

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/236145257>

# DNA barcoding invasive insects: Database roadblocks

Article *in* Invertebrate Systematics · October 2012

Impact Factor: 2.26 · DOI: 10.1071/IS12025

CITATIONS

14

READS

237

4 authors:



[Laura M Boykin](#)

University of Western Australia

48 PUBLICATIONS 1,740 CITATIONS

[SEE PROFILE](#)



[Karen F. Armstrong](#)

Lincoln University New Zealand

81 PUBLICATIONS 1,513 CITATIONS

[SEE PROFILE](#)



[Laura S Kubatko](#)

The Ohio State University

51 PUBLICATIONS 2,293 CITATIONS

[SEE PROFILE](#)



[Paul J De Barro](#)

The Commonwealth Scientific and Industri...

139 PUBLICATIONS 4,553 CITATIONS

[SEE PROFILE](#)

# DNA barcoding invasive insects: database roadblocks

Laura M. Boykin<sup>A,D</sup>, Karen Armstrong<sup>A</sup>, Laura Kubatko<sup>B</sup> and Paul De Barro<sup>C</sup>

<sup>A</sup>Bio-Protection Research Centre, PO Box 84, Lincoln University, Lincoln 7647, New Zealand.

<sup>B</sup>Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, USA.

<sup>C</sup>CSIRO Ecosystem Sciences, GPO Box 2583, Brisbane, Qld 4001, Australia.

<sup>D</sup>Corresponding author. Email: lboykin@mac.com

**Abstract.** This study examines the genetic data coverage and availability in the Barcode of Life Database (BOLD), versions 2.5 and 3.0, and GenBank for the 88 invasive insects listed in the Global Invasive Species Database (<http://www.issg.org>). No data are recorded in either BOLD or GenBank for seven of those species. As a dedicated repository of curated barcode data BOLD is either missing data or contains inaccessible private data for 37 (42%) of the species while no data are available in GenBank for nine (8%) of the species. An evaluation of the Barcode Identification Number (BIN) scheme in BOLD ver. 3.0 was also evaluated and in 41% of cases the BIN contained more than one species. This essentially arose due to the 1% delimitation thresholds associated with the BINs and would result in misidentifications. Overall, more information is available from GenBank for the 88 invasive species listed on the Global Invasive Species Database, but quality checking is required to ensure that the data extracted from GenBank are of sufficient quality to make it useful. The implications of these results are discussed, with investment in parallel data silos suggested to be both costly and potentially an inefficient use of resources that may lead to loss of data if the means needed to maintain these databases become unavailable.

**Additional keywords:** Barcode of Life Database (BOLD), *Bemisia tabaci* Gennadius, CBOL, iBOL, QBOL.

Received 13 April 2012, accepted 16 September 2012, published online 19 December 2012

## Introduction

Invasive pests are a major threat to biodiversity, conservation and agriculture (Mack *et al.* 2000). The rapid and accurate identification of recent arrivals or recently established populations of such species is essential to maximise the likelihood of favourable biosecurity outcomes such as successful eradication. Many molecular-based methods for identifying such organisms, using various tools and target loci, have been developed for this purpose where conventional approaches are ineffective (Clarke *et al.* 2005; De Barro *et al.* 2011). However, consistent methods that can be applied to a range of targets are desirable to deal with the considerable increase in and globalisation of potential targets for identification (Bonants *et al.* 2010). In this respect, DNA barcoding has been promoted as a standard tool for the identification of economically important arthropods (Floyd *et al.* 2010). The strength of the method is use of the same genetic region to obtain DNA sequences that uniquely characterise each species (Hebert *et al.* 2003). Using this to distinguish native from invasive biota for habitats across the world would be of significant benefit to biosecurity, but its potential can only be realised with the appropriate level of species and global population coverage of data availability in DNA barcode libraries.

Armstrong and Ball (2005) were the first to publish a DNA barcoding-based study for species identification in a biosecurity

context and they concluded, among other things, that an agreed framework for the sharing and quality control of sequence data needed to be given serious consideration for adoption in such a critical application. Since then, numerous campaigns have continued to collect and register DNA barcodes (see Table 1), all of which are closely associated with the key consortia, being the Consortium for the Barcode of Life (CBOL, <http://www.barcodeoflife.org/>), the International Barcode of Life Project (iBOL, <http://ibol.org/>) and the Barcode of Life Database Systems (BOLD, <http://www.barcodeoflife.org/>), the latter being the common repository for data generated as well as an online workbench that aids collection, management, analysis, and use of DNA barcodes (Ratnasingham and Hebert 2007). Several of these campaigns incorporate groups of highly invasive species e.g. Tephritid Barcode Initiative (TBI), Quarantine Barcode of Life (QBOL), Mosquito Barcode Initiative (MBI) and HealthBOL focusing on vectors, pathogens, and parasites related to human health. Of these, QBOL has an overt biosecurity focus, aiming to obtain DNA barcode data for key species of plant health biosecurity significance, as based on European Union quarantine target lists across fungi, arthropods, bacteria, nematodes, viruses and phytoplasmas (Bonants *et al.* 2010). In addition, the working group for Agricultural and Forestry Pests and their Parasitoids in iBOL are planning to barcode a total of 25 000 species, including aphids, thrips, true fruit flies,

