Identification of Ornamental Fishes for Biosecurity

A thesis
submitted in partial fulfilment
of the requirements for the Degree of
Doctor of Philosophy
at
Lincoln University

by

Rupert A. Collins

Lincoln University
2012

Abstract

Introduction: Poorly regulated international trade in ornamental fishes poses risks to both biodiversity and economic activity via invasive alien species and exotic pathogens. Border security officials need robust tools to confirm identifications, often requiring hard-to-obtain taxonomic literature and expertise. DNA barcoding offers a potentially attractive tool for quarantine inspection, but has yet to be scrutinised for many fishes in the aquarium trade. This research examines a DNA barcoding approach for ornamental cyprinid fishes (Teleostei: Cypriniformes), an important group in terms of biosecurity risk.

Methodology and results: A reference library of fishes purchased from the international aquarium trade was assembled, and the specimens were identified to species using morphological characters derived from taxonomic literature. Many species were found to be misidentified in the trade. DNA barcodes were then generated using standardised protocols, and the efficacy of the reference library in making species level identifications was assessed. A total of 172 ornamental cyprinid fish species were sampled, providing baseline molecular data for 91 species currently unrepresented in public reference libraries. DNA barcodes were found to be highly congruent with the morphological assignments, with identification success rates of up to 99%. The cyprinid fish dataset was augmented with sequences from GenBank for an additional 157 species, the benefit of which was additionally evaluated. Here, it was observed that the inclusion of GenBank data resulted in a more comprehensive library, but at a cost to success rate due to the increased number of singleton species.

Identification success rates are known to be sensitive to the choice of identification criterion, and because this may be important for biosecurity applications, a specific focus of this research was to assess these procedures. Here, a variety of different techniques were applied (neighbour-joining monophyly, bootstrap, nearest neighbour, GMYC, percent threshold), and it was found that identification success rates varied between 87% and 99%, according to the method used. The appropriateness of the commonly employed Kimura two-parameter (K2P) model was also examined using an information-theoretic model-selection approach. Despite its ubiquity in the DNA

barcooding literature, the K2P model was not found to be well supported as an appropriate substitution model at the species level. However, using this model did not affect identification success rates overall.

Standard DNA barcoding techniques are known to be inappropriate and potentially misleading in situations where interspecific hybridisation has occurred. Similarly, where cryptic species are suspected, mitochondrial DNA is sometimes insufficient to robustly recognise lineages. As both of these situations are believed to occur in the ornamental fish trade, and using a genomic dataset, a range of candidate nuclear loci were assessed as a complementary marker to COI. The rhodopsin gene was shown to be variable between closely related species, and with 200 sequences from cyprinid fishes, interspecific hybridisation events were confirmed, and unrecognised diversity was highlighted within popular aquarium species.

Traces of degraded environmental DNA present in water can now be used to detect the presence of aquatic species, so diagnostic tests for fish identification were investigated with the aim of developing a new, more efficient biosecurity quarantine tool. The COI barcode library was mined for informative short-length markers using a sliding window analysis of variation through the gene. Species-specific DNA sequences were successfully amplified from aquarium water samples, and at relatively low densities of the target species.

Conclusions: This study demonstrates that DNA barcoding can provide a highly effective biosecurity tool for rapidly identifying ornamental fishes. In the small number of cases where DNA barcodes are unable to offer a species level identification, previous studies are improved upon by consolidating supplementary information from multiple data sources in the form of specimen images, morphological characters, taxonomic bibliography, and preserved voucher material. Reference libraries can be utilised to develop new diagnostic approaches using environmental DNA, allowing quarantine facilities to capitalise on non-invasive techniques for detecting high-risk fishes. The biggest obstacles, however, to an operational implementation of DNA barcoding and any future expansions of the reference libraries, are the combined problems of misidentification of reference specimens between labs, and a lack of access to appropriate taxonomic literature to first identify the fishes. If these problems are not addressed by the barcoding and taxonomic communities respectively, this will ultimately compromise the ability of biosecurity agencies to use a DNA barcoding tool.

Acknowledgements & Preface

Lincoln University: I thank first of all, Karen Armstrong and Rob Cruickshank for their continued support and contributions throughout the duration of the work. I also thank: Andrew Holyoake for his efficient lab management and technical expertise; Samuel Brown for his endless patience and help in all things R related (as well as encouraging me to use R, LaTeX, and Linux); Laura Boykin for many helpful comments and suggestions on manuscript drafts; James Ross for statistical advice; Norma Merrick for the smooth operation of the DNA sequencing facility; Jagoba Malumbres-Olarte, Stephane Boyer, Emily Fountain and the rest of the Molecular Ecology Lab Group for many thoughtful meetings and discussions; and Elizabeth Wandrag & Kirsty McGregor for their excellent proofreading skills.

MAF Biosecurity New Zealand: I thank my PhD advisors Suzanne Keeling & Colin Johnston, especially for arranging with MAF Biosecurity New Zealand the generous extension to my scholarship after the difficult circumstances of the Canterbury earthquakes of 2010/2011.

National University of Singapore and the Raffles Museum of Biodiversity: Here, I thank: Rudolf Meier & Youguang Yi for advice and suggestions on manuscripts, as well as contributing data (137 of the COI sequences used in Chapter 2 were provided by Youguang); and Kelvin Lim, Heok Hui Tan & Heok Hee Ng for logistical support and a warm welcome during my visit.

Natural History Museum, London: Thanks are due to: James Maclaine & Oliver Crimmen, for being jolly helpful and accommodating during my visits; Lisa Di Tommaso for taking care of all my literature requirements; and Patrick Cambell for providing tissue samples of a hybrid *Clarias* catfish.

Other acknowledgements: Acknowledgements are also to due to: the thesis examiners Robert Ward (CSIRO) and Ian Hogg (University of Waikato); the anonymous reviewers and editors who offered improvements to the manuscripts, and ultimately

the thesis; Richard Broadbent (Warwick, UK) and Neil Woodward (Pier Aquatics, Wigan, UK) for help with sourcing some of the fishes; Olivier David (INRA, France) for kindly providing a *k*-NN script for R; Samuel Smits (San José State University) for helping improve the online phenograms; Jon Banks (University of Waikato) for advice on the eDNA work; Matt Ford (seriouslyfish.com) for proving essential literature, and helping to promote my paper on his Web site; Peter Cottle (danios.info) for sharing his enthusiasm and expertise with *Danio*; Jeremy Wright (University of Michigan) for assistance with *Synodontis*; Bob McDowall (1939–2011) at NIWA for taking the time to meet up and share his thoughts on the project; and lastly Bill Eschmeyer and the *Catalog of Fishes* team for making my life considerably easier when searching for taxonomic fish literature. The final and possibly the most important thanks are due to Philip and Katherine Collins for all their moral and financial support over the years.

Funding: This work was funded by a Ministry of Agriculture and Forestry Biosecurity New Zealand (MAFBNZ) scholarship, and was completed at the Bio-Protection Research Centre, Lincoln University, New Zealand between November 2008 and March 2012.

Publications: From this thesis, three articles have been published in academic journals (Brown *et al.*, 2012; Collins *et al.*, 2012a,b). Two further articles have been provisionally accepted for publication, pending revisions (as of 02/09/12).

List of acronyms and symbols

Δ	Delta
Γ	Gamma
µL	Microlitre
µM	Micromolar
E	Evidence ratio
P	Probability
g	Gravity
p	Proportion
w	Akaike weight
aff.	Affinis (Latin)
AIC	Akaike Information Criterion
BCM	Best Close Match
BI	Bayesian Inference
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
BOLD	Barcode of Life Data Systems
bp	Base pair
CART	Classification and Regression Trees
cf.	Confer (Latin)
COI	Mitochondrial cytochrome *c* oxidase subunit I
Cyt *b*	Mitochondrial cytochrome *b*
DNA	Deoxyribose Nucleic Acid
DOI	Digital Object Identifier
eDNA	Environmental DNA
ERMA	Environmental Risk Management Authority
ESU	Evolutionary Significant Unit
FNZAS	Federation of New Zealand Aquatic Societies
g	Gram
GMYC	General Mixed Yule Coalescent
H_0	Null hypothesis

HSNO	Hazardous Substances and New Organisms act
HTML	HyperText Markup Language
IHS	Import Health Standard
Indel	Insertion-deletion event
IRBP	Interphotoreceptor Retinoid-binding gene
***k*-NN**	*k*-Nearest Neighbour
K2P	Kimura 2-parameter
LSU	Large Subunit 28S rDNA
M	Molar
MAFBNZ	Ministry of Agriculture and Forestry Biosecurity New Zealand
MCMC	Markov chain Monte Carlo
min	Minute
ML	Maximum Likelihood
MLL	Mixed-lineage Leukemia-like gene
mM	Millimolar
MP	Maximum Parsimony
mtDNA	Mitochondrial DNA
nDNA	Nuclear DNA
NGS	Next Generation Sequencing
NJ	Neighbour joining
NN	Nearest Neighbour
NUMT	Nuclear-mitochondrial Pseudogene
PCR	Polymerase Chain Reaction
QBOL	Quarantine Barcode of Life
RAG1	Recombination Activating Gene 1
rDNA	Ribosomal DNA
RHO	Rhodopsin gene
s	Second
sp.	Species (singular)
spp.	Species (plural)
SVG	Scalable Vector Graphics
Tm	Oligonucleotide melting temperature
UPGMA	Unweighted Pair Group Method with Arithmetic means
URL	Uniform Resource Locator

Contents

List of Figures

List of Tables

Chapter 1

An introduction to DNA barcoding for biosecurity

1.1 Biosecurity in New Zealand

In contrast to many countries, New Zealand has a stringent legal framework for biosecurity, with specific legislation and Acts of Parliament (Meyerson & Reaser, 2002). In 1993, the Biosecurity Act was introduced—legislation administrated by the Ministry of Agriculture and Forestry Biosecurity New Zealand (MAFBNZ)—to provide a "legal basis for excluding, eradicating and effectively managing pests and unwanted organisms" (Ministry of Agriculture and Forestry, 2011). Unwanted organisms are defined as those "capable or potentially capable of causing unwanted harm to any natural and physical resources or human health" (Ministry of Agriculture and Forestry, 2011). The protection of biodiversity, livelihoods, health, and culture, is central to the remit of MAFBNZ. In the context of ornamental fish importation to New Zealand and worldwide, the primary risks regard the introduction of invasive alien species and exotic aquatic pathogens.

1.1.1 Invasive alien species

In 2005, the Millennium Ecosystem Assessment (2005) highlighted the economic and social benefits of biodiversity and associated ecosystem services. Along with climate change, habitat destruction, pollution and over-exploitation, the impacts of alien invasive species are frequently cited as a major cause of the anthropogenic biodiversity crisis (Chapin III *et al.*, 2000; Pimentel *et al.*, 2005; Vitousek *et al.*, 1997). This human interference has seen the biotic homogenisation of aquatic communities, with assemblages of cosmopolitan species replacing more complex, unique communities of native, often endemic fishes (Dudgeon *et al.*, 2006; Rahel, 2002, 2007). By breaching natural barriers, global trade has transported freshwater species beyond both administrative boundaries and their biogeographical confines

(Hulme, 2009). Overall, economic losses associated with invasive alien species are significant, and have been estimated at up to US$120 billion per year in the USA (Pimentel *et al.*, 2000, 2005). Several pathways for the global introduction of non-native freshwater fish species have been identified, and include but are not limited to: (1) deliberate legal/illegal introduction for recreational angling; (2) escaped or released bait fish for recreational angling (Rahel, 2007); (3) contaminant species in fish stocking events (Rahel, 2007); (4) escapes from aquaculture facilities and retailers (Naylor *et al.*, 2001; Rixon *et al.*, 2005); (5) creation of canals and waterways linking drainages (Rahel, 2007); (6) discharge of ballast water from shipping (Ricciardi & MacIsaac, 2000); (7) deliberate release for cultural/religious reasons (Lintermans, 2004); and (8) the release of ornamental species by aquarists (McDowall, 2004; Padilla & Williams, 2004; Rixon *et al.*, 2005).

