Cameron CM & Frampton ER. 1997. Fruit fly (Diptera: Tephritidae) species identification: a rapid molecular diagnostic technique for quarantine application. *Bulletin of Entomological Research* **87**, 111–118.

Armstrong KF & Cameron CM. 1998. Species identification of tephritids across a broad taxonomic range using ribosomal DNA. In: *Area-wide control of fruit flies and other insect pests* (ed KH Tan), pp. 703–710. CABI Publishing, New York.

Armstrong KF & Ball SL. 2005. DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society B* **360**, 1813–1823.

Barr NB, Islam MS, Meyer MD & McPheron BA. 2012. Molecular identification of *Ceratitis capitata* (Diptera: Tephritidae) using DNA sequences of the CO1 barcode region. *Annals of the Entomological Society of America* **105**, 339–350.

Blacket MJ, Semeraro L & Malipatil MB. 2012. Barcoding Queensland fruitflies (*Bactrocera tryoni*): impediments and improvements. *Molecular Ecology Resources* **12**, 428–436.

Bolger AM, Lohse M & Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120.

Boykin LM, Schutze MK, Krosch MN *et al*. 2014. Multi-gene phylogenetic analysis of the south-east Asian pest members of the *Bactrocera dorsalis* species complex (Diptera: Tephritidae) does not support current taxonomy. *Journal of Applied Entomology* **138**, 235–253.

Calla B, Hall B, Hou S & Geib SM. 2014. A genomic perspective to assessing quality of mass-reared SIT flies used in Mediterranean fruit fly (*Ceratitis capitata*) eradication in California. *BMC Genomics* **15**, 98.

Carstens BC, Pelletier TA, Reid NM & Satler JD. 2013. How to fail at species delimitation. *Molecular Ecology* **22**, 4369–4383.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17**, 540–552.

Clarke AR & Schutze MK. 2014. The complexities of knowing what it is you are trapping. In: *Trapping and the Detection, Control, and Regulation of Tephritid Fruit Flies* (eds T Shelly, N Epsky, EB Jang, J Reyes-Flores & R Vargas), pp. 611–632. Springer, Dordrecht, The Netherlands.

Coissac R, Hollingsworth PM, Lavergne S & Taberlet P. 2016. From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology* **25**, 1423–1428.

Collins RA, Armstrong KF, Meier R *et al*. 2012. Barcoding and border biosecurity: identifying cyprinid fishes in the aquarium trade. *PLoS One* **7**, e28381.

Collins RA & Cruickshank RH. 2013. The seven deadly sins of DNA barcoding. *Molecular Ecology Resources* **13**, 969–975.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM & Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499–510.

De Salle R, Egan MG & Siddall M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B* **360**, 1905–1916.

Dhami MK, Gunawardana DN, Voice D & Kumarasinghe L. 2016. A real-time PCR toolbox for accurate identification of invasive fruit-fly species. *Journal of Applied Entomology* **140**, 536–552.

Dohino T, Hallman GJ, Grout TG *et al*. 2017. Phytosanitary treatments against *Bactrocera dorsalis* (Diptera: Tephritidae): current situation and future prospects. *Journal of Economic Entomology* **110**, 67–79.

Dong Y, Desneux N, Lei C & Niu C. 2014. Transcriptome characterization analysis of *Bactrocera minax* and new insights into its pupal diapause development with gene expression analysis. *International Journal of Biological Sciences* **10**, 1051–1063.

Donoghue MJ. 1985. A critique of the biological species concept and recommendations for a phylogcnetic alternative. *Bryologist* **88**, 172–181.

Dowton M, Meiklejohn K, Cameron SL & Wallman J. 2014. A preliminary framework for DNA barcoding, incorporating the multispecies coalescent. *Systematic Biology* **63**, 639–644.

Drew R, Romig MC 2013. *Tropical fruit flies of South-East Asia*. CABI Press, Wallingford, UK. 664 pp.

Dupuis JR, Roe AD & Sperling FAH. 2012. Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. *Molecular Ecology* **21**, 4422–4436.

Duran C, Appleby N, Vardy M, Imelfort M, Edwards D & Batley J. 2009. Single nucleotide polymorphism discovery in barley using autoSNPdb. *Plant Biotechnology Journal* **7**, 326–333.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT & Glenn TC. 2012. Ultraconserved elements anchore thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* **61**, 717–726.

Frey JE, Guillén L, Frey B, Samietz J, Rull J & Aluja M. 2013. Developing diagnostic SNP panels for the identification of true fruit flies (Diptera: Tephritidae) within the limits of COI-based species delimitation. *BMC Evolutionary Biology* **13**, 106.