**Table 1. List of DNA barcoding campaigns compiled from sources collected on 7 March 2012**

Project Name	Target Species	Website
Source: <a href="http://www.barcodeoflife.org/">http://www.barcodeoflife.org/</a> :		
CBOL Protist Working Group (ProWG)	Protists	unable to locate electronically
Lepidoptera Barcode of Life	Butterflies and moths	<a href="http://lepbarcoding.org/">http://lepbarcoding.org/</a>
Trichoptera Barcode of Life (TrichopteraBOL)	Caddisflies	<a href="http://trichopteraBOL.org/">http://trichopteraBOL.org/</a>
Tephritid Barcode Initiative (TBI)	Fruit Flies	unable to locate electronically
Sponge Barcoding Project (SpongeBOL)	Sponges	<a href="http://www.spongebarcoding.org/">http://www.spongebarcoding.org/</a>
Shark Barcode of Life (SharkBOL)	Sharks	<a href="http://www.sharkbol.org/">http://www.sharkbol.org/</a>
Quarantine Barcode of Life (QBOL)	Plant pathogenic quarantine organisms	<a href="http://www.qbol.org/UK/">http://www.qbol.org/UK/</a>
Polar Barcode of Life (PolarBOL)	Bioinventory Arctic and Antarctic	unable to locate electronically
Mosquito Barcode Initiative (MBI)	Mosquitos	unable to locate electronically
Marine Barcode of Life (MarBOL)	Marine species	<a href="http://www.marinebarcoding.org/">http://www.marinebarcoding.org/</a>
Mammalia Barcode of Life Campaign	Mammals	<a href="http://www.mammaliabol.org/">http://www.mammaliabol.org/</a>
HealthBOL	Vectors, pathogens, and parasites	<a href="http://www.healthbol.org/">http://www.healthbol.org/</a> (no content)
Fish Barcode of Life Campaign (FISH-BOL)	Fish	<a href="http://www.fishbol.org/">http://www.fishbol.org/</a>
European Consortium for the Barcode of Life (ECBOL)	DNA barcoding in Europe	<a href="http://www.ecbol.org/">http://www.ecbol.org/</a>
Coral Reef Barcode of Life	Reef species	<a href="http://www.reefbarcoding.org/">http://www.reefbarcoding.org/</a>
Bee Barcode of Life Initiative (Bee-BOL)	Bees	<a href="http://www.bee-bol.org/">http://www.bee-bol.org/</a>
CBOL Fungal Working Group	Fungus	<a href="http://www.fungalbarcoding.org/">http://www.fungalbarcoding.org/</a>
All Birds Barcoding Initiative (ABBI)	Birds	<a href="http://www.barcodingbirds.org/">http://www.barcodingbirds.org/</a>
Source: <a href="http://lepbarcoding.org/">http://lepbarcoding.org/</a> :		
Formicidae barcode of life	Ants	<a href="http://www.formicidaebol.org/">http://www.formicidaebol.org/</a>
Termitidae Barcode of life	Termites	coming soon
Source: <a href="http://ibol.org/about-us/campaigns/">http://ibol.org/about-us/campaigns/</a> :		
Tipulidae Barcode of Life	Crane flies	coming soon
Odonata Barcode of Life	Dragonflies and damselflies	coming soon
Ephemeroptera Barcode of Life	Mayflies	coming soon
Lumbricidae Barcode of Life	Earthworms	coming soon

scale insects, sawflies and gall wasps (<http://ibol.org/about-us/campaigns/>), all of which are of biosecurity interest.

DNA barcoding has great potential as a species identification tool because it is practical and affordable to perform and, more often than not, shows species-level separations (Ross *et al.* 2008) that mirror established taxonomy. This is despite its use being subject to criticism, for example within the context of DNA taxonomy (Will *et al.* 2005; Rubinoff *et al.* 2006), the evolutionary pitfalls of mitochondrial DNA (Galtier *et al.* 2009), misidentifications based on distance thresholds (Cameron *et al.* 2006), single versus multiple loci (Elias *et al.* 2007), and its potential technological obsolescence (Taylor and Harris 2012), all of which could affect its utility for the identification of invasive species. Not least, however, is the lack of sufficient taxonomic coverage in the reference datasets leading to vulnerability to Type II errors (misidentification of queries without conspecifics in the database). This risk was a key concern in a recent study testing the performance of DNA barcoding and BOLD across insect orders (Virgilio *et al.* 2010). A subsequent study on invasive African fruit flies tested the identification of an unknown using an incomplete DNA barcode library (Virgilio *et al.* 2012). They outlined a general working strategy to deal with incomplete libraries. However, this relied heavily on the matching criteria chosen and still produced an error rate of 5%, which in biosecurity is likely to be unacceptably high when decisions on market access are to be made.

Given the need for appropriate taxonomic coverage of reliable reference barcodes, how useful are current databases for identifying species of biosecurity concern? To explore this, the

current utility of two key global databases, the Barcode of Life Database (BOLD vers 2.5 and 3.0) and GenBank as sources of data relevant to invasive insect species are examined. Given that there are many thousands of invasive species, we limit our assessment to the 88 highly invasive insect species listed in the Global Invasive Species Database (<http://www.issg.org>).

## Materials and methods

### Databases

All 88 insect species listed in the Global Invasive Species Database (<http://www.issg.org/>) as highly invasive were included in this study. BOLD ver. 2.5 (<http://www.boldsystems.org/>), BOLD ver. 3.0 (<http://v3.boldsystems.org/>) and GenBank (Taxonomy search) were queried for availability of all DNA sequence data for each of the species listed in Table 2. BOLD ver. 2.5 was replaced by ver. 3.0 in April 2012. Data collected from BOLD ver. 2.5 included numbers of specimen records, specimens with sequences, specimens with barcodes, and public records. BOLD ver. 3.0 data collection included, in addition to the ver. 2.5 dataset, the public Barcode Identification Number (BIN) and the number of times a BIN contained a species that was not the query sequence (possibly leading to misidentification). GenBank taxonomy browser data collected included the number of GenBank nucleotide records and all gene regions. Species with no records in BOLD, GenBank or both were investigated for synonyms that might be misleading the search for DNA sequences. Subsequent synonyms were searched in both databases and no sequence data

Table 2. Sequence data available (compiled March 2012) in BOLD versions 2.5 and 3.0 and GenBank for the 88 most invasive insects identified by the global invasive species database (<http://www.issg.org/>)