A total of 233 aquatic species are known to have been introduced outside their native range worldwide by 1988, but 49% of the introduction events comprised eighteen common species (Rahel, 2007). The ornamental industry is implicated as the primary transport vector in 37 of the 59 fish introductions in the United States (Rahel, 2007), while more generally across North America, approximately 100 species have been introduced via the aquarium trade, with 40 species having become established (Rixon *et al.*, 2005). In Singapore—a global aquarium fish trading hub—at least 14 invasive ornamental fish species were reported to be resident in 1993 (Ng *et al.*, 1993). In Florida—the centre of the U.S. ornamental aquaculture industry—greater than 75% of freshwater fish introductions have been associated with releases from private aquariums (Padilla & Williams, 2004). A similar figure is reported in Australia, at 65% of 34 species (Lintermans, 2004). Although New Zealand's narrow climatic/habitat range, and isolated drainage basins make it less vulnerable to fish invasions, it does not diminish the potential harm from the invasion of a more limited selection (McDowall & James, 2005). Geothermal waters in New Zealand have been colonised by three species of "tropical" ornamental fishes: *Poecilia latipinna*, *P. reticulata*, and *Xiphophorus helleri* (McDowall, 2004). These fishes have so far not spread from geothermal sites. However, their impacts although localised, are unknown (McDowall, 2004; McDowall & James, 2005).

1.1.2 Exotic aquatic pathogens

The risks presented by the ornamental industry are not, however, limited to traded invasive fishes. Associated pathogenic organisms such as protozoa, bacteria and

viruses are equally undesirable (Smith *et al.*, 2012), with these exotic pathogens known to cause harm to native species (Gozlan *et al.*, 2005), industrial food aquaculture (Go & Whittington, 2006; McDowall, 2004; Whittington & Chong, 2007), and also the ornamental fish trade itself (Ploeg *et al.*, 2009). The impacts of exotic fish diseases have the potential to interfere with New Zealand's tourism market (e.g. to close trout fisheries), as well as decrease the production capacity of export industries such as fish farming (Murray & Peeler, 2005). New Zealand's biosecurity strategy aims to minimise this risk and prevent the transfer of exotic aquatic pathogens to: (1) populations of native fishes and amphibians; (2) populations of non-native but economically important fishes (e.g. salmonids for recreational angling); (3) aquaculture facilities; and (4) ornamental fishes already present in New Zealand.

The ornamental fish industry is recognised as a significant disease pathway (Hine & Diggles, 2005; Whittington & Chong, 2007), with for example in Sri Lanka, 23 of 26 ornamental fish farms being infected with one or more parasites (Thilakaratne *et al.*, 2003). Streptococcal infections of aquarium danios (*Danio* spp.) imported into Canada were shown to be transmittable to the rainbow trout *Oncorhynchus mykiss*, an important food fish (Ferguson *et al.*, 1994). In Australia, an outbreak of *Megalocytivirus* (*Iridoviridae*) at a *Maccullochella peelii* (Murray cod) aquaculture facility was likely to have been passed across the species barrier by imported ornamental *Colisa lalia* (dwarf gourami) from Asia (Go *et al.*, 2006; Go & Whittington, 2006). New and harmful pathogens are also often associated with invasive species. For example, the introduction of *Pseudorasbora parva* (topmouth gudgeon) into the River Danube has led to local extirpation of *Leucaspius delineatus* (sunbleak) due to a rosette-like intracellular eukaryotic parasite, leading to conservation concerns (Gozlan *et al.*, 2005). The pathogenic organisms of interest to New Zealand biosecurity are listed by Hine & Diggles (2005), and include a broad range of groups including viruses, bacteria, fungi, protozoans, myxozoans, monogeneans and crustaceans. Fishes are often mixed at breeding and wholesale export facilities before they are shipped abroad, and it is difficult to predict which pathogens they may have been in contact with. Pathogens can also be host-taxon specific, and possibly require special quarantine measures for some species or groups (MAF Biosecurity New Zealand, 2011; Ploeg *et al.*, 2009; Whittington & Chong, 2007). Compounding this, some pathogens can be vectored by carrier hosts with no clinical signs of disease (Gozlan *et al.*, 2005; Ploeg *et al.*, 2009; Whittington & Chong, 2007).

1.1.3 International trade and the ornamental fish industry

The ornamental aquatic industry is among the world's largest transporters of live animals and plants[1], with an annual trade volume estimated at US$15–25 billion (Padilla & Williams, 2004; Ploeg *et al.*, 2009). Aquarium fishes are both wild caught, and captive bred at aquaculture facilities, with over one billion fishes traded through more than 100 countries in 2000 (Whittington & Chong, 2007). In the case of freshwater fishes, $\geq$ 90% of the trade volume is in a relatively small number of popular species sourced from commercial farms (Gerstner *et al.*, 2006), while more diverse wild caught exports contribute the remainder. A complex supply chain exists for these ornamental fishes, and before they arrive at a retailer they may have passed though a series of regional and international distribution centres where consignments can be consolidated, reconsolidated and subdivided (Ploeg *et al.*, 2009). This potentially increases the number of access points for undesirable organisms to enter each shipment (Ploeg *et al.*, 2009), as well as opportunities for mislabelling. Figure 1.1 shows such a centre in Singapore.

While statistics are available on total volumes sold, little quantitative data exist on the number and composition of species involved in the aquarium trade, but it has been estimated that over 5,000 species have been available at some point (Hensen *et al.*, 2010; McDowall, 2004). The industry in wild aquatic ornamentals for the aquarium hobby is a dynamic business, with new and undescribed species frequently appearing from new areas. As an example, some, such as *Puntius denisonii* (redline torpedo barb) have quickly moved from obscurity to becoming a major Indian export and a conservation concern within relatively few years (Ali *et al.*, 2010; Raghavan *et al.*, 2007).

1.1.4 Biosecurity management of ornamental species

Biosecurity challenges exist in effectively monitoring and managing the complex pathways involved in international trade (Hulme, 2009; Rubinoff *et al.*, 2011; Wong *et al.*, 2010), with a key issue for risk assessment being the identification of traded biological materials to species (Armstrong & Ball, 2005; Darling & Blum, 2007; deWaard *et al.*, 2010). Effective cataloguing of both known problematic species, and potential propagules (all traded species), can inform risk analyses and facilitate pre- or post-border control measures (i.e. import restrictions and quarantine).

[1] Of additional concern are the introductions of incidental fauna such as invertebrate plankton associated with aquarium fish imports and the aquarium hobby (Duggan, 2010).

Figure 1.1. An export facility in Singapore showing rows of hundreds of stock tanks and fishes bagged and ready for dispatch. © Rupert A. Collins, 2012.

Currently in New Zealand, when fishes are inspected by customs officials they are identified visually using morphological features, but there are multiple difficulties associated with this method: (1) literature and keys pertaining to the taxa in question may be unobtainable or inadequate for diagnosis; (2) identifications can be non-standardised and liable to subjectivity between examiners; (3) undescribed species are commonly traded, with little literature published to discern them from currently described species; (4) aquarium guide books are frequently inaccurate for many groups; (5) consultation with appropriate taxonomic expertise can be impossible or time consuming; and (6) specimens may lack important differentiating characters due to factors such as stress during shipment, age, sexual dimorphism or selective breeding. Reviews have identified that fish identification should be a key priority in risk assessment and monitoring procedures in New Zealand (Hine & Diggles, 2005; McDowall, 2004).

Approaches to addressing biosecurity threats from ornamental fishes are varied; the United States and United Kingdom adopt a "blacklist" system, whereby a small group of known high risk species are subject to controls (Copp *et al.*, 2010; Ploeg, 2008). For countries such as Australia and New Zealand who view this industry as a greater biosecurity threat, only fishes included on a "whitelist" of manageable species are permitted, and all others are by default disallowed (MAF Biosecurity New

Zealand, 2011; McDowall, 2004; Ploeg, 2008; Whittington & Chong, 2007). Under Section 22 of the Biosecurity Act 1993 (Ministry of Agriculture and Forestry, 2011), the current allowable imports list comprise 1,451 (1,010 freshwater and 441 marine) fish species on the Import Health Standard (MAF Biosecurity New Zealand, 2011, accessed December 2011). For the enforcement of these restrictions, an effective biosecurity procedure requires fast and accurate early detection of potentially harmful fishes at the pre-retail quarantine stage. Biological attributes such as disease vectoring potential and invasiveness are associated with the nomenclature of the species, and it is therefore important that names be both accurate and harmonised throughout the process of risk management, import, and quarantine.

1.2 Molecular diagnostics for biosecurity

Molecular diagnostic technologies are becoming an increasingly important part of biosecurity procedures, especially with regard to economically important agricultural insect pests (Armstrong & Ball, 2005; deWaard *et al.*, 2010). These molecular methods circumvent some of the problems with identifying specimens morphologically in situations when discriminating characters are absent (e.g. immature life stages). Most methods rely on species-specific DNA-sequence variation detected by PCR amplification (e.g. RFLP, RAPD, Multiplex-PCR, SSCP, AFLP), and have been reviewed by Darling & Blum (2007), Ali *et al.* (2004), Armstrong & Ball (2005), Teletchea (2009), Le Roux & Wieczorek (2009), and Rasmussen & Morrissey (2008). The restriction fragment length polymorphism (RFLP) method has been the most widely used for identifying commercial food fishes (Rasmussen & Morrissey, 2008). This method, which relies on presence/absence of diagnostic restriction sites, allows confirmation of specimen identity due to length variation in cleaved fragments. The primary weakness identified with this, and other previously used methods, is the group specificity of the procedures (e.g. primer design and PCR conditions), the requirement of *a priori* knowledge of the sequence variation, and therefore the limited size of the species pool for which identifications can be made. Because infrastructure may not be in place for directly comparing data shared between laboratories, this reduces the anticipatory aspect in adapting to changing biosecurity threats and priorities (Armstrong & Ball, 2005). When data are not able to be effectively shared, identification of an unanticipated pest would be potentially time consuming, as new

experimental procedures using restriction enzymes or multiplex PCR reactions would need to be developed.

1.2.1 DNA barcoding as an identification tool

1.2.1.1 Standardisation and scalability

DNA sequence data contain a higher resolution of information (i.e. discrete nucleotide polymorphisms) when compared to methods such as RFLP fragment length variation. The development of the DNA barcoding method (*sensu* Hebert *et al.*, 2003a) has facilitated a standardised technique using sequence data, overcoming some of the problems identified with previous methods. For animals, DNA barcoding uses sequence data from a short ∼650 bp fragment from the 5′ region of the protein-coding mitochondrial cytochrome *c* oxidase I gene (COI). The key benefit of a DNA barcoding approach is its standardisation: universal, conserved primers are able to amplify a positionally homologous gene region across diverse realms of life, and further standardisation is achieved through shared lab protocols and data management systems. With each new sequence, the reference database can then be improved and refined in terms of both intra- and interspecific variation (Armstrong & Ball, 2005; deWaard *et al.*, 2010). The Barcode of Life Data System BOLD (Ratnasingham & Hebert, 2007), represents the centralised, international workbench/portal for barcode data, and can be used in conjunction with the GenBank repository (Federhen, 2011). Such are the benefits in scalability that systems like BOLD offer, automated pipelines can also now be implemented for vast biodiversity assessment projects, or bulk routine identifications (Borisenko *et al.*, 2009).