Geib SM, Calla B, Hall B, Hou S & Manoukis NC. 2014. Characterizing the developmental transcriptome of the oriental fruit fly, *Bactrocera dorsalis* (Diptera: Tephritidae) through comparative genomic analysis with *Drosophila melanogaster* utilizing modENCODE datasets. *BMC Genomics* **15**, 942.

Gillett CP, Crampton-Platt A, Timmermans MJ, Jorda BH, Emerson BC & Vogler AP. 2014. Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution* **31**, 2223–2237.

Haas BJ, Papanicolaou A, Yassour M *et al*. 2013. *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature Protocols* **8**, 1494–1512.

Haider N, Nabulsi I & al-Safadi B. 2012. Identification of meat species by PCR-RFLP of the mitochondrial *CO1* gene. *Meat Science* **90**, 490–493.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95–98.

Hebert PD, Cywinska A, Ball SL & deWaard JR. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B* **270**, 313–321.

Hollingsworth PM, Graham SW & Little DP. 2011. Choosing and using a plant DNA barcode. *PLoS One* **6**, e19254.

Huelsenbeck JP & Ronquist F. 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* **17**, 754–755.

Jiang F, Jin Q, Liang L, Zhang AB & Li ZH. 2014. Existence of species complex largely reduced barcoding success for invasive species of Tephritidae: a case study in *Bactrocera* spp. *Molecular Ecology Resources* **14**, 1114–1128.

Jiang F, Fu W, Clarke AR *et al*. 2016. A high-throughput detection method for invasive fruit-fly (Diptera: Tephritidae) species based on microfluidic dynamic array. *Molecular Ecology Resources* **16**, 1378–1388.

Junier T & Zdobnov EM. 2010. The Newick Utilities: high-throughput phylogenetic tree processing in the UNIX Shell. *Bioinformatics* **26**, 1669–1670.

Katoh K & Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–780.

Kress WJ, Wurdack KJ, Zimmer EA, Weig LA & Janzen DH. 2005. Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences USA* **102**, 8369–8374.

Krosch MN, Schutze MK, Armstrong KF, Graham GC, Yeates DK & Clarke AR. 2012. A molecular phylogeny for the Tribe Dacini (Diptera: Tephritidae): systematic and biogeographic implications. *Molecular Phylogenetics and Evolution* **64**, 513–523.

Krosch MN & Bryant LM. 2015. A note on sampling chironomids for RNA-based studies of natural populations that retain critical morphological vouchers. *CHIRONOMUS Journal of Chironomid Research* **28**, 4–11.

Kumaran N, Prentis PJ, Mangalam KP, Schutze MK & Clarke AR. 2014. Sexual selection in true fruit flies (Diptera: Tephritidae): transcriptome and experimental evidences for phytochemicals increasing male competitive ability. *Molecular Ecology* **23**, 4645–4657.

Lardeux F, Aliaga C, Tejerina R & Ursic-Bedoya R. 2012. Development of exon-primed intron-crossing (EPIC) PCR primers for the malaria vector *Anopheles pseudopunctipennis* (Diptera: Culicidae). *Comptes Rendus Biologies* **335**, 398–405.

Lemmon AR, Emme SA & Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* **61**, 727–744.

Li C, Riethoven J-JM & Ma L. 2010. Exon-primed intron crossing (EPIC) markers for non-model teleost fishes. *BMC Evolutionary Biology* **10**, 90.

Lohse K, Sharanowski B, Blaxter M, Nicholls JA & Stone GN. 2011. Developing EPIC markers for chacidoid Hymenoptera from EST and genomic data. *Molecular Ecology Resources* **11**, 521–529.

Luo A, Zhang A, Ho SYW *et al*. 2011. Potential efficacy of mitochondrial genes for animal DNA barcoding. *BMC Evolutionary Biology* **12**, 84.

Mattock CJ, Morris MA, Matthijs G *et al*. 2010. A standardised framework for the validation and verification of clinical molecular genetic tests. *European Journal of Human Genetics* **18**, 1276–1288.

Mayer C, Sann M, Donath A *et al*. 2016. BaitFisher: a software package for multi-species target DNA enrichment probe design. *Molecular Biology and Evolution* **33**, 1875–1886.

Meier R, Shiyang K, Vaidya G & Ng PKL. 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* **55**, 715–728.

Miller MA, Pfeiffer W, Schwartz T. 2010 Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: *Proceedings of the Gateway Computing Environments Workshop (GCE)*, 14 Nov. 2010, New Orleans, LA pp 1–8.