\*\* indicates there are no sequences in GenBank or BOLD

Species	BOLD ver. 2.5: specimens with sequences	BOLD ver. 2.5: public records	BOLD ver. 3.0: specimens with sequences	BOLD ver. 3.0: public sequences	Difference between ver. 2.5 and 3.0	BOLD ver. 3.0: public BINs	BOLD ver. 3.0: wrong species in BIN	Gene regions in GenBank	GenBank nucleotide records	Difference in BOLD ver. 3.0 and GenBank public sequences
<i>Acnemia bifida</i> Zaitzev**	0	0	0	0	0	0	0	None	0	0
<i>Acronymex octospinosus</i> Reich	0	0	0	0	0	0	0	Efla and COI-COII	25	-25
<i>Adelges piceae</i> Ratzeburg	14	13	15	13	0	1	1	Barcode	19	-6
<i>Adelges tsugae</i> Annand	174	119	174	119	0	4	0	Barcode, COII, cytb, ND1, EFla, microsats	222	-103
<i>Aedes aegypti</i> Linnaeus	97	31	92	31	0	3	1	Many	32 749	-32 718
<i>Aedes albopictus</i> Skuse	77	22	76	26	4	1	1	Many	68 773	-68 747
<i>Agrius planipennis</i> Fairmaire	18	3	18	4	1	0	0	Barcode, microsats, 16S	60	-56
<i>Anopheles quadrimaculatus</i> Say	94	12	94	12	0	1	0	Barcode, single copy nt genes	23	-11
<i>Anoplolepis gracilipes</i> Fr. Smith	65	25	66	26	1	1	0	Barcode, EFla, cytb, 28S, wng, microsats	62	-36
<i>Anoplophora chinensis</i> Forster	5	5	5	5	0	1	1	Barcode, 16S, COI-COII	58	-53
<i>Anoplophora glabripennis</i> Motschulsky	439	430	439	430	0	1	1	Barcode, 28S, 16S, COI-COII, ND4, microsats	469	-39
<i>Anthonomus grandis</i> Boheman	23	22	23	22	0	0	0	Many (ITS1 and 2, COI, COII)	2785	-2763
<i>Apis mellifera scutellata</i> Lepelletier	51	0	51	46	46	1	0	cytb, ITS1 and 2	9	37
<i>Aulacaspis yasumatsui</i> Takagi	0	0	0	0	0	0	0	Barcode, 28S, Efla	3	-3
<i>Bactrocera tryoni</i> Froggatt	11	3	11	3	0	1	1	Barcode, cytb, COI, CAD, transposons, microsats	161	-158
<i>Bemisia tabaci</i> Gemadius	200	193	199	153	-40	8	0	Many	134 529	-134 376
<i>Brontispa longissima</i> Gestro	0	0	0	0	0	0	0	microsats, COI, p450	63	-63
<i>Cactoblastis cactorum</i> Berg	59	1	59	1	0	0	0	Barcode, COI	136	-135
<i>Calliphora vicina</i> Robineau-Desvoidy	30	23	30	26	3	1	1	Barcode, cytb, 16S 5.8S, ITS2	239	-213
<i>Ceratitis capitata</i> Wiedemann	170	108	170	122	14	2	1	Barcode, microsats, ND4, ND5	1144	-1022
<i>Cinara cupressi</i> Buckton	1	0	2	0	0	0	0	Barcode, ATP6	2	-2
<i>Compsilura concinnata</i> Meigen	2	0	2	0	0	0	0	None	0	0
<i>Coptotermes formosanus</i> Shiraki	0	0	0	0	0	0	0	16S, COII, 18S, 5.8S, 12S	292	-292
<i>Culex quinquefasciatus</i> Say	113	33	113	34	1	1	1	Barcode, 28S, COI, COII, microsats, EFla, 5.8S, 28S	26 021	-25 987
<i>Dendroctonus valens</i> LeConte	8	6	8	6	0	0	0	Barcode, COI, 5.8S, 28S	171	-165
<i>Dendrolimus sibiricus</i> Chetverikov	7	6	7	6	0	0	0	Barcode, ITS2, COI	24	-18
<i>Diaphorina citri</i> Kuwayama	211	210	211	210	0	1	1	Barcode, CYP4DB1, COI, microsats	360	-150
<i>Dysdera crocata</i> C. L. Koch	11	7	11	7	0	2	1	Barcode, 28S, H3, 16S	23	-16
<i>Gonipterus scutellatus</i> Gyllenhal**	0	0	0	0	0	0	0	Barcode	0	0
<i>Harmonia axyridis</i> Pallas	75	18	78	18	0	1	0	Barcode, 16S, 12S	343	-325

(continued next page)

Table 2. (continued)

Species	BOLD ver. 2.5: specimens with sequences	BOLD ver. 2.5: public records	BOLD ver. 3.0: specimens with sequences	BOLD ver. 3.0: public sequences	Difference between ver. 2.5 and 3.0	BOLD ver. 3.0: public BINs	BOLD ver. 3.0: wrong species in BIN	Gene regions in GenBank	GenBank nucleotide records	Difference in BOLD ver. 3.0 and GenBank public sequences
<i>Hemiberlesia pitysophila</i> Takagi**	0	0	0	0	0	0	0	None	0	0
<i>Homalodisca viripennis</i> Gemar	9	3	5	3	0	1	0	Barcode, genes from an EST library	69	-66
<i>Hoplochelus marginalis</i> Fairmaire**	0	0	0	0	0	0	0	None	0	0
<i>Hylastes ater</i> Paykull	4	0	4	0	0	0	0	28S	2	-2
<i>Hyphantria cunea</i> Drury	194	69	194	69	0	2	1	Barcode, pgi	142	-73
<i>Icerya purchasi</i> Maskell	6	1	6	1	0	0	0	Barcode, 18S, 28S, EF1a	10	-9
<i>Ips typographus</i> DeGreer	4	1	4	1	0	1	0	Barcode, microsats, COI, 12S	117	-116
<i>Lasius neglectus</i> Van Loon	19	18	19	19	1	1	0	Barcode, COI, microsats	24	-5
<i>Linepithema humile</i> Mayr	66	22	66	25	3	1	0	Genomic scaffold, wnt-1, EF1a, 28S, 18S	6570	-6545
<i>Lymantiria dispar</i> Linnaeus	267	6	268	193	187	1	1	Barcode, COI	333	-140
<i>Lymantiria mathura</i> Moore	26	0	26	26	26	1	1	Barcode	28	-2
<i>Lymantiria monacha</i> Linnaeus	77	4	77	63	59	1	1	Barcode	77	-14
<i>Maconellicoccus hirsutus</i> Green	5	1	5	5	4	1	0	COI, microsats, EF1a	342	-337
<i>Monomorium destructor</i> Jerdon	59	1	59	7	6	2	2	Barcode, 28S	13	-6
<i>Monomorium floricola</i> Jerdon	114	0	114	7	7	1	1	Barcode	7	0
<i>Monomorium pharaonis</i> Linnaeus	48	0	48	0	0	0	0	microsats, EF1a	11	-11
<i>Myrmica rubra</i> Linnaeus	72	64	72	64	0	1	1	Barcode, EF1a, cytb, ITS1	179	-115
<i>Nylanderia</i> (= <i>Paratrechina</i> ) <i>pubens</i> Forel**	0	0	0	0	0	0	0	None	0	0
<i>Ochlerotatus japonicus japonicus</i> Theobald	0	0	0	0	0	0	0	NADH, 28S, COI, microsats	30	-30
<i>Oopteris soledadinus</i> Guerin-Ménéville	14	14	14	14	0	0	0	Barcode, 16S, microsat, COI	75	-61
<i>Oracella acuta</i> (= <i>Pseudococcus acutus</i> ) Lobdell**	0	0	0	0	0	0	0	None	0	0
<i>Orthetia insignis</i> Browne	5	0	5	0	0	0	0	None	0	0
<i>Orthotomicus erosus</i> Wollaston	0	0	0	0	0	0	0	COI, EF1a,	5	-5
<i>Oryctes rhinoceros</i> Linnaeus	0	0	0	0	0	0	0	Random genes	14	-14
<i>Pachycondyla chinensis</i> Emery	36	34	36	34	0	2	1	Barcode	38	-4
<i>Paratarchardina pseudolobata</i> Kondo & Gullan**	0	0	0	0	0	0	0	None	0	0
<i>Paratrechina longicornis</i> Latreille	104	6	104	9	3	2	2	Barcode, microsats, 18S, 28S, CAD, EF1a	82	-73
<i>Pheidole megacephala</i> Fabricius	233	45	233	48	3	1	1	Barcode, cytb, 28S, microsats	81	-33
<i>Philornis downsi</i> Dodge & Aitken	0	0	0	0	0	0	0	microsats, 5.8S, 18S ITS1 and 2	13	-13
<i>Polistes chinensis</i> Perez	5	3	5	5	2	1	0	Barcode, COI, microsat	30	-25
<i>Quadrastichus erythrinae</i> Kim	1	1	1	1	0	0	0	Barcode, COI	7	-6
<i>Radumeris tasmaniensis</i> Saussure	1	0	1	1	1	1	0	None	0	1
<i>Rhynchohorus ferrugineus</i> Olivier	0	0	0	0	0	0	0	Microsats, cytb, COI	387	-387
<i>Scolytus multistriatus</i> Marsham	9	1	9	1	0	0	0	Barcode, 18S	3	-2
<i>Scyphophorus acupunctatus</i> Gyllenhal	5	1	5	1	0	1	0	Barcode, 28S, EF1a	3	-2