1.2.1.2 Mitochondrial DNA as a molecular marker

The use of a mitochondrial gene is important, as mitochondrial DNA molecules are vastly more abundant in the cell (∼1,000×), when compared to the nuclear DNA (Avise, 2009; Teletchea, 2009). This improves PCR success in the laboratory, and offers greater chance of recovery from poorly preserved or degraded samples (Linacre & Tobe, 2011; Teletchea, 2009). Due to a lack of DNA repair enzymes (Brown *et al.*, 1979; Joseph & Omland, 2009), and/or possible environmental selection (Lane, 2009), mitochondrial genes have high nucleotide substitution rates. In salamanders and beetles, COI has been shown to have one of the fastest mutation rates for a mitochondrial gene, especially at the third position (Mueller, 2006; Pons *et al.*, 2010).

For diploid organisms, mitochondrial loci also reach coalescence generally four times faster than nuclear genes, due to their smaller effective population size (Joseph & Omland, 2009; Zink & Barrowclough, 2008). Protein-coding mitochondrial genes typically lack introns, greatly reducing alignment ambiguity when compared to 12S or 16S rDNA, for example (Hebert *et al.*, 2003a). The largely maternal inheritance of mitochondrial genes and lack of recombination and heterozygosity, further simplifies analytical procedures. Despite these benefits, several complications can arise when making inferences with mtDNA (see Section 1.3).

Historically, sequence data from gene regions other than COI have also been utilised as DNA barcode markers *sensu lato*, the most significant in species-level fish research being mitochondrial cytochrome *b* (Johns & Avise, 1998; Page & Hughes, 2010; Sevilla *et al.*, 2007; Teletchea, 2009). Consequently, there are a large number of sequences for this gene available on GenBank for fishes (Johns & Avise, 1998; Page & Hughes, 2010; Teletchea, 2009). Some studies have shown that cytochrome *b* may be more discriminating, and perform better than COI for specimen identification in some mammal species (Tobe *et al.*, 2010). However, COI was not chosen as the *de facto* animal barcode for an *a priori* assumption of its superior variability over any of the other 12 mitochondrial protein-coding genes; it was chosen due to its highly constrained amino acid sequence, and therefore the reliability of available primer sets to amplify across much of the Metazoa (Hebert *et al.*, 2003a). Importantly, and in contrast to the barcode application of COI, many of the cyt *b* data in GenBank frequently lack the associated voucher specimens essential for a reference library, and are not from consistent regions of the ∼1,140 bp gene (Broughton *et al.*, 2001; Dawnay *et al.*, 2007; Ward *et al.*, 2009). Now, and primarily as a result of the FISH-BOL initiative to DNA barcode all fish species, COI has recently overtaken cyt *b* in terms of number of sequences on GenBank (Becker *et al.*, 2011; Ward *et al.*, 2009). For many taxa, COI barcodes have shown adequate resolution of even closely related species, and especially so for many fishes (Ward, 2009; Ward & Holmes, 2007).

DNA barcoding has now been demonstrated as an effective fish identification tool in food-product consumer protection (Cohen *et al.*, 2009; Lowenstein *et al.*, 2009, 2010), with the U.S. Food and Drug Administration (FDA) recently validating DNA barcoding as an identification tool for marketplace seafood (Becker *et al.*, 2011; Stoeckle, 2012; Yancy *et al.*, 2008). A critical benefit of DNA barcoding in this scenario is the possibility to successfully retrieve and amplify full or partial barcodes from cooked, processed, or otherwise degraded samples (Becker *et al.*, 2011; Huxley-Jones *et al.*, 2012; Teletchea, 2009). Other applications for fisheries management

and conservation have also been demonstrated (Holmes *et al.*, 2009; Ogden, 2008; Steinke *et al.*, 2009b; Wong *et al.*, 2009), while the study of Steinke *et al.* (2009b) applied the technique to identify fishes in the marine ornamental trade .

1.2.2 DNA barcoding as a biosecurity tool

Armstrong & Ball (2005) were the first to apply a DNA barcoding approach to a biosecurity question; they found potentially invasive organisms—including morphologically indistinct immature life stages such as insect eggs or larvae—could be reliably identified to species level, an invaluable benefit to biosecurity. Even some of the strongest critics of DNA barcoding have supported its application in these kind of situations (e.g. Cameron *et al.*, 2006; Rubinoff *et al.*, 2006). Now, DNA barcoding is demonstrated to be an essential part of the toolkit for the management of invasive species (Darling & Blum, 2007). As part of this, the QBOL (Quarantine Barcode of Life) initiative aims to set up a "sustainable diagnostic resource to enable 'DNA-barcode identification' ultimately for all quarantine plant pests or pathogens of statutory importance" through targeted acquisition of pest species and collaboration in data sharing (Bonants *et al.*, 2010).

Classic barcoding for biosecurity may involve identifying to species the hitchhikers on an imported agricultural product, for example, and thereby informing an appropriate biosecurity response based on the pest status of the organism concerned (Armstrong & Ball, 2005; deWaard *et al.*, 2010). For ornamental fish quarantine, it is usually the status of the traded species themselves that is of concern. An extension of this is the use of DNA barcoding for wildlife forensics, where controlled and often endangered species are traded (Alacs *et al.*, 2010; Dawnay *et al.*, 2007; Linacre & Tobe, 2011; Ogden, 2008; Reid *et al.*, 2011). Legal cases involving trade in illicit animals or derivatives thereof, are similar to that of biosecurity, with stakes and responsibilities being considerable, i.e. incorrect prosecutions or valuable shipments unnecessarily destroyed. Validation of the method is therefore important for the admissibility of a DNA test in court (Dawnay *et al.*, 2007). The process of validation is to ensure "that a laboratory procedure is robust, reliable, and reproducible" (Alacs *et al.*, 2010). Dawnay *et al.* (2007) provided a validation study of laboratory procedures in generating DNA barcode identifications, and examined "reproducibility, heteroplasmy, mixed DNA, DNA template concentration, chemical treatments, substrate variation, environmental conditions and thermocycling parameters"; they reported their protocols as generally robust to these factors.

1.3 Problems with DNA barcoding

Several challenges to the use of DNA barcodes have been identified since the inception of the method, and important caveats and assumptions need to be made when using these data—and sometimes when using mitochondrial DNA data in general (Funk & Omland, 2003; Galtier *et al.*, 2009; Rubinoff, 2006). Some of these problems that need to be considered with regard to their impact on identification success are outlined below.

1.3.1 NUMTs and heteroplasmy

Mitochondrial genes can be duplicated into parts of the nuclear genome, becoming paralogous copies—NUMTs (nuclear-mitochondrial pseudogenes)—of their cytoplasmic equivalent (Buhay, 2009; Song *et al.*, 2008).Typically, they are relaxed from the strong selection of the functional mitochondrial protein, and are altered substantially by random mutational events, giving rise to length variation, indels, and the presence of in-frame stop codons (Buhay, 2009; Song *et al.*, 2008). Therefore, if NUMTs are confused with authentic mtDNA sequences in reference datasets, identification success may decrease. While a potentially significant pitfall when studying insects or crustaceans, NUMTs have not been identified as a critical issue in fish barcoding (Ward *et al.*, 2009), provided vigilance and quality control of sequences is maintained (Song *et al.*, 2008). However, so-called "cryptic NUMTs" have recently been identified in a beetle species, differing from their orthologues by only 1–3 non-synonymous changes (Bertheau *et al.*, 2011). It is not clear how widespread these are and if they will become a problem, but providing authentic mtDNA is co-amplified, their presence can be identified by double peaks in the sequence chromatograms (Bertheau *et al.*, 2011).

Intra-individual polymorphism in mitochondrial DNA from heteroplasmic tissues can cause ambiguity and bias in estimates of molecular diversity (Magnacca & Brown, 2009; Rubinoff *et al.*, 2006). While this phenomenon has been reported in fishes (Hoarau *et al.*, 2002), it has not been flagged by reviews of fish mtDNA studies as being a significant occurrence (Becker *et al.*, 2011; Teletchea, 2009; Ward, 2009; Ward *et al.*, 2009).

1.3.2 Non-neutrality

Mitochondrial genes involved in metabolic processes such as respiration (e.g. COI), are assumed to be nearly neutrally evolving, i.e. the protein sequence remains static while synonymous substitutions accumulate at third and first codon positions (Galtier *et al.*, 2009). However, widespread selective sweeps and instances of non-neutrality have been documented (Bazin *et al.*, 2006; Wares, 2009). Through the reduction of intraspecific variation, these phenomena may generally be of benefit to specimen identification using DNA barcodes, but conversely, positive selection from maternally-inherited intracellular endosymbionts such as *Wolbachia*, is believed to cause both inflated intraspecific divergences and haplotype sharing between species (Hurst & Jiggins, 2005). Although endosymbionts have been reported in vertebrates (Werren & Baldo, 2008), the problem again appears more significant for invertebrates (Galtier *et al.*, 2009; Hurst & Jiggins, 2005).

1.3.3 Rate variation

Mitochondrial evolution does not always occur in a consistent or clock-like manner; some lineages may display significantly faster rates than others (Drummond & Suchard, 2010; Galtier *et al.*, 2009; Hendrich *et al.*, 2010; Rutschmann, 2006). This lack of a constant mutation rate calls into question whether a universal divergence threshold (e.g. Hebert *et al.*, 2003a) can be used to delimit species or even identify specimens (Cognato, 2006; Rubinoff *et al.*, 2006; Vogler & Monaghan, 2007). Furthermore, speciation is independent of mitochondrial sequence divergence (but see Lane, 2009; Shiyang *et al.*, 2012), and perhaps more importantly there may not be an *a priori* reason to assume all taxa in a group diverged from one another at an equivalent time, i.e. the depth of the coalescent may vary considerably between species (Monaghan *et al.*, 2009).

1.3.4 Non-monophyly

The non-monophyly of mitochondrial DNA trees has been well documented (Funk & Omland, 2003; Joseph & Omland, 2009; McKay & Zink, 2010). Patterns of phylogenetic relationships have therefore been uncovered for some taxa in which putative organismal phylogeny is not reflected in the mtDNA genes sampled. Definition and illustration of following terminology is shown in Figure 1.2. In terms of DNA barcoding, most interpretations of the method require a “barcoding gap”, which

is essentially the same representation of monophyly minus the phylogenetic tree, where all members of each species must be more similar to each other than to a different species (Meyer & Paulay, 2005). When using monophyly as an identification criterion, as is commonly conducted (Meier, 2008; Ross *et al.*, 2008), incorrect or ambiguous identifications can occur when querying para- or polyphyletic species (Meier, 2008). The oft-cited article by Funk & Omland (2003), reported a 23.1% rate of para-/polyphyly across a variety of animals in 584 studies of mtDNA. Reasons for this discord are also presented by Joseph & Omland (2009), as well as Funk & Omland (2003), and are broken down in the following sections.

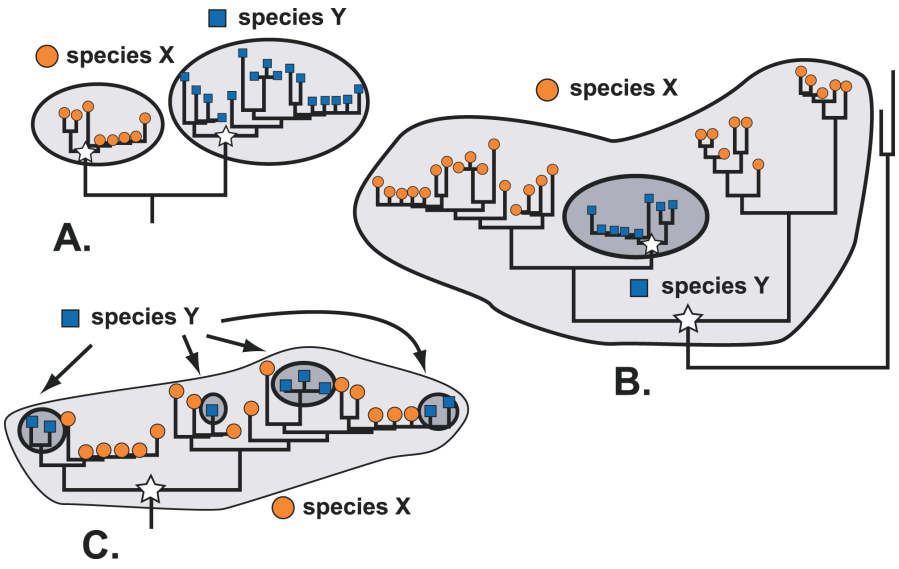

Figure 1.2. Three examples of non-monophyletic relationships: Figure (A) shows monophyly of species X and species Y; Figure (B) shows a paraphyletic species X with regard to a monophyletic species Y; and Figure (C) shows polyphyly of both species X and Y. Coalescent points are shown with white star. Figure copyright ⓒⓘ (Meyer & Paulay, 2005).