Mishler BD. 1985. The morphological, developmental and phylogenetic basis of species concepts in bryophytes. *Bryologist* **88**, 207–214.

Mishler BD & Donoghue DJ. 1982. Species concepts: a case for pluralism. *Systematic Zoology* **31**, 491–503.

Morrow JL, Frommer M, Royer JE, Shearman DCA & Reigler M. 2015. *Wolbachia* pseudogenes and low prevalence infections in tropical but not temperate Australian fruit flies: manifestations of lateral gene transfer and endosymbiont spillover? *BMC Evolutionary Biology* **15**, 202.

Mueller S, Handy SM, Deeds JR *et al*. 2015. Development of a *COX1* based PCR-FRLP method for fish species identification. *Food Control* **55**, 39–42.

Muraji M & Nakahara S. 2002. Discrimination among pest species of *Bactrocera* (Diptera: Tephritidae) based on PCR-RFLP of the mitochondrial DNA. *Applied Entomology and Zoology* **37**, 437–446.

Nabhan AR & Sarkar IN. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics* **13**, 122–134.

Nelson LA, Lambkin CL, Batterham P *et al*. 2012. Beyond barcoding: genomic approaches to molecular diagnostics in blowflies (Diptera: Calliphoridae). *Gene* **511**, 131–142.

Nixon KC & Wheeler QD. 1990. An amplification of the phylogenetic species concept. *Cladistics* **6**, 211–223.

Notredame C, Higgins DG & Heringa J. 2000. T-Coffee: a novel method for multiple sequence alignments. *Journal of Molecular Biology* **302**, 205–217.

Parra G, Bradnam K & Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067.

Pavlidi N, Dermauw W, Rombauts S, Chrisargiris A, Van Leeuwen T & Vontas J. 2013. Analysis of the olive fruit fly *Bactrocera oleae* transcriptome and phylogenetic classification of the major detoxification gene families. *PLoS One* **18**, e66533.

Plant Health Australia. 2016. *The Australian Handbook for the Identification of Fruit Flies*, Version 2.1. Plant Health Australia. ACT, Canberra.

Pons J, Barraclough TG, Gomez-Zurita J *et al*. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* **55**, 595–609.

Price MN, Dehal PS & Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* **26**, 1641–1650.

Renner SS. 2016. A return to Linnaeus's focus on diagnosis, not description: the use of DNA characters in the formal naming of species. *Systematic Biology* **65**, 1085–1095.

Ronquist F & Huelsenbeck JP. 2003. MrBayes3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

Rosen DE. 1979. Fishes from the uplands and intermontane basins of Guatemala: revisionary studies and comparative geography. *Bulletin of the American Museum of Natural History* **162**, 267–376.

Rubinoff D, Cameron SL & Will K. 2006a. A genomic perspective on the shortcomings of mitochondrial DNA for barcoding and DNA taxonomy. *Journal of Heredity* **97**, 581–594.

Rubinoff D, Cameron SL & Will K. 2006b. Are plant DNA barcodes a search for the Holy Grail? *Trends in Ecology and Evolution* **21**, 1–2.

Rutschmann S, Detering H, Simon S, Fredslund J & Monoghan MT. 2017. DISCOMARK: nuclear marker discovery from orthologous sequences using draft genome data. *Molecular Ecology Resources* **17**, 257–266.

Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E & Crozier RH. 2010. Integrative taxonomy: a multisource approach to exploring biodiversity. *Annual Review of Entomology* **55**, 421–438.

Schutze MK, Mahmood K, Pavasovic A *et al*. 2015. One and the same: integrative taxonomic evidence that the African invasive fruit fly, *Bactrocera invadens* (Diptera: Tephritidae), is the same species as the Oriental fruit fly, *Bactrocera dorsalis*. *Systematic Entomology* **40**, 472–486.

Schutze MK, Virgilio M, Norrbom A & Clarke AR. 2017. Integrative taxonomy: where are we now, with a focus on the resolution of three tropical fruit fly species complexes. *Annual Review of Entomology* **62**, 147–164.

Schwarz D, Robertson HM, Feder JL *et al*. 2009. Sympatric ecological speciation meets pyrosequencing: sampling the transcriptome of the apple maggot *Rhagoletis pomonella*. *BMC Genomics* **10**, 633.

Seifert KA. 2009. Progress towards DNA barcoding of fungi. *Molecular Ecology Resources* **9**, 83–89.

Sim SB, Calla B, Hall B, DeRego T & Geib SM. 2015. Reconstructing a comprehensive transcriptome assembly of a white-pupal translocated strain of the pest fruit fly *Bactrocera cucurbitae*. *Gigascience* **4**, 14.