(continued next page)

Table 2. (continued)

Species	BOLD ver. 2.5: specimens with sequences	BOLD ver. 2.5: public records	BOLD ver. 3.0: specimens with sequences	BOLD ver. 3.0: public sequences	Difference between vers 2.5 and 3.0	BOLD ver. 3.0: public BINs	BOLD ver. 3.0: wrong species in BIN	Gene regions in Genebank	GenBank nucleotide records	Difference in BOLD ver. 3.0 and GenBank public sequences
<i>Sirex noctilio</i> Fabricius	17	1	17	1	0	1	0	Barcode, 16S, 18S	14	-13
<i>Solenopsis geminata</i> Fabricius	100	26	100	27	1	1	1	Barcode, odorant binding protein, 5.8S, ITS1 and 2	59	-32
<i>Solenopsis invicta</i> Buren	4	1	4	2	1	1	1	Barcode, many	69 988	-69 986
<i>Solenopsis papuana</i> Emery	4	0	4	0	0	0	0	None	0	0
<i>Solenopsis richteri</i> Forel	1	0	1	0	0	0	0	Barcode, genome	35	-35
<i>Tapinoma melanocephalum</i> Fabricius	141	0	141	1	1	1	0	EF1a, wing, 28S, COI	11	-10
<i>Technomyrmex albipes</i> Smith	133	26	133	27	1	3	1	Barcode, 18S, wing,	37	-10
<i>Tetropium fuscum</i> Fabricius	3	0	3	0	0	0	0	Barcode, 5.8S, ITS1 and 2, 28S	15	-15
<i>Thaumetopoea pityocampa</i> Denis & Schifferrmüller	7	0	7	1	1	1	0	Barcode, microsats, COII,	148	-147
<i>Tomocis piniperda</i> Linnaeus	42	41	42	41	0	0	0	Barcode, 28S, COI, EF1a, COII	161	-120
<i>Toumeyella parvicornis</i> Cockerell	2	0	2	0	0	0	0	None	0	0
<i>Trechisibus antarcticus</i> Dejean	21	21	21	21	0	0	0	Barcode, 16S	85	-64
<i>Trogoderma granarium</i> Everts	4	0	4	0	0	0	0	Barcode, cytb	14	-14
<i>Vespa velutina</i> de Buysson	7	3	7	3	0	0	0	Barcode	7	-4
<i>Vespula germanica</i> Fabricius	12	2	12	2	0	1	1	Barcode, 28S, 16S	21	-19
<i>Vespula pensylvanica</i> Rohwer	0	0	0	0	0	0	0	18S, 5.8S, ITS2	4	-4
<i>Vespula vulgaris</i> Linnaeus	14	0	14	1	1	1	1	Barcode, other gene products, microsat	58	-57
<i>Wasmannia auropunctata</i> Roger	107	90	107	90	0	5	0	Barcode, wing, EF1a, microsats, COII, cytb	171	-81
<i>Xyleborus glabratus</i> Eichhoff	0	0	0	0	0	0	0	Barcode, CAD, 28S	3	-3
<i>Xylosandrus compactus</i> Eichhoff	0	0	0	0	0	0	0	Barcode, CAD, EF1a, 28S	9	-9
<i>Xylosandrus mutilatus</i> Blandford	0	0	0	0	0	0	0	CAD, EF1a, COI, 28S	6	-6
Total	3957	1796	3953	2134	338	68	28	None	348 303	-346 169



were recovered for the synonyms listed in Table 2. The QBOL (Q-Bank: <http://www.q-bank.eu/>) database was excluded because the database is not fully populated and DNA sequence data cannot yet be extracted.