1.3.4.1 Inadequate phylogenetic signal

If too small a fragment of DNA is used for phylogenetic inference, insufficient information (i.e. synapomorphies) may be present to resolve groups, and the result may also be confounded by homoplasy (Funk & Omland, 2003). Paternal inheritance issues aside (see Zhao *et al.*, 2004), all mitochondrial genes have the same matrilineal history (Avise, 2009), but patterns in single genes or gene fragments can be obscured due to stochastic processes, saturation of substitution, or idiosyncratic rates of mutation (Hendrich *et al.*, 2010; Mueller, 2006). This is a potential problem for

recently diverged groups, but in some situations, sampling further mitochondrial genes may improve the likelihood of recovering reciprocal monophyly (Elias *et al.*, 2007).

1.3.4.2 Incomplete lineage sorting

Patterns similar to those caused by inadequate phylogenetic signal can be observed in mtDNA trees due to incomplete lineage sorting. Under coalescent theory, the time for reproductively isolated lineages to become reciprocally monophyletic (i.e. fixation of exclusive haplotypes), is dependent on the effective population size (Avise, 2009; Funk & Omland, 2003). Thus, recently divergent sister species, or sister species with exceptionally large population sizes may retain some ancestral polymorphisms causing para-/polyphyly. Contrary to patterns caused by inadequate phylogenetic signal, sampling further mtDNA will not resolve monophyletic groups. McKay & Zink (2010) estimate 15.6% of the non-monophyletic patterns they examined from bird studies were caused by incomplete lineage sorting; an additional 21.3% of cases could not be distinguished between hybridisation and incomplete lineage sorting.

1.3.4.3 Introgression

Due to the maternal inheritance of mtDNA, interspecific hybridisation events can obscure true genealogical histories, and may not be detected at all depending on the direction of the introgression (Avise, 2001; Scribner *et al.*, 2001). The most common pattern is with haplotype sharing between species, although this may be difficult to distinguish from incomplete lineage sorting for species with a long history of backcrossing and introgressive hybridisation (Funk & Omland, 2003; Joly *et al.*, 2009). Hybridisation events are additionally difficult to reconcile with standard bifurcating phylogenetic trees, especially where single gene trees are concerned (Kubatko, 2009). Incongruences due to hybridisation are sometimes documented in the DNA barcoding literature, and in particular for birds, which are well studied in this respect (Kerr *et al.*, 2009a). McKay & Zink (2010) estimate 5.7% of the non-monophyletic patterns they examined from bird studies were caused by hybridisation; again an additional 21.3% of cases could not be distinguished between hybridisation and incomplete lineage sorting. The extent to which this affects other groups is less clear, but more broadly, Mallet (2005) estimated at least 10% of animal species hybridise. Regardless, introgressed individuals create problems for mtDNA based

identification systems (Le Roux & Wieczorek, 2009; Moritz & Cicero, 2004; Teletchea, 2009).

1.3.4.4 Taxonomy

Problems of non-monophyly can arise through human interpretations, and specifically as expressed through taxonomy. McKay & Zink (2010) estimate that 55.7% of the non-monophyletic patterns in the bird studies examined were caused by incorrect taxonomy. This is significant when compared to the lower rates estimated from incomplete lineage sorting and hybridisation (see above). These taxonomic discrepancies can occur in the following ways.

Firstly, imperfect taxonomy: the species hypotheses generated as part of taxonomic studies—and almost exclusively using morphological data—may not be congruent with patterns observed in mtDNA gene trees. Biological reasons that cause these incongruences can exist, and could be due, for example, to incomplete lineage sorting as explained above, or a lack of molecular divergence between the nominal taxa. In these cases, and given that few concepts of species require monophyly at mtDNA loci (Barraclough & Nee, 2001; Meier, 2008), a lack of monophyly cannot refute a hypothesis of speciation in light of other data (de Queiroz, 2007; McKay & Zink, 2010). On the other hand, the taxonomy could simply be incorrect, and the mtDNA tree shows a more accurate relationship (Funk & Omland, 2003); this may be the case in groups that have not received a modern treatment.

Secondly, due to nomenclatural rules and the changing of taxonomic hypotheses, there are more names available than currently valid taxa, i.e. synonyms are prevalent (Eschmeyer, 2010b). If not treated correctly, these kind of discrepancies can create artificially non-monophyletic groups. For undescribed taxa, the situation is worse, with no standardisation between informal "tag-names" (Leschen *et al.*, 2009). The management of names is now becoming a significant hurdle to biodiversity informatics and also DNA barcoding (Patterson *et al.*, 2010). This is potentially a significant problem in ornamental fishes from diverse tropical regions, where taxonomy is yet to stabilise.

Thirdly, when specimens are gathered for molecular study, they may not have been identified competently, by for example, a taxonomist or specialist on the group (Bortolus, 2008; Nilsson *et al.*, 2006; Steinke & Hanner, 2011). Therefore any misidentifications at this stage can again create artificial patterns of non-monophyly similar to that observed in the biological ways listed above.

1.4 Analysing DNA barcode data

Despite much of the standardisation that DNA barcoding has achieved, the methods of data analysis often differ considerably between studies (Casiraghi *et al.*, 2010). In one respect this is to be expected, as individual objectives will differ to some extent. However, a more overarching target is usually to simply calculate the effectiveness of a reference library, i.e. how accurate are the identifications using barcode data. It is here that it is less clear as to what are the accepted methods. Generally, identification success is measured as the overall degree of congruence between *a priori* specimen identifications based on morphological data (Vogler & Monaghan, 2007). The taxonomic names provide the index for matching the morphological with the DNA barcode identifications. Although unquantified here, there appears to be a discrepancy between studies critically analysing the practical effectiveness and theoretical validity of various methods (e.g. Austerlitz *et al.*, 2009; van Velzen *et al.*, 2012; Virgilio *et al.*, 2010), and the many studies just reporting and describing barcode data. These latter studies will provide a descriptive summary of the data, including for example, mean, minimum and maximum intra-/interspecific variation among taxa, and a histogram showing a distribution of the same data (see Cawthorn *et al.*, 2011); few studies explicitly quantify identification accuracy (Little & Stevenson, 2007). Outlined below are several methods used to measure identification accuracy. This is not intended to be an exhaustive list; Casiraghi *et al.* (2010), van Velzen *et al.* (2012), and Goldstein & DeSalle (2011) provide more information.

1.4.1 Similarity methods

1.4.1.1 Genetic distances

Similarity methods using genetic distances are generally the backbone of most DNA barcoding studies. A distance matrix is constructed, with the variable sites between each pairwise comparison within the total ~651 bp alignment providing the proportion of difference between two comparisons (Nei & Kumar, 2000). Therefore, an alignment with 10 base pair differences over 651 sites has a raw genetic distance of 0.0154 (or 1.54%). In most studies, the Kimura two-parameter (K2P) model is used to correct for unobserved substitutions (Casiraghi *et al.*, 2010; Hebert *et al.*, 2003a). An important problem with distances, are that they are phenetic, i.e. they compress multiple individual changes (character state differences) into a single value of overall similarity. Therefore, potentially valuable information can be lost with

this approach (DeSalle, 2007; Will & Rubinoff, 2004; Zhang *et al.*, 2008), especially when the number of nucleotides diagnosing species is small (Lowenstein *et al.*, 2009). Distances can be used directly for identification purposes, or used for constructing phylogenetic trees (see Section 1.4.2). For specimen identification using distance data, there are a variety of different criteria which can be applied (e.g. "best match" or "best close match"); these are outlined by Meier *et al.* (2006) and Virgilio *et al.* (2010). Commonly, a per cent threshold or cut-off value is used to distinguish intra- from interspecific variation (also see Section 1.3.3).

1.4.1.2 BLAST

The Basic Local Alignment Search Tool, BLAST, in its many incarnations (initially Altschul *et al.*, 1990), is another similarity method used in barcoding studies (Little & Stevenson, 2007). Unlike the standard genetic distance measures above, it does not require a pre-aligned database, and sequences of variable length can be queried. BLAST searches short motif patterns and scores its closest hits by similarity (Casiraghi *et al.*, 2010). The algorithm is frequently used to match queries against the GenBank database (Lowenstein *et al.*, 2009). BLAST has, however, an array of different parameter settings, and as such is reported to be incorrect and inconsistent under certain conditions (Anderson & Brass, 1998; Koski & Golding, 2001; Munch *et al.*, 2008; Ratnasingham & Hebert, 2007). The results can also be ambiguous to interpret when an identical match is not found in the database (Goldstein & DeSalle, 2011). Little (2011) compared some of the implementations of BLAST, simulating DNA barcoding scenarios using different markers and querying mini-barcodes versus full length sequences.

1.4.2 Tree-based methods

Tree-based methods operate by the hierarchical clustering of sequences, and are visualised in terms of phylogenetic relations in a dendrogram (Page, 2012). Trees can be created using a variety of methods (Baldauf, 2003; Nei, 1996), and these fall into two categories: distance methods, and discrete data methods. The latter includes maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI), and the former include neighbour-joining (NJ) and UPGMA (Unweighted Pair Group Method with Arithmetic means). The discrete data methods resolve more accurate phylogenies, especially for deeper branches (Austerlitz *et al.*, 2009), but are computationally the most demanding (Baldauf, 2003; Nei, 1996). DNA barcode

datasets are typically larger than most phylogenetic datasets, so NJ clustering is the most popular (Casiraghi *et al.*, 2010; Goldstein & DeSalle, 2011). For identification purposes, tree-based methods require monophyletic groupings; thus, species are required to be monophyletic with regard to the query for tree-based methods to give a correct identification (Goldstein & DeSalle, 2011). For this reason and others, tree-based methods have been repeatedly criticised on both philosophical (DeSalle *et al.*, 2005; Goldstein & DeSalle, 2011; Little & Stevenson, 2007; Meier *et al.*, 2008; Will & Rubinoff, 2004), and empirical grounds (Little, 2011; Lowenstein *et al.*, 2009; Virgilio *et al.*, 2010).

1.4.3 Character-based methods

Unlike the phenetic approaches, character-based methods use each nucleotide as an independent source of information (DeSalle *et al.*, 2005). These rely on shared similarity rather than overall similarity (Little, 2011), and are reported to work better for closely related taxa with few or conflicting sequence information separating species (Lowenstein *et al.*, 2009). The most common implementation of character diagnostics is via the CAOS program (Sarkar *et al.*, 2008), but also see DNA-BAR (DasGupta *et al.*, 2005), and DOME-ID (Little & Stevenson, 2007). Character methods are often reported to be superior over distance approaches (DeSalle *et al.*, 2005; Goldstein & DeSalle, 2011; Lowenstein *et al.*, 2009; Rach *et al.*, 2008). However, there have been few studies comparing the two approaches directly (but see Little, 2011; Rach *et al.*, 2008; Reid *et al.*, 2011; Yassin *et al.*, 2010; Zou *et al.*, 2011).