Sites JW & Marshall JC. 2004. Operational criteria for delimiting species. *Annual Review of Ecology, Evolution and Systematics* **35**, 199–227.

Song H, Buhay JE, Whiting MF & Crandall KA. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences USA* **105**, 13486–13491.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690.

Tamura K, Dudley J, Nei M & Kumar S. 2007. *MEGA4*: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**, 1596–1599.

Taylor HR & Harris WE. 2012. An emergent science on the brink of irrelevance: a review of DNA barcoding. *Molecular Ecology Resources* **12**, 377–388.

Vences M, Thomas M, Bonett R & Vieites DR. 2005. Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philosophical Transactions of the Royal Society of London B* **360**, 1859–1868.

Virgilio M, Jordaens K, Verwimp C, White IM & De Meyer M. 2015. Higher phylogeny of frugivorous flies (Diptera: Tephritidae: Dacinae): localised conflicts and a novel generic classification. *Molecular Phylogenetics & Evolution* **85**, 171–179.

White IM & Elson-Harris MM. 1992. *Fruit Flies of Economic Significance: Their Identification and Bionomics*. Canberra, Australia, ACIAR and Wallingford, U.K., CAB International.

White VL, Endersby NM, Chan C, Hoffmann AA & Weeks AR. 2015. Developing exon-primed intron-crossing (EPIC) markers for population genetic studies in three *Aedes* disease vectors. *Insect Sci.* **22**, 409–423.

Will KW & Rubinoff D. 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* **20**, 47–55.

Yong HS. 1995. Genetic differentiation and relationships in five taxa of the *Bactrocera dorsalis* complex (Insecta: Diptera: Tephritidae). *Bulletin of Entomological Research* **85**, 431–435.

Accepted for publication 17 October 2017.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

Figure S1 Maximum likelihood topology for NOMO1. Node support values represent Bayesian posterior probabilities (left) and maximum likelihood bootstrap scores (right); posterior probabilities of 1.0 or bootstrap scores of 100 were relabeled as '*', and any scores <50 were coded as '-'. Sample codes and Genbank accession numbers correspond to Supplementary Tables 1 and 2. The *B. dorsalis* genomic scaffold used for mapping exon boundaries in NOMO1 is indicated by its Genbank accession.

Figure S2 Maximum likelihood topology for DDOSTs2. Node support values represent Bayesian posterior probabilities (left) and maximum likelihood bootstrap scores (right); posterior probabilities of 1.0 or bootstrap scores of 100 were relabeled as '*', and any scores <50 were coded as '-'. Sample codes and Genbank accession numbers correspond to Supplementary Tables 1 and 2. The *B. dorsalis* genomic scaffold used for mapping exon boundaries in DDOSTs2 is indicated by its Genbank accession.

Figure S3 Maximum likelihood topology for RPA2. Node support values represent Bayesian posterior probabilities (left) and maximum likelihood bootstrap scores (right); posterior probabilities of 1.0 or bootstrap scores of 100 were relabeled as '*', and any scores <50 were coded as '-'. Sample codes and Genbank accession numbers correspond to Supplementary Tables 1 and 2. The *B. dorsalis* genomic scaffold used for mapping exon boundaries in RPA2 is indicated by its Genbank accession.

Table S1 Genbank accession numbers for representative sequences (COI) and transcript identifiers that had best reciprocal hits to five target genes (POP4, RPA2, NOMO1, DDOSTs2, CPT2), from all target species used for pairwise genetic distance estimates and phylogeny reconstruction.

Table S2 Alignment details for the remaining 40 loci retained post-filtering that did not have greater average percent difference than COI. Alignment length is given in amino acids for protein alignments and base pairs for nucleotide alignments. Lengths of primed regions correspond to the region spanned by primers designed in Primer3 on the nucleotide alignment and are given in base pairs. Average percent difference was calculated on the nucleotide alignment. Gene function was assigned by BLAST searches to the *B. dorsalis* genome: Annotations and accessions are provided here.

Table S3 Sample details and Genbank Accession numbers for selected sibling species representatives used to test the efficacy of newly developed putative diagnostic loci. Specimen codes correspond to tip labels in Figure 3. Gene acronyms are given in the text. CPT2 failed to amplify and was not sequenced. An 'x' denotes failed sequencing or PCR.

Table S4 Polymerase chain reaction details for each of the five selected novel diagnostic loci. Gene acronyms are given in the text.

Table S5 PCR cycle protocols for each of the five selected novel diagnostic loci. Gene acronyms are given in the text.