#### Example query: *Bemisia tabaci*

Of the species on the Global Invasive Species list, *Bemisia tabaci* (authorities are listed in Table 2 for all species) was selected as a test case for querying the BOLD and GenBank databases because there is substantial understanding of the biology, taxonomy and phylogeny of this species (Liu *et al.* 2007, 2012; Boykin *et al.* 2007, 2012; Dinsdale *et al.* 2010; De Barro and Ahmed 2011; De Barro *et al.* 2011) to assist in interpretation of the results. A known sequence from the global population alignment of *B. tabaci* (maintained by PDB and LMB) was extracted and used as the query specimen (MidEastAm1\_Syria\_Al\_Hasakeh\_Hap2\_AB\_473559) to test how well BOLD and GenBank identify it. Other information that could be extracted from the search results (for example: authors, host information, publication citation, etc) was also collected.

## Results

There are more DNA sequences available in GenBank (348 303) than BOLD ver. 2.5 (3957) or BOLD ver. 3.0 (3953) for the 88 species described in Table 2. Twelve species have no GenBank records and 16 have no BOLD records, while seven have no DNA sequence records in either BOLD or GenBank (Table 2). In addition, there are 21 species in BOLD for which the sequences are private and were therefore classed as not being present for the purposes of this study; this increased the total to 37 of the 88 not being represented in BOLD (Table 2).

#### BOLD vers 2.5 and 3.0

GISD species missing data in BOLD include *Acnemia bifida*, *Acromyrmex octospinosus*, *Brontispa longissima*, *Goniapterus scutellatus*, *Hemiberlesia pitysohila*, *Hoplochelus marginalis*, *Nylanderia* (= *Paratrechina*) *pubens*, *Ochlerotatus japonicus japonicus*, *Oracella acuta*, *Orthotomicus erosus*, *Oryctes rhinoceros*, *Paratachardina pseudolobata*, *Philornis downsi*, *Rhynchophorus ferrugineus*, *Vespula pensylvanica*, *Xylosandrus mutilates*.

GISD species with data in BOLD, but not open to the public include *Apis mellifera scutellata*, *Aulacaspis yasumatsui*, *Cinara cupressi*, *Hylastes ater*, *Lymantria mathura*, *Monomorium floricola*, *Monomorium pharaonis*, *Orthezia insignis*, *Radumeris tasmaniensis*, *Solenopsis papuana*, *Solenopsis richteri*, *Tapinoma melanocephalum*, *Tetropium fuscum*, *Thaumetopoea pityocampa*, *Toumeyella parvicornis*, *Trogoderma granarium*, *Vespula vulgaris*.

Several issues were encountered when searching BOLD ver. 3.0. For example, the BIN numbers for each species are not consistent when using the 'taxonomy' and 'BIN' search options. The taxonomy search appears to be adding BINs to each species. Specifically, a taxonomy search for *Aedes aegypti* returns three BINs while the BIN search options reports two BINs only. There also appears to be an inconsistency with the results returned for the number of publically available sequences using the taxonomy and public data search options. For example, a taxonomy search in

BOLD for *L. dispar* yielded six public sequences while the public database option search returned 193 sequences (see Table 3 for more examples). Another concern noted for BOLD ver. 3.0 was the number of times the BINs contained a species other than the query sequence. A total of 68 BINs were found for the 88 species and in 28 cases there was a species other than the query sequence in the BIN, meaning that 41% of the time a misidentification of the query sequence could occur for these 88 species.

For 21 species, BOLD ver. 3.0 has more sequence data than GenBank (*Pheidole megacephala*, *Tapinoma melanocephalum*, *Monomorium floricola*, *Techomyrmex albipes*, *Anopheles quadrimaculatus*, *Hyphantria cunea*, *Monomorium destructor*, *Apis mellifera scutellata*, *Solenopsis geminate*, *Monomorium pharaonis*, *Paratrechina longicornis*, *Scolytus multistriatus*, *Orthezia insignis*, *Anoplolepis gracilipes*, *Solenopsis papuana*, *Sirex noctilio*, *Compsilura concinnata*, *Hylastes ater*, *Scyphophorus acupunctatus*, *Toumeyella parvicornis* and *Radumeris tasmaniensis*), but data for 11 of these are private and so effectively unavailable.

#### GenBank

GISD species missing data in GenBank include *Acnemia bifida*, *Compsilura concinnata*, *Hoplochelus marginalis*, *Nylanderia* (= *Paratrechina*) *pubens*, *Oracella acuta* (= *Pseudococcus acutus*), *Orthezia insignis*, *Paratachardina pseudolobata*, *Radumeris tasmaniensis*, *Solenopsis papuana*, *Toumeyella parvicornis*.

All GISD species with data in GenBank are publically available.

GISD species with no records in either GenBank or BOLD include *Acnemia bifida*, *Goniapterus scutellatus*, *Hemiberlesia pitysohila*, *Hoplochelus marginalis* (= *Empecta nudiplaga*), *Nylanderia* (= *Paratrechina*) *pubens*, *Oracella acuta* (= *Pseudococcus acutus*), *Paratachardina pseudolobata*.

#### DNA barcoding campaigns

DNA barcoding campaigns are listed in Table 1. The following campaigns are of interest to this study: Lepidoptera Barcode of Life, the Tephritid Barcode Initiative, the Mosquito Barcode Initiative, Bee Barcode of Life Initiative, Formicidae Barcode of Life, Termitidae Barcode of Life. All of the data collected from these campaigns are, or will be, deposited in BOLD. Table 1 also

**Table 3. Species for which there is a discrepancy between the numbers of sequences available using the taxonomy search and public data search options (GenBank ver. 3.0)**

Species	No. of sequences found using the taxonomy search option	No. of sequences found using the public data search option
<i>Lymantria dispar</i>	6	193
<i>Lymantria mathura</i>	0	26
<i>Lymantria monarcha</i>	4	63
<i>Monomorium destructor</i>	1	7
<i>Monomorium floricola</i>	0	7
<i>Paratrechina longicornis</i>	6	9
<i>Pheidole megacephala</i>	45	48

contains the Quarantine Barcode of Life (QBOL) campaign that has its own database, Q-Bank, but this is not publically available. Peter Bonants (Coordinator QBOL and Q-bank) states, 'The databases are meant for NPPOs (National Plant Protection Organisations) to correctly identify quarantine species using the Q-bank database' (pers. comm.) and are therefore not available to the public.