1.4.4 Statistical, coalescent, and machine learning methods

An increasing level of sophistication can be applied to the question of specimen identification, and techniques using methods other than those based on phylogenetics are being developed. Some of these methods are based directly on the sequence data (i.e. character-based), while others operate upon distance matrices (i.e. phenetic). Zhang *et al.* (2008) and Zhang & Savolainen (2009), presented an artificial intelligence approach using back-propagating neural networks; their method appears effective and promising for cases where species are not monophyletic. Austerlitz *et al.* (2009) presented a range of supervised classification methods (CART, random forest, and kernel); they found no one method was best in all simulations. Logic methods have also been used, and can offer the desirable quality of a measure of confidence in

each specimen assignment; Bertolazzi *et al.* (2009) developed a character-based logic mining approach, while Zhang *et al.* (2012) used a distance-based fuzzy logic technique. Probabilities of identification can also be generated using genealogical and population-genetic approaches, and include the Bayesian-coalescent methods of Nielsen & Matz (2006) and Abdo & Golding (2007), or the statistical phylogenetic methods of (Munch *et al.*, 2008). Statistical methods are particularly attractive to important biosecurity, quarantine, or forensic applications, as a measure of group membership probability can be incorporated (Boykin *et al.*, 2012; van Velzen *et al.*, 2012). However, due to the relatively small amounts of information content in DNA barcodes of closely related species, difficulties may arise in parameterising these probablistic models (van Velzen *et al.*, 2012). A coalescent technique is also used by Pons *et al.* (2006) and Monaghan *et al.* (2009); the general mixed Yule-coalescent (GMYC) models the probability of transition between speciation-level (Yule model) and population-level (coalescent model) processes of lineage branching, and offers a likelihood based test of biological pattern in the data, i.e. approximating the "barcoding gap" of intraspecific versus interspecific variation. The problem of heterogeneous coalescent depth is also overcome with the GMYC, as multiple thresholds can be incorporated. Unlike the other methods mentioned above, the GMYC was not designed as a identification method, but as a parataxonomic or primary species delimitation tool; it can, however, be used for identification purposes.

1.5 Opportunities for new diagnostic approaches

Using novel DNA barcoding *sensu lato* techniques, and capitalising upon the wealth of information and experimental protocols created from DNA barcoding studies, new possibilities are opening for data to be applied to previously difficult questions (Frézal & Leblois, 2008; Valentini *et al.*, 2009). For forensic applications, and where sufficient population level sampling has taken place, identification of specimens can now proceed without the need for sequencing, with identifications carried out by DNA hybridisation on microarray chips (Hajibabaei *et al.*, 2007; Kochzius *et al.*, 2010; Summerbell *et al.*, 2005; Teletchea, 2009; Teletchea *et al.*, 2008). For more specific/local applications, DNA melting analyses from DNA barcode data can now be used to get rapid identifications with a simple PCR protocol (Winder *et al.*, 2011). Novel ways of extracting DNA from damaged or valuable specimens have also emerged, including from formalin-preserved fish specimens (Zhang, 2010), 80 year

old fish tissues (de Bruyn *et al.*, 2011), live beetle larvae (Lefort *et al.*, 2011), and the ethanol preservative many specimens are stored in (Shokralla *et al.*, 2010). More generally, prospects for making identifications from degraded tissues are improving, with mini-barcode methods shown to be surprisingly effective (Dubey *et al.*, 2010; Hajibabaei *et al.*, 2006b).

Detection of species is now no longer limited to their physical collection. Surveying environmental DNA (eDNA) can reveal the presence of rare or invasive species, and even monitor their movements through ecosystems (Darling & Mahon, 2011; Dejean *et al.*, 2011; Ficetola *et al.*, 2008; Goldberg *et al.*, 2011; Jerde *et al.*, 2011; Minamoto *et al.*, 2012; Thomsen *et al.*, 2012). Next generation sequencing techniques are now capable of producing vast quantities of data compared to standard Sanger methods (Mardis, 2008; Taylor & Harris, 2012). This now allows a meta-barcoding approach, whereby entire faunas can be surveyed by proxy through environmental DNA traces (Andersen *et al.*, 2012). Further mitochondrial protein-coding genes can also now be sequenced using a ROCHE 454 platform at relatively low cost, providing markers for systematics applications, and additional data to potentially improve DNA barcode resolution (Timmermans *et al.*, 2010).

1.6 Research rationale, outline, and objectives

As previously stated, international trade in ornamental fishes is a high volume industry distributing millions of aquatic organisms throughout the world each year. However, it is poorly known in terms of the composition of species that are traded. Consequently there is a requirement in New Zealand, and internationally, for there to be a more rigorous assessment of the identity of these potentially invasive and disease carrying imports. Molecular approaches (DNA barcoding) have been promoted as a solution, offering potentially both precise and accurate biological identifications. However, limitations may apply to the usefulness of the method, and these are to be explored in this thesis.

Specific objectives and experimental hypotheses are outlined in each chapter introduction. Overall, the aim of the thesis is to assess how effective a DNA barcoding approach can be for a specific biosecurity application: the identification of fishes traded in the ornamental industry. In Chapter 2, the primary objective will be to assemble a DNA barcode reference library of a target fish group, and therefore provide a long term resource for MAFBNZ and other biosecurity agencies to use

and build upon. Ornamental fishes will be collected from the trade, identified using morphological data, and then barcoded using standardised protocols. A descriptive summary of the molecular data will be provided, and biological or taxonomic issues such as those highlighted in Section 1.3 will be assessed and discussed in relation to biosecurity priorities.

An important aspect in assessing the utility of molecular data for biosecurity is to thoroughly evaluate the relative merits of current analytical methods for DNA barcoding. Particularly, factors including identification criterion, choice of nucleotide substitution model, presence of singleton species, and data quality from third party sources such as GenBank, have the potential to influence or bias identification success rates in a practical context. Therefore, in Chapter 3 and Chapter 4, identification success will be tested under a variety of these criteria, assumptions and scenarios in order to gauge how robust DNA barcode data are to alternative methods of inference. Recommendations will also be made in respect to the appropriate use of identification and analytical criteria for biosecurity applications.

As discussed in Section 1.3, issues such as interspecific hybridisation—which is not uncommon in ornamental aquaculture—can create potential pitfalls when employing a solely mitochondrial approach to specimen identification in the ornamental fish trade. One solution to this problem is with the addition of genetic data from a nuclear gene. However, standardised nuclear "barcoding" genes have received little attention, so in Chapter 5, candidate nuclear markers will be assessed for suitability, and the resulting data will be applied to assisting with the recognition of both interspecific hybrids, and the putative cryptic species frequently encountered in DNA barcoding studies.

In Chapter 6, new and promising avenues in diagnostic research will be investigated, potentially providing novel methods for improving the capacity of biosecurity agencies to more effectively solve problems emerging from the ornamental fish trade. Specifically, environmental DNA technologies could be a useful quarantine tool, with the potential for reliably detecting high risk organisms in ornamental fish quarantine centres, simply through water sampling. Therefore, whether such a non-invasive sampling approach is effective in providing identifications will be tested. Additionally, factors important when recovering degraded DNA from environmental samples will be explored, and in particular, the variability of small-fragment molecular-markers to make species level identifications will also be assessed.

Chapter 2

An evaluation of DNA barcoding for the identification of ornamental cyprinid fishes

2.1 Introduction

2.1.1 Ornamental cyprinid fishes

Freshwater ornamental fishes comprise a diverse group, with up to 150 families reported to be represented by Hensen *et al.* (2010). One of the common families is the Cyprinidae (Teleostei: Cypriniformes), and Hensen *et al.* (2010) record 333 species of this group in the aquarium trade. The global diversity is far higher, however, at over 2,400 species (Nelson, 2006). Many, such as the barbs, danios and rasboras are popular aquarium and pond fishes, being ubiquitously available at low prices from aquarium and general pet-retailers. In particular, the danios and barbs are frequently promoted as being suitable for beginner aquarists.

Cyprinid fishes are naturally found across Africa, Europe, North America and Asia, although many have been introduced outside this range (Berra, 2007). The majority of wild ornamental species are sourced from India, Burma, Thailand, Indonesia, and occasionally Africa (Nigeria or Congo). Farmed species usually arrive in New Zealand via transshippers in Singapore, and are sourced from farms in Florida, Sri Lanka, Israel, and across Southeast Asia (Ploeg *et al.*, 2009).

2.1.2 Biosecurity risk

Cyprinid fishes represent risk in terms of both their potential as invasive species, and as vectors of exotic pathogens (MAF Biosecurity New Zealand, 2011; Ploeg *et al.*, 2009; Rahel, 2007; Whittington & Chong, 2007). In terms of potential for invasiveness among all potential aquarium species imported into New Zealand,

McDowall & James (2005) presented a thorough review. Their key recommendations were that likelihood of invasion is unpredictable, and a precautionary approach should be taken. This meant restricting the breadth of imported fishes at the point of entry, and ascertaining which species were already present at that time in New Zealand. However, in this respect, taxonomic capacity was identified as a limiting factor in MAFBNZ's ability to respond to difficulties in identifying the vast number of potentially traded species. This is particularly the case where fishes are poorly known or undescribed, their nomenclature has changed, or are traded as juveniles.

Hine & Diggles (2005) made parallel assessment in terms of disease risk of ornamental fish imports to New Zealand. In particular, temperate and subtropical cyprinid fishes such as some *Puntius* and *Barbus* species were identified as a substantial threat in terms of pathogen vectoring, carrying zoonotic diseases such as the bacterium *Edwardsiella*. The study also recommended that species not already present in the country should be determined as new organisms under ERMA (Environmental Risk Management Authority) regulations and the Hazardous Substances and New Organisms (HSNO) Act.

Subsequent to both of these reports (Hine & Diggles, 2005; McDowall & James, 2005), the list of New Zealand permitted species was updated in light of a survey of fishes present in the country, with help of the FNZAS (Federation of New Zealand Aquatic Societies). An Import Health Standard (IHS) is now in place permitting only the import of the species listed (as opposed to genera previously). There are 82 permitted cyprinid fish species now listed on the IHS for import, with 27 of these in the IHS Appendix 2 "high risk" category (in terms of exotic diseases). Imported fishes are now subjected to a four week quarantine period, with additional risk mitigation procedures and targeted disease surveillance in place for the IHS Appendix 2 species (MAF Biosecurity New Zealand, 2011).

2.1.3 Sampling strategies and GenBank data

Due to the difficulties in morphological/visual fish identification outlined above, and in Section 1.1.4, molecular methods can be therefore be recommended here, assuming the reference library is correct and the data are able to discriminate effectively. Steinke *et al.* (2009b) provided barcode data for 391 species available in the marine trade, but for freshwater ornamental species, and especially cyprinid fishes, few molecular data are currently available. The sequences available from GenBank are from a variety of mtDNA markers (frequently cyt *b*), and often have no voucher

material associated with them. Therefore their use is limited for diagnostic purposes (Ward *et al.*, 2009). Ornamental cyprinid fish species are also under-represented in the BOLD database, and the possibilities of making accurate species level identifications solely using this resource are currently poor. DNA barcodes generated in this study will provide the basis for an improved ornamental fish reference library, and will be uploaded to BOLD, along with supplementary information.

Overall, a number of cyprinid fish species are, however, represented with COI sequences on GenBank. Many of these may not be available in the aquarium trade, but a proportion will be congeners to those which are. Therefore, in order to expand taxon coverage, and to assist in identification of target species, the utility of extra data for non-target species in GenBank will be assessed. There will be sequences available for additional, new species, but the databases may also include sequences from misidentified specimens or specimens collected from otherwise unsampled, divergent populations (Harris, 2003; Meier *et al.*, 2006; Ward *et al.*, 2009).

2.1.4 Data presentation

With advances in technology, and subsequently increasing amounts of data, new bioinformatic problems are emerging: one of these is the way in which to effectively present phylogenetic hypotheses (Page, 2012). Typically, in published DNA barcoding studies, NJ phenograms (trees) are displayed as embedded image files. However, embedding text into flattened raster images (image rendered pixel-by-pixel) removes local as well as global (Internet) search engine visibility for those taxa. Vector graphic (image rendered by paths) solutions overcome this problem, but large trees remain unwieldy, even as appendices or supplemental data. There is also the problem of tables of species lists; see Lakra *et al.* (2011) as an example of where much of the article is occupied with rasterised NJ trees and lists of species sampled. As studies use more and more data, these problems become increasingly untenable. A significant challenge will be the linking of biodiversity information from primary research to that already present in databases, and for it to therefore remain future-proof in terms of nomenclatural stability, and be accessible over time (Patterson *et al.*, 2010). A recently proposed method could potentially address some of these problems simultaneously; Smits & Ouverney (2010) presented a javascript library for scalable vector graphics (SVG), allowing phylogenetic trees to be displayed in a Web browser rather than a document viewer. Importantly, the trees are interactive, containing within the HTML code persistent URLs leading to the database records for

each specimen. This serves as both a phenogram, a list of species which can easily be searched, and a stable link to additional online resources such as GenBank or BOLD.

2.1.5 Objectives

The primary objective of this chapter is to sample the cyprinid fishes currently found in the aquarium trade internationally, identify them to species using taxonomic literature, test a fit-for-purpose lab protocol for generating DNA barcodes, and assemble a reference library on BOLD. The DNA barcodes will then be assessed by comparing patterns of congruence with the taxonomic identifications. Summary statistics will be generated along with measures of sampling effort, and taxonomic inconsistencies will also be discussed. New methods of data presentation will also be explored. This chapter is primarily methodological and descriptive, and so does not attempt to quantify identification success (see Chapter 3).

2.2 Materials and methods

2.2.1 Specimen sampling

2.2.1.1 Specimen acquisition

Specimens of ornamental cyprinid fishes were acquired from aquarium retailers, wholesalers and exporters in the United Kingdom, Singapore and New Zealand during 2008 to 2010. The non-cyprinid taxa *Gyrinocheilus* and *Myxocyprinus* were also included due to their ubiquity in the trade and superficial morphological similarity to some cyprinid fishes. Specimens were euthanised with MS-222 (tricaine methane sulfonate), before a tissue sample was excised from the right-hand caudal peduncle and stored at $-20°C$ in 100% ethanol. Specimens were subsequently formalin fixed and preserved in 70% ethanol as vouchers, following the procedures outlined by Kottelat & Freyhof (2007). At least one specimen from each sample was photographed alive (left-hand side) prior to tissue sampling, with the remainder photographed after preservation. See Appendix A for further details of how tissue samples were taken, and voucher material preserved. Voucher specimens for each COI barcode were deposited at the Raffles Museum of Biodiversity Research (ZRC), National University of Singapore.

2.2.2 Assessment of sampling strategy

Whenever possible, multiple individuals of each species were sampled. In order to better assess intraspecific genetic diversity, multiple specimens were purchased at different times and from different vendors. Sampling efficiency was tested by correlating the number of haplotypes observed in each species with the number of individuals collected and the number of samples taken. For this purpose, a sample was considered as all conspecific specimens acquired from the same holding tank at the same premises on the same visit. These analyses were carried out in R version 2.12.1 (R Development Core Team, 2010), using a generalised, linear regression model with poisson distributions for count data; singleton species (species represented by one individual) were omitted. A haplotype accumulation/rarefaction curve was generated to make an assessment of intraspecific variation captured (cf. Gotelli & Colwell, 2001; Zhang *et al.*, 2010). To assess the coverage of the project in terms of species-level sampling, a list of species believed to be in the aquarium trade was consulted as the most up-to-date and accurate guide available at this time (Hensen *et al.*, 2010); the MAFBNZ Import Health Standard list of species was also used to gauge coverage in terms of biosecurity risk species (MAF Biosecurity New Zealand, 2011).

2.2.3 Morphological identification

Specimens were identified using morphological characters from the scientific literature relevant to the group. A bibliography was therefore first assembled by searching the *Catalog of Fishes* (Eschmeyer, 2010a) for the genera and possible species encountered. Original descriptions were consulted where possible. The taxonomic publications were obtained from current journal subscriptions, hobbyist/scientist contacts, or when out-of-copyright, via the Biodiversity Heritage Library (URL: http://www.biodiversitylibrary.org/). Much of the essential literature was still unavailable, however, through these channels. Therefore, a visit to the Natural History Museum, London was made to access the remaining literature from their extensive library[1].

The use of "sp.", "cf." and "aff." notation in reference specimen identification follows Kottelat & Freyhof (2007). For analytical purposes, individuals designated "cf." are treated as conspecific with taxa of the same specific name, while those

[1] It must be noted that this was round trip of over 40,000 km (roughly the circumference of the Earth), and produced over 3.42 metric tons of carbon dioxide.

designated "aff." are treated as non-conspecific. Nomenclature follows Eschmeyer (2010a), unless otherwise stated.

2.2.4 DNA protocols

2.2.4.1 DNA extraction and PCR

Approximately 2–3 mm^2 of white muscle tissue was prepared for genomic DNA extraction using the Quick-gDNA spin-column kit (ZYMO RESEARCH CORPORATION) following the manufacturer's protocol, but scaled to use a 50% volume of pre-elution reagents. Optimised PCR reactions were carried out using a GeneAmp 9700 thermocycler (APPLIED BIOSYSTEMS) in 10 µl reactions[1]. Amplification of the COI barcode marker comprised reactions of the following reagents: 2.385 µl ultrapure water; 1.0 µl Expand High Fidelity 10× PCR buffer (ROCHE DIAGNOSTICS); 0.54 µl $MgCl_2$ (25.0 mM); 2.0 µl dNTPs (1.0 mM); 1.5 µl forward and reverse primer (2.0 µM); 1.0 µl DNA template; 0.075 µl Expand High Fidelity polymerase (ROCHE DIAGNOSTICS).

The COI fragment was amplified using one of the following primer pairs: FishF1 and FishR1 (Ward *et al.*, 2005), LCO1490 and HCO2198 (Folmer *et al.*, 1994), or LCO1490A and HCO2198A (Tang *et al.*, 2010). Thermocycler settings for COI amplification were as follows: 2 min at 94°C; 40 cycles of 15 s at 94.0°C, 30 s at 48.0°C (LCO/HCO) or 52.0°C (FishF1/R1), and 45 s at 72.0°C; 7 min at 72.0°C; ∞ at 4.0°C.

2.2.4.2 Sequence data

Prior to sequencing, PCR products were checked visually for quality and length conformity on a 1% agarose gel. Bidirectional sequencing was carried out following the manufacturer's protocol on a Prism 3130xl Genetic Analyser (APPLIED BIOSYSTEMS) using the BigDye Terminator v3.1 Cycle Sequencing Kit (APPLIED BIOSYSTEMS). The same primer combinations as for PCR amplification were used for sequencing. Sequencing products were purified using the Agencourt CleanSEQ system (BECKMAN COULTER GENOMICS). Steps undertaken here to avoid or identify cross-amplification of nuclear mitochondrial pseudogenes (NUMTs) are outlined by Buhay (2009) and Song *et al.* (2008). Sequence chromatograms were inspected visually for quality and

[1]Final concentrations of reagents are as follows: 1× buffer; 2.85mM $MgCl_2$; 0.2 mM dNTPs; 0.3 µM per primer; 0.26 U polymerase.

exported using FinchTV 1.4 (GEOSPIZA). Trimmed nucleotide sequences were aligned according to the translated vertebrate mitochondrial amino acid code in the program MEGA 4.1 (Tamura *et al.*, 2007). The resulting COI fragment comprised a sequence read length of 651 base pairs (bp), positionally homologous to nucleotides 6,476 through 7,126 of the *Danio rerio* mitochondrial genome presented by Broughton *et al.* (2001). Sequence data, chromatogram trace files, images and supplementary information were uploaded to BOLD, and are publicly available in the "Ornamental Cyprinidae" [RCYY] project. See also Appendix B.

2.2.5 GenBank data search

In addition to sequence data generated here, public databases including GenBank and BOLD were searched under the following terms: "Cyprinidae", "COI", "CO1" and "COX1". Records were retained if the taxon in question was believed to occur in the aquarium trade (Hensen *et al.*, 2010), or if congeneric to a species that had already been collected during sampling. For the purposes of simplification, these data are herein termed "GenBank", although they comprise data from both the GenBank and BOLD databases. To facilitate analysis, nomenclature and spellings of GenBank records were updated or corrected following Eschmeyer (2010a).

2.2.6 Summary statistics

All descriptive statistics and analyses were conducted using SPIDER, the DNA barcode analysis package for R (Brown *et al.*, 2012; Paradis *et al.*, 2004). Distance matrices and NJ phenograms were generated under Kimura's two-parameter model (K2P/K80) using the APE package (Paradis *et al.*, 2004), with missing data treated under the "pairwise deletion" option. The K2P model was used to ensure consistency and comparability with other barcoding studies, but see Chapter 4 for an analysis of the applicability of the K2P model. Summary statistics were generated using the *checkDNA*, *dataStat*, *seqStat*, *nonConDist* and *maxInDist* functions of SPIDER. Negative branch lengths were set to zero (Ross *et al.*, 2003; Saitou & Nei, 1987). Terminology of topological relationships follows phylogenetic nomenclature consistent with literature (e.g. monophyly, paraphyly, polyphyly); however, this does not imply explicit evolutionary relationship. The barcoding gap is defined as the proportion of individuals for which the minimum non-conspecific (i.e. interspecific) distance is greater than the maximum intraspecific distance for that species.

2.2.7 Data presentation

NJ phenograms were rendered in Web-based jsPhyloSVG format (Smits & Ouverney, 2010), following conversion from NEXUS format into PHYLOXML using ARCHAEOPTERYX (Han & Zmasek, 2009). This creates an interactive vector-graphic phenogram with links to specimen database records and supplementary data (e.g. images) via embedded URLs. Further instructions for viewing the phenogram can be found in Appendix Section B.3.

2.3 Results

2.3.1 Morphological identifications and taxon sampling

A total of 678 cyprinid specimens were collected during the study from the UK (11 retailers throughout the country), Singapore (3 wholesalers, 3 retailers) and New Zealand (6 retailers in Christchurch). These specimens were identified to 172 species in 45 genera using morphological characters from 156 taxonomic references. Ten species were found to differ substantially from published literature and are believed to be possible new species (labelled "sp." or "aff."); four could not be assigned to any species given the literature available (labelled "sp. undetermined"); and 29 examples were uncertain members of a species (labelled "cf."). Refer to Appendix C for a full list of the assignments, characters used for identification, taxonomic comments, and bibliography.

The survey of GenBank and BOLD databases contributed a further 562 COI sequences from 157 species, with 81 of the species represented in both GenBank data and the new data presented here (Table 2.1). With regard to the aquarium trade, the taxon coverage of this study represents 131 (39%) of the 333 aquarium cyprinids listed in Hensen *et al.* (Hensen *et al.*, 2010), a proportion which increased to 56% coverage when GenBank data were also included. An additional 41 species not present in this inventory (Hensen *et al.*, 2010) were reported from the survey of the trade presented here. In terms of biosecurity risk, the taxon sampling of this study covered 78% (85% including GenBank) of the 27 cyprinid species listed as high-risk allowable imports to New Zealand (MAF Biosecurity New Zealand, 2011); of the total 82 permitted cyprinids, our data represented 79% of these (90% including GenBank).

Table 2.1. Summary of descriptive statistics for DNA barcodes from the three data partitions analysed in the study.