#### *Example query: Bemisia tabaci*

As of 7 March 2012 there were 3127, 193 and 153 publically available mtCOI sequences in GenBank, BOLD ver. 2.5 and BOLD ver. 3.0, respectively. The query DNA sequence from the global population alignment of *B. tabaci* for MidEastAm1\_Syria\_Al\_Hasakeh\_Hap2\_AB473559 resulted in the identification results for BOLD vers 2.5 and 3.0 and GenBank. Beyond the first few hits, the BOLD searches are substantially different from the GenBank results (data not shown). For example, all of the top hits in the GenBank search are *B. tabaci* in comparison to the BOLD results where other genera and species are listed, and likely reflects the fewer barcode sequences in BOLD than in GenBank. Second, *B. tabaci* is a cryptic species complex (Dinsdale *et al.* 2010; De Barro and Ahmed 2011; De Barro *et al.* 2011; Boykin *et al.* 2012) and so the additional information regarding the genetic groups can be obtained from the GenBank sequences via the author codes assigned to the sequences, while this information is lacking in BOLD. For the *B. tabaci* community, a database is needed to store all of the curated 3' COI data. This is the gene region that is used to identify unknowns and conduct phylogenetics – ignoring these data will lead to incorrect identifications.

## Discussion

Change is needed for DNA barcoding to remain relevant to biosecurity, especially regarding data handling of invasive insects. Tables 1 and 2 show that there is room for improvement in species coverage and data availability for these 88 species for both BOLD and GenBank. Researchers query these databases for several reasons, but the two reasons of interest for this study are (1) identification of an unknown and (2) ability to obtain sequences to conduct further studies, i.e. phylogenetic analyses. Both of these reasons are affected by the quality and availability of the data in each database. The results of this study are clear: GenBank has more sequences available for the 88 invasive insects identified by the Global Invasive Species Database than BOLD (Table 2). It is important to note that the data that are set to private in BOLD are still used when identifying unknowns (Ratnasingham and Hebert 2007) regardless, there are still fewer sequences available.

Identification success using GenBank (Blast) and BOLD (distance-based identification) has been analysed in detail via simulation and real data (Ross *et al.* 2008), which showed that no single method of species identification was superior across the range of scenarios studied. In other words, using BOLD to identify unknowns is not superior to GenBank's Blast for identification of unknowns. In fact, successful identification of plants (Lactuceae and Anthemideae) using only GenBank has been reported to be as high as 91% (Gemeinholzer *et al.* 2006) with the current quality of data. GenBank has also been used for

successful identifications of Phytoseiidae (Acari : Mesostigmata) (Tixier *et al.* 2011). The degree to which species are genetically differentiated appears to be the critical determinant of success. Correct identification is also dependent on representation in the reference dataset, but Ross *et al.* (2008) found that a correct identification was possible, even in the absence of the species in the reference data, using a strict tree-based approach to identification.

In order to conduct a tree-based approach, data must be available and proper phylogenetic analyses must be carried out. Most of the criticism of DNA barcoding falls on the methods used by DNA barcoders, particularly distance-based tree-reconstruction methods (DeSalle *et al.* 2005; Prendini 2005; Will *et al.* 2005), but they continue to dominate the methods in BOLD and the DNA barcoding literature, beyond identification (see fig. 4 in Taylor and Harris 2012). The ability to obtain sequences to conduct further studies (i.e. phylogenetic analyses) is a second reason that researchers query these databases. Does the researcher choose quantity (GenBank) or quality (BOLD)? The argument for using BOLD is that it contains quality data (Ratnasingham and Hebert 2007), but quality data (BOLD) is not useful when 37 of the 88 highly invasive species have no data or the data are private (Table 2). With BOLD missing 42% of the invasive insect species included in this study, we are left to rely on GenBank data, of variable quality. This raises the further issue of GenBank and the presumption of quality, or lack thereof. Early assessment of the quality of data in GenBank includes a review of human mtDNA in GenBank (Forster 2003) where the study concluded that up to half of the sequences contained mistakes (Harris 2003). Sources of error include amplification of numts (Song *et al.* 2008) and manual editing and sequencing errors (Harris 2003; Buhay 2009). A more recent review of 'COI-LIKE' sequences (the 'like' tag is given by GenBank staff when a stop codon is detected in the mtCOI region) shows that the tag 'like' is fairly common (22 of the 86 species included in the study) in Crustacean barcodes (Buhay 2009). There have been requests to GenBank to let users make changes themselves to the data, but GenBank's managers are yet to grant this access (Pennisi 2008). Recent improvements to GenBank include the 'barcode' tag on associated DNA barcoding data (to indicate quality), and also the UNVERIFIED tag before odd sequences flagged by GenBank staff (Benson *et al.* 2011, 2012). The user can also manage GenBank sequence quality. By way of example, De Barro and Ahmed (2011) downloaded 2325 *B. tabaci* mtCOI sequences from GenBank and then removed any sequences with gaps, stop codons, short sequences or those with ambiguous bases to obtain a final dataset of 1436 sequences. This demonstrates that one can manage quality (GenBank), but cannot do anything if the data are not available (BOLD or Q-bank).

As outlined above, both databases have issues with either data quality (GenBank) or data availability (BOLD). For the 88 invasive insect species included in this study, conducting an identification search or investigative search for DNA sequence data should include a GenBank search due to the lack of public sequences in BOLD. A GenBank search also provides more information on a particular species. For example, GenBank lists other gene regions that have been sequenced, author names, and author codes for each sequence, which are useful when going beyond identification. In BOLD ver. 2.5, most of the author



information is private – all that is recovered is the author's institution. BOLD ver. 3.0 is an improvement on ver. 2.5 as it includes author information for public and private sequence.

Another issue that needs consideration is the wealth of data already available in non-barcoding regions. Table 2 shows that several highly invasive insects have more sequence data for regions other than the barcoding region. For example, *B. tabaci* has the most sequences listed in GenBank (Table 1) for any of the 88 invasive insect species considered in this study. A key component of this dataset is mtCOI, but in this case it is the 3' end as this was the region chosen by some of the earliest studies to consider the genetic diversity in what has now become known as a cryptic species complex (Frohlich *et al.* 1999; Boykin *et al.* 2007, 2012; Dinsdale *et al.* 2010; De Barro *et al.* 2011). It would be beneficial for BOLD to link to the most recent papers on the identification and genetic research for a given species, even if these studies do not include the barcoding region of mtCOI.