Statistic	This study	GenBank	Combined
Individuals	678	562	1240
Species (no. unique sp.)	172 (91)	238 (157)	329
Mean individuals per sp. (range)	3.9 (1–12)	2.4 (1–42)	3.8
Singletons	20	125	97
Genera	45	63	65
Mean sampling events per sp. (range)	2.32 (1–8)	-	-
Mean seq. length bp (range)	645 (378–651)	639 (441–651)	643 (378–651)
No. barcodes < 500 bp	5	1	6
Mean haplotypes per species	1.97 (1–7)	1.61 (1–8)	2.07 (1–10)
Mean intraspecific dist. (range)	0.90% (0–14.7%)	0.86% (0–24.1%)	1.13% (0–24.1%)
Mean smallest interspecific dist. (range)	9.11% (0–23.2%)	8.40% (0–26.0%)	8.06% (0–26.0%)
95% intraspecific var. $\leq$	5.48%	2.13%	6.85%
95% smallest interspecific dist. $\geq$	1.72%	0.00%	0.15%
Prop. intraspecific dist. > 1%	19.0%	32.2%	28.3%
Prop. intraspecific dist. > 2%	13.5%	5.90%	12.7%

Ranges or subsets are presented in parentheses. Abbreviations: dist. = distance(s); no. = number; prop. = proportion; seq. = sequence; sp. = species; tot. = total; var. = variation. “Combined” refers to data generated in this study combined with collected GenBank/BOLD data.

2.3.2 Barcode sampling

DNA barcodes were successfully amplified from all samples in the study with at least one of the primers reported. All nucleotides translated into functional protein sequences in the correct reading frame, with no stop codons or indels observed in the data. Regarding sequence quality, 100% scored as "high quality" by BOLD ($< 1\%$ Ns). In terms of trace quality, 94.6% of the chromatograms (trace files) scored as "high quality" according to BOLD's criteria. In the COI barcode dataset, each species was represented by an average of 3.9 individuals (2.32 sampling events), with twenty species by one individual (11.6%), and 102 (59%) by ≥ 3 individuals (Table 2.1). The average number of haplotypes per species was 1.97, with sampling effort (sampling events and number of individuals per sp.) and haplotype diversity correlated ($P < 0.001$). The accumulation/rarefaction curve of haplotypes (Figure 2.1) shows no asymptote as sample size increases, with an almost linear relationship.

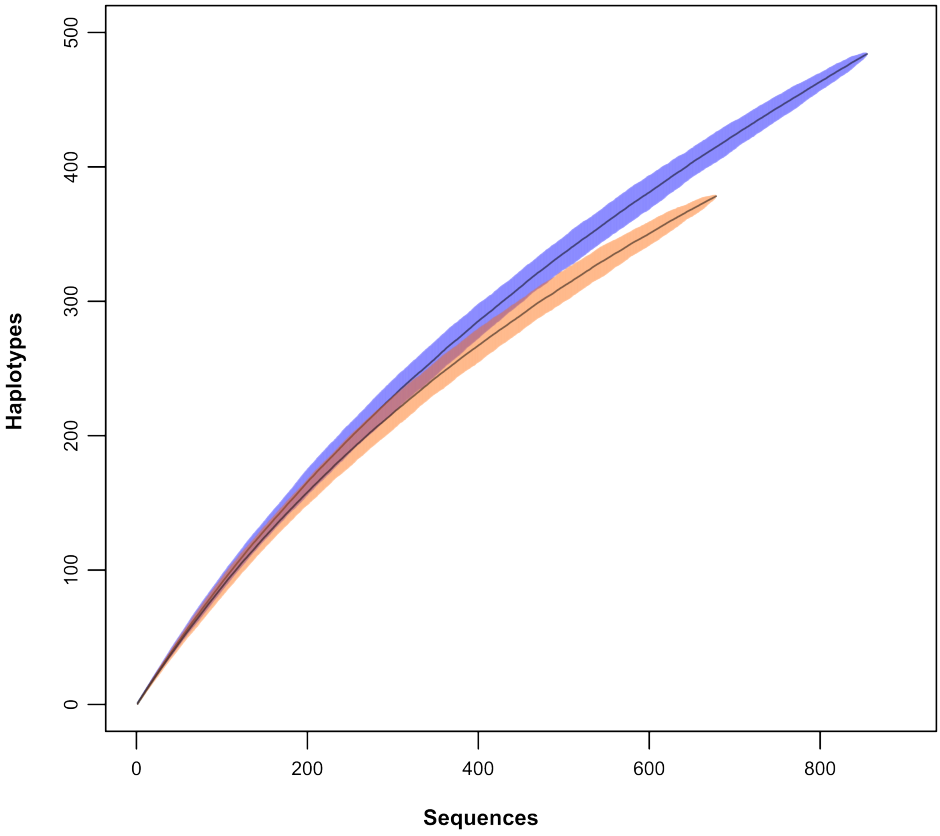

Figure 2.1. Haplotype accumulation curve for sequences generated in this study (orange) and when combined with sequences from the same species in the GenBank data (blue). Confidence intervals are shown by coloured shading.

2.3.3 Description of barcode data

A full description of the data, partitioned by source, is given in Table 2.1. Genetic diversity was generally lower within species than between, with 95% of total intraspecific variation less than 5.48% K2P distance. Of the interspecific distances to a closest non-conspecific neighbour—the "smallest interspecific distance" of Meier *et al.* (2008)—95% were above 1.72% K2P distance. Mean distance to closest non-conspecific was 10× mean intraspecific distance. Of the intraspecific values, 13.5% were over 2% K2P distance, while 19.0% were above 1%. A total of 167 of the total 172 species (97%) were recovered as monophyletic for the data generated in this study. When combined with GenBank data 287 of 329 species (87%) were found to be monophyletic. A barcoding gap was reported for 655 of the 678 individuals in this study (97%), and for 1054 of the 1240 individuals when GenBank data were added (85%). A dotplot representation of the barcoding gap is shown in Figure 2.2. Species that fell on or below the barcoding-gap line are discussed in Section 2.3.4. See Chapter 3 for discussion of identification success.

Graphical structure of the distance data (total dataset including GenBank) is shown in the NJ phenogram presented in online Appendix Section B.3, and indicates cohesive clusters for the majority of species. This includes many morphologically similar species such as the *Puntius* spp. shown in Figure 2.3, which were well differentiated with DNA barcodes. Links to BOLD and GenBank database records for all sequences used here are presented as URLs in online Appendix Section B.3. Sequence data are provided as a text file in FASTA format, and are available in online Appendix Section B.1.

2.3.4 Incongruences between data

Cases of incongruence and inconsistency for some common aquarium species are presented in a reduced NJ phenogram (Figure 2.4). These are illustrated by barcode sharing observed in two groups: between two *Eirmotus* species (*E.* cf. *insignis* and *E.* cf. *octozona*), and between two *Rasbora* species (*R. brigittae* and *R. merah*). Additionally, a polyphyletic species was observed: an individual of *Danio* cf. *dangila* (RC0343) clustered closer to *D. meghalayensis* than to other *D. dangila*.

When GenBank data were added, several additional species were also non-monophyletic on the COI gene tree, with these added data conflicting with some barcodes generated in this study. For example, *D. albolineatus* became polyphyletic with the inclusion of *D. albolineatus* HM224143, as did *D. roseus* when *D. roseus*

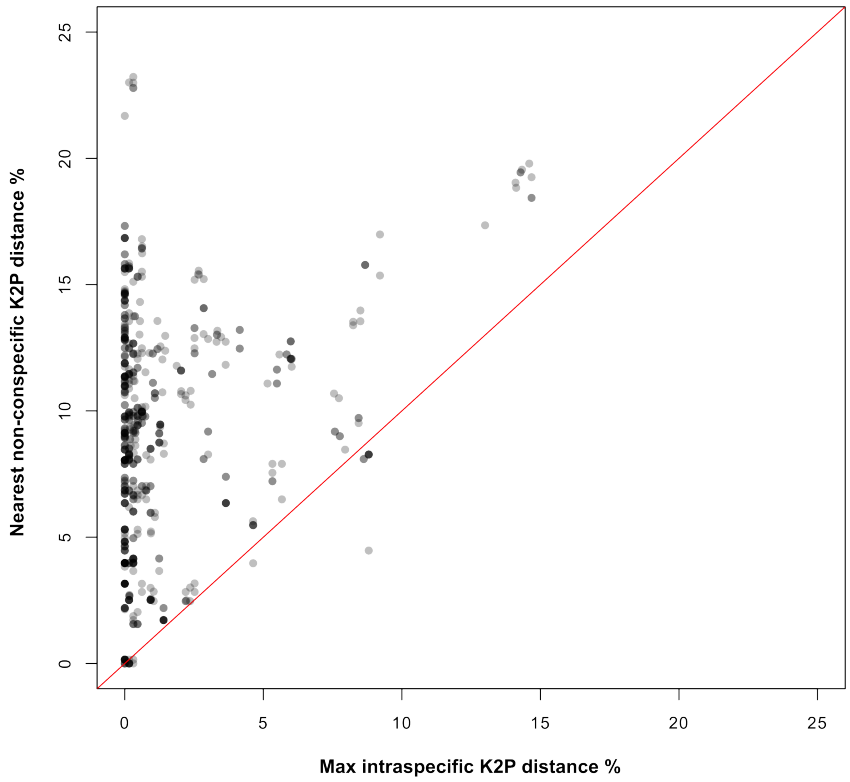

Figure 2.2. Dotplot illustrating the barcoding gap for sequences generated in this study. For each individual, the maximum distance to an intraspecific individual is plotted in relation to the distance to the nearest non-conspecific individual (minimum interspecific distance). The red line shows a 1:1 relationship of intra- and interspecific distances, i.e. above the line the interspecific distances are greater than intraspecific (barcoding gap present), and those on or below the line are where interspecific distances are equal to or less than intraspecific distances (barcoding gap absent). Density of points is shown by colour (dark = overplotted points).

HM224151 was added. In regard to these species, the topology of the NJ phenogram (Figure 2.4) is misleading for identification purposes, however; all *D. roseus* remain diagnosable from *D. albolineatus* by a single transversion at position 564, while the remaining differences in *D. roseus* HM224151 are autapomorphies. Other aquarium species that were affected by GenBank data inclusion include (refer to Figure 2.4): haplotype sharing between a possibly undescribed *Devario* ("TW04") and *D. annandalei* HM224155; haplotype sharing and polyphyly of *R. daniconius* and *R.* cf. *dandia*; paraphyly of *Barbonymus schwanenfeldii* by *Balantiocheilos melanopterus* HM536894; paraphyly of *Devario* cf. *devario* by *D. devario* EF452866; polyphyly of *Paedocypris carbunculus*; paraphyly of *Puntius stoliczkanus* with polyphyletic *P. ticto*; polyphyly of *R. paviana* with regard to *R. hobelmani* HM224229 and *R. vulgaris*

Figure 2.3. Illustrating the utility of DNA barcodes in biosecurity. *Puntius filamentosus* (A) and *P. assimilis* (B) are two species strikingly similar in appearance; morphological differences are especially difficult to discern when these are exported as juveniles. Here, we demonstrate they can be readily separated by DNA barcodes, with the two specimens pictured here differing by a 17.6% divergence in K2P distance for COI. Also see Appendix B for NJ phenogram.

HM224243; polyphyly of *Esomus metallicus*. It is important to note that this is not a full description of all ambiguous clusters in the full NJ phenogram (Appendix Section B.3). Only a subsample of aquarium species where data were conflicting are described, while conflict between non-aquarium species represented by GenBank data are not discussed.