Considering the diversity of campaigns listed in Table 1, there is no doubt that there is considerable interest in the application of the barcoding approach. Many of the campaigns include invasive species and some have an overt focus on biosecurity. So again, there is clear evidence that the philosophy behind barcoding encompasses a biosecurity focus. The key limiting issue is availability of data and the need for coordination to ensure that effort is appropriately targeted. QBOL is the campaign designed to handle invasive species through the development of Q-Bank. This initiative does not appear to contain data for the included invasive species, even though there is information in both BOLD and GenBank. If Q-Bank were to become useful to the biosecurity community, it would need to include all the data available in BOLD and also that from other gene regions, not just barcoding data (GenBank). The fracturing of the data from GenBank to BOLD to Q-Bank might lead to lack of enthusiasm (researchers having to submit data multiple times) and funding. A key issue that needs to be considered in the context of biosecurity is whether a stand-alone database is necessarily the best option. Q-Bank is a stand-alone database that delivers to QBOL, a project financed by the 7th Framework Program of the European Union. Once this project is completed, funding will cease and the question remains as to whether future funding can be secured to maintain the Q-Bank. If long-term funding is not secured it will fail to become useful. One argument for establishing Q-Bank as a stand-alone database is the particular need for security of data in the context of biosecurity regulation. However, there are other mechanisms of securing data without resorting to setting up a database *de novo*. Another example of the consequences of the lack of long-term funding on the usefulness and availability of data is the Influenza Sequence Database (Macken *et al.* 2001). This database began as an open access resource, but subsequent lack of funding forced the transition to fee for access. Conversely, there are several very successful sequence databases such as the HIV sequence database (Gaschen *et al.* 2001), the hepatitis C sequence database (Kuiken *et al.* 2005), and FlyBase (Ashburner and Drysdale 1994), that have maintained funding for over 20 years. One common theme across these is that they are not restricted to one gene region. Rather, they each encompass a broad range of regions and include biological data. Such an approach may improve the utility and relevance of databases such as Q-Bank and BOLD, which will help ensure their long-term viability. One can argue that this is

also behind the success of GenBank and is a challenge for BOLD to overcome.

In summary, BOLD is a relatively new database but for it to remain beneficial to the biosecurity community the following suggestions should be considered:

- All data in BOLD should be available to the public.
- When there is a positive identification of an unknown or query that falls within one of the campaign initiatives (Table 1), a link from BOLD to the campaign would be useful.
- Percentage thresholds (or BINs) should be avoided in identification of invasive insects as it is misleading and may lead to unintended biosecurity outcomes (Boykin *et al.* 2012).
- Creation of compartments for non-barcoding regions in BOLD would be useful for invasive species.
- Links to phylogenetic software from the BOLD website would be helpful for those wanting to go beyond identification and build more complex phylogenetic trees (Taylor and Harris 2012).
- Q-bank needs to incorporate all data from BOLD and GenBank for these invasive insects, regardless of gene region.

## Acknowledgements

The Tertiary Education Commission of New Zealand funded KA and LMB. LMB thanks Rupert Collins, The Bio-Protection Research Centre, Lincoln University, Lincoln, New Zealand for helpful discussions regarding BINs in BOLD.

## References

- Armstrong, K. F., and Ball, S. L. (2005). DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society of London B Biological Science* **360**, 1813–1823.
- Ashburner, M., and Drysdale, R. (1994). FlyBase – the *Drosophila* genetic database. *Development* **120**, 2077–2079.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2011). GenBank. *Nucleic Acids Research* **39**, D32–D37. doi:10.1093/nar/gkq1079
- Benson, D. A., Karsch-Mizrachi, I., Clark, K., Lipman, D. J., Ostell, J., and Sayers, E. W. (2012). GenBank. *Nucleic Acids Research* **40**, D48–D53. doi:10.1093/nar/gkr1202
- Bonants, P., Groenewald, E., Rasplus, J. Y., Maes, M., Vos, P. D., Frey, J., Boonham, N., Nicolaisen, M., Bertacini, A., Robert, V., Barker, I., Kox, L., Ravnikar, M., Tomankova, K., Caffier, D., Li, M., Armstrong, K., Freitas-Astúa, J., Stefani, E., Cubero, J., and Mostert, L. (2010). QBOL: a new EU project focusing on DNA barcoding of Quarantine organisms. *EPPO Bulletin* **40**. doi:10.1111/j.1365-2338.2009.02350.x
- Boykin, L. M., Shatters, R. G. Jr, Rosell, R. C., McKenzie, C. L., Bagnall, R. A., De Barro, P., and Frohlich, D. R. (2007). Global relationships of *Bemisia tabaci* (Hemiptera: Aleyrodidae) revealed using Bayesian analysis of mitochondrial COI DNA sequences. *Molecular Phylogenetics and Evolution* **44**, 1306–1319. doi:10.1016/j.ympev.2007.04.020
- Boykin, L. M., Armstrong, K. F., Kubatko, L., and De Barro, P. (2012). Species delimitation and global biosecurity. *Evolutionary Bioinformatics* **8**, 1–37.
- Buhay, J. E. (2009). “COI-LIKE” sequences are becoming problematic in molecular systematic and DNA barcoding studies. *Journal of Crustacean Biology* **29**, 96–110. doi:10.1651/08-3020.1
- Cameron, S., Rubinoff, D., and Will, K. (2006). Who will actually use DNA barcoding and what will it cost? *Systematic Biology* **55**, 844–847. doi:10.1080/10635150600960079

- Clarke, A. R., Armstrong, K. F., Carmichael, A. E., Milne, J. R., Raghu, S., Roderick, G. K., and Yeates, D. K. (2005). Invasive phytophagous pests arising through a recent tropical evolutionary radiation: the *Bactrocera dorsalis* complex of fruit flies. *Annual Review of Entomology* **50**, 293–319. doi:10.1146/annurev.ento.50.071803.130428
- De Barro, P., and Ahmed, M. (2011). Genetic networking of the *Bemisia tabaci* cryptic species complex reveals pattern of biological invasions. *PLoS Biology* **6**, e25579.
- De Barro, P. J., Liu, S. S., Boykin, L. M., and Dinsdale, A. B. (2011). *Bemisia tabaci*: a statement of species status. *Annual Review of Entomology* **56**, 1–19. doi:10.1146/annurev-ento-112408-085504
- DeSalle, R., Egan, M. G., and Siddall, M. (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **360**, 1905–1916.
- Dinsdale, A., Cook, L., Riginos, C., Buckley, Y. M., and De Barro, P. (2010). Refined global analysis of *Bemisia tabaci* (Gennadius) (Hemiptera: Sternorrhyncha: Aleyroidea) mitochondrial CO1 to identify species level genetic boundaries. *Annals of the Entomological Society of America* **103**, 196–208. doi:10.1603/AN09061
- Elias, M., Hill, R. I., Willmott, K. R., Dasmahapatra, K. K., Brower, A. V., Mallet, J., and Jiggins, C. D. (2007). Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proceedings. Biological Sciences* **274**, 2881–2889. doi:10.1098/rspb.2007.1035
- Floyd, R., Lima, J., deWaard, J. R., Humble, L. M., and Hanner, R. H. (2010). Common goals: policy implications of DNA barcoding as a protocol for identification of arthropod pests. *Biological Invasions* **12**, 2947–2954. doi:10.1007/s10530-010-9709-8
- Forster, P. (2003). To err is human. *Annals of Human Genetics* **67**, 2–4. doi:10.1046/j.1469-1809.2003.00002.x
- Frohlich, D. R., Torres-Jerez, I. I., Bedford, I. D., Markham, P. G., and Brown, J. K. (1999). A phylogeographical analysis of the *Bemisia tabaci* species complex based on mitochondrial DNA markers. *Molecular Ecology* **8**, 1683–1691. doi:10.1046/j.1365-294x.1999.00754.x
- Galtier, N., Nabholz, B., Glemin, S., and Hurst, G. D. (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology* **18**, 4541–4550. doi:10.1111/j.1365-294X.2009.04380.x
- Gaschen, B., Kuiken, C., Korber, B., and Foley, B. (2001). Retrieval and on-the-fly alignment of sequence fragments from the HIV database. *Bioinformatics* **17**, 415–418. doi:10.1093/bioinformatics/17.5.415
- Gemeinholzer, B., Oberprieler, C., and Bachmann, K. (2006). Using GenBank data for plant identification: possibilities and limitations using the ITS1 of Asteraceae species belonging to the tribes Lactuceae and Anthemidae. *Taxon* **55**, 173–187. doi:10.2307/25065539
- Harris, J. D. (2003). Can you bank on GenBank? *Trends in Ecology & Evolution* **18**, 317–319. doi:10.1016/S0169-5347(03)00150-2
- Hebert, P. D., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **270**, 313–321. doi:10.1098/rspb.2002.2218
- Kuiken, C., Yusim, K., Boykin, L., and Richardson, R. (2005). The Los Alamos hepatitis C sequence database. *Bioinformatics* **21**, 379–384. doi:10.1093/bioinformatics/bth485
- Liu, S., De Barro, P., Jing, X., Luan, J. B., Zang, L. S., and Ruan, Y. M. (2007). Asymmetric mating interactions drive widespread invasion and displacement in a whitefly. *Science* **318**, 1769–1772. doi:10.1126/science.1149887
- Liu, S. S., Colvin, J., and De Barro, P. (2012). Species concepts as applied to the whitefly *Bemisia tabaci* systematics: how many species are there? *Journal of Integrative Agriculture* **11**, 176–186. doi:10.1016/S2095-3119(12)60002-1
- Mack, R. N., Simberloff, D., Lonsdale, W. M., Evans, H., Clout, M., and Bazzaz, F. A. (2000). Issues in ecology. Biotic invasions: causes, epidemiology, global consequences, and control. *Ecological Applications* **10**, 689–710. doi:10.1890/1051-0761(2000)010[0689:BICEGC]2.0.CO;2
- Macken, C., Lu, H., Goodman, J., and Boykin, L. (2001). The value of a database in surveillance and vaccine selection. *International Congress Series* **1219**, 103–106. doi:10.1016/S0531-5131(01)00330-2
- Pennisi, E. (2008). DNA data. Proposal to ‘Wikify’ GenBank meets stiff resistance. *Science* **319**, 1598–1599. doi:10.1126/science.319.5870.1598
- Prendini, L. (2005). Identifying spiders through DNA barcodes. *Canadian Journal of Zoology* **83**, 498–504. doi:10.1139/z05-025
- Ratnasingham, S., and Hebert, P. D. (2007). BOLD: The Barcode of Life Data system (<http://www.barcodinglife.org>). *Molecular Ecology Notes* **7**, 355–364. doi:10.1111/j.1471-8286.2007.01678.x
- Ross, H. A., Murugan, S., and Li, W. L. (2008). Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology* **57**, 216–230. doi:10.1080/10635150802032990
- Rubinoff, D., Cameron, S., and Will, K. (2006). A genomic perspective on the shortcomings of mitochondrial DNA for “barcoding” identification. *The Journal of Heredity* **97**, 581–594. doi:10.1093/jhered/esl036
- Song, H., Buhay, J. E., Whiting, M. F., and Crandall, K. A. (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 of the United States of America* **105**, 13486–13491. doi:10.1073/pnas.0803076105
- Taylor, H. R., and Harris, W. E. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources* **12**, 377–388. doi:10.1111/j.1755-0998.2012.03119.x
- Tixier, M. S., Hernandez, F. A., Guichou, S., and Kreiter, S. (2011). The puzzle of DNA sequences of Phytoseiidae (Acari: Mesostigmata) in the public Genbank database. *Invertebrate Systematics* **25**, 389–406. doi:10.1071/IS11013
- Virgilio, M., Backeljau, T., Nevado, B., and De Meyer, M. (2010). Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics* **11**, 206. doi:10.1186/1471-2105-11-206
- Virgilio, M., Jordaens, K., Breman, F. C., Backeljau, T., and De Meyer, M. (2012). Identifying insects with incomplete DNA barcode libraries, african fruit flies (Diptera: Tephritidae) as a test case. *PLoS ONE* **7**, e31581. doi:10.1371/journal.pone.0031581
- Will, K. W., Mishler, B. D., and Wheeler, Q. D. (2005). The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology* **54**, 844–851. doi:10.1080/10635150500354878