2.4 Discussion

2.4.1 Morphological identification

Accurately assigning correct taxonomic names to voucher specimens and DNA barcodes is a critical first step in assembling a useful reference library for non-expert users. Unlike previous studies of regional ichthyofaunas (e.g. Hubert *et al.*, 2008; Valdez-Moreno *et al.*, 2009), scientific publications covering all taxa likely to be

Figure 2.4. Incongruences and inconsistencies in barcode data. This reduced-taxon NJ phenogram highlights cases of haplotype sharing and paraphyly/polyphyly between nominal species. The phenogram shows the same "ingroup" patterns of relationship as the full NJ tree (Appendix Section B.3); i.e. removing taxa did not influence relationships discussed. Data generated in this study are prefixed "RC0", "YGN" and "EUN" (otherwise GenBank), with anomalous individuals represented in red.

encountered in the aquarium trade were not available. Even after extensive literature was gathered at great expense, identifying some of the specimens remained difficult. Liberal use of the "cf." notation where specimens examined differed from diagnoses in the literature is testament to the uncertainty in identification based on these data. In some cases, reliable guides to local faunas and up-to-date revisions existed. However, in other cases such as Indian fishes, little taxonomic research has been conducted since the original descriptions from the early 19th century.

Frequently, the morphological characters recorded in early taxonomic works are inadequate for diagnosis, being heavily reliant on subjective terminology, missing explicit comparisons, and often being incompatible with more modern techniques making use of data sources such as colour pattern (e.g. Tan & Kottelat, 2008). Morphometric characters such as relative proportions of anatomical features (e.g. depth of caudal peduncle compared to body length), were found to be almost useless for identification due to the considerable variation observed in small sample sizes and often juvenile material (see Kottelat & Freyhof, 2007, for further discussion). Meristic measurements such as scale and fin ray counts are also common in the literature (Kottelat & Freyhof, 2007). However, these are difficult to accurately take on small fishes, and frequently the distributions between species overlap, and the measurements taken invariably tended to fall within this overlap. Therefore, morphometrics and meristics were avoided where possible. Presence, absence, position, or qualitative description of shape/colour-markings (i.e. cladistic character states), were found to be most informative, but only when these were well documented in the literature.

2.4.2 Assessment of sampling strategy

The survey of the trade revealed that 24% of species available were not listed in the most recent and thorough reference list for the trade (Hensen *et al.*, 2010), indicating a mismatch between actual availability and published literature. Conversely, many species listed in this reference did not appear to be available at the wholesalers and retailers visited. Some of these discrepancies surely arise from identification and nomenclatural issues, but are otherwise likely due to changing export patterns through different regions and time, as data in Hensen *et al.* (2010) was compiled from historical information.

A strong relationship between haplotype diversity and sample frequency was observed, indicating that expanding the reference library will result in the discovery of further genetic variability. Given the relatively small sample sizes taken here (breadth

in favour of depth), it is to be expected that intraspecific sampling would fail to uncover much of the genetic diversity. Zhang *et al.* (2010) report that depending on evolutionary and demographic histories, a sample size between 4.5 and 332.9 individuals per species will estimate when most of the diversity has been sampled (i.e. ≤ 10 new haplotypes per 100 individuals sampled).

In terms of the patterns of trade, it is predicted that farmed species will have a lower genetic diversity and fewer observed haplotypes than those of wild caught species, which may make them easier to identify with DNA barcodes. Preliminary investigations have suggested that this may well be the case. However, due to difficulties obtaining reliable source information through the supply chain, and problems with establishing independence of samples (i.e. "independent" samples may have derived from a single source), these observations should be investigated further.

2.4.3 DNA barcode generation and description

The laboratory protocols provided worked effectively, resulting in high quality DNA barcodes from all specimens tested. The data generated here were considered "barcode compliant" (Hanner, 2009) by BOLD in terms of all criteria, except one: collection geo-location. This was unfortunately unavoidable due to the nature of the collection method—specimens sampled from aquarium retailers—and so lacked the important country-code data for the real distribution of the organism (Hanner, 2009). The choice of three primer pairs was perhaps unnecessary, but reflected the publication of a new cyprinid-fish primer set (Tang *et al.*, 2010) during laboratory work. The majority of the samples amplified well, however, using the general fish primers FishF1 and FishR1 of Ward *et al.* (2005). Those which did not amplify well for this set worked with either the standard Folmer *et al.* (1994) primers, or the Tang *et al.* (2010) primer pair. Use of these three primer pairs could be recommended as an appropriate procedure. However, the use of the M13-tailed fish primer cocktails presented by Ivanova *et al.* (2007) could also be considered for high throughput work. The PCR mastermix and cycling parameters appeared not to be an important factor, and generation of the DNA barcodes was found to be robust to variation as far as these were concerned; following the manufacturer's instructions, most proprietary products should give similar results. A more important consideration, however, is that of DNA extraction, with significantly better results being obtained

using a spin-column kit over some of the lower cost alternatives such as single-tube digestion methods such as PREPGEM (data not shown).

2.4.4 Patterns in DNA barcode data

Broadly the DNA barcode data agrees with the names provided during the morphological identification process, with the majority of species recovered as monophyletic. The variation within and between species was well separated, and the presence of a barcode gap suggests identification is possible (but see Chapter 3). When using the kind of sampling strategy adopted here—relatively small intraspecific sample sizes from a small number of species comprising a much larger group—the presence of well separated intra- and interspecific diversity is to be expected (Moritz & Cicero, 2004). It is anticipated, however, that intra- and interspecific variation will increase and decrease respectively, when both species and population level sampling increases (Meier *et al.*, 2008).

2.4.5 Incongruences between data

Although few in number, cases of incongruence between barcodes require careful interpretation, especially where the inclusion of GenBank data result in some common aquarium species becoming ambiguous to distinguish. However, with some background knowledge inferences can be made, and incongruence falls broadly into two categories: taxonomic uncertainty (or genetic para-/polyphyly), and conflict due to misidentifications. In the example of barcode sharing in *Eirmotus*, despite good quality specimens and the availability of a thorough, modern revision of the genus (Tan & Kottelat, 2008), our morphological identifications were uncertain (see Appendix C). DNA barcodes from this cluster could belong to either *E. octozona* or *E. insignis*, which is likely the result of these taxonomic/identification problems. Topotypic specimens would be required for a better understanding of the problem. Likewise in the case of *Rasbora brigittae* and *R. merah*, individuals of both species were observed to be inconsistent in diagnostic morphological character states (see Appendix C). Again, specimens clustering in this group could belong to either species, a finding which certainly warrants further taxonomic investigation. Haplotype sharing between the possibly undescribed *Devario* sp. “TW04” and GenBank *D. annandalei* is likely explained also by uncertainty in our identification of this individual, or the misidentification of the GenBank specimen. Due to the large number of undescribed

Devario species in Asia, and few modern treatments, identification of many wild caught *Devario* is difficult. The aberrant specimen of *Danio dangila* (RC0343) displayed slight morphological differences to the other *D. dangila*, but with only one individual available, it was conservatively regarded as conspecific (see Appendix C). A similar observation was made with *Devario* cf. *devario* having divergent barcodes from GenBank *D. devario*, and an inconsistent morphology to that of the published *D. devario* literature. The example of *Danio albolineatus* and *D. roseus* shows a situation where all specimens from the trade are homogeneous and diagnosable; however, they are rendered polyphyletic when data are included from other GenBank populations. This finding is perhaps expected given *D. albolineatus* (*sensu lato*) is a variable species with three synonyms, distributed across much of Southeast Asia (Fang & Kottelat, 2000).

Some examples certainly represent cases of misidentification, with specimens of GenBank "*Puntius ticto*" from the Mekong, grouping closer to *P. stoliczkanus*, a species with which it is often confused (Linthoingambi & Vishwanath, 2007). Other examples such as the paraphyly of *Barbonymus schwanenfeldii* by a GenBank *Balantiocheilos melanopterus* individual (HM536894), is probably a case of human error and poor quality control of data, given the marked morphological differences between the two species. Identifications made prior to recently published taxonomic works may also be subject to error. This may explain GenBank's sequences of *Rasbora daniconius*, a species formerly considered to be widely distributed but now likely restricted to the Ganges drainage of northern India (Silva *et al.*, 2010).

2.5 Summary

This chapter provides tested laboratory protocols for sampling tissues, imaging and storing specimens, and PCR amplification. DNA barcode data for 678 specimens from 172 species of ornamental cyprinid fish are now published and freely accessible on BOLD. Of these, 91 species were not previously present in GenBank or BOLD. The majority of the recognised biosecurity risk species were represented, and this will contribute greatly towards building a long term library for ornamental fish biosecurity. DNA barcode data were largely congruent with taxonomy. Issues for specific taxa are discussed where barcodes were ambiguous, and/or conflicted with GenBank data. Using morphological characters the identification of voucher specimens to species was difficult, but this process now provides a tangible benefit to both border security

and future taxonomic or barcoding studies by associating this additional data with the vouchered museum specimens as well as the DNA barcodes, trace files, and other supplementary data.

When the morphological identifications were compared to trade names or names in popular references used by the trade (e.g. Baensch & Fischer, 2007), it is estimated that up to 25% of cyprinid species could be mislabelled. The DNA barcode library generated in this study provides an ideal tool to test this preliminary observation in more detail, and provide a future quantified study of supplier mislabelling in the ornamental industry; this work is currently in progress in association with researchers at the National University of Singapore.

Finally, new methods of presenting barcode data were explored, with Web based methods using URLs to link to corresponding database entries and supporting information providing a vast improvement over traditional ways to represent large trees and share data.

Chapter 3

An evaluation of methods for quantifying identification success in DNA barcoding

3.1 Introduction

As discussed in Chapter 1, not all DNA barcoding studies aim to quantify identification success. An effective biosecurity tool incorporating molecular data such as DNA barcodes relies on making accurate identifications to species level, so explicitly making an assessment of how the data perform in identification scenarios is desirable and necessary. For studies where identification is a possible use of the data generated, then an evaluation of identification success should accompany the standard summary statistics (e.g. Chapter 2). As outlined in Section 1.3, issues such as NUMTs, incomplete lineage sorting and conflicting taxonomy can influence identification success. Here the focus will be upon the analytical methods used, however. Three testable factors with the potential to influence identification success in DNA barcoding studies have been identified. These are: (1) the choice of identification criterion, or analytical method; (2) conflict between datasets, especially where third-party data such as those from GenBank are used; and (3) the effect of singleton species (one specimen per species) in the dataset.

3.1.1 Identification criteria

An overview of the broad categories of methods used to measure identification success was presented in Section 1.4. In order to draw conclusions as to which method(s) is/are best for biosecurity situations, a total of six were chosen to test, and are described below. The most widely used measures of specimen identification were selected, as well as some relatively newer ones. More precise details of how each of the criteria are defined and implemented is presented in Section 3.2.2. It

is important to note that with the exception of the GMYC, all analyses are initially based on genetic distances, using the K2P genetic distance matrix (see Chapter 4).

3.1.1.1 Tree-based monophyly

Firstly, the phylogenetic measure of species monophyly method was tested. Although criticised (see Section 1.4.2), this is a commonly used metric (Casiraghi *et al.*, 2010; Goldstein & DeSalle, 2011), with nearly all barcoding studies reporting some kind of assessment of monophyly, even if just discussing patterns in NJ phenograms. There is an implicit assumption using this method that all species are monophyletic at mtDNA loci, and that identifications can be made by clustering in NJ trees (Meier, 2008). Testing whether the criticisms are valid is an important step. Another common procedure here is to use bootstrap resampling on the NJ phenograms to gauge support for the identifications made using the criterion of monophyly. Recent studies (Zhang *et al.*, 2012) have reported that success rates are low with a bootstrap approach, as it is a conservative measure. Again, however, it is important to make further assessments of this frequently used technique in the context of biosecurity.

3.1.1.2 Distance/threshold methods

The BOLD-IDS identification engine (Ratnasingham & Hebert, 2007) is the main portal for DNA barcode end users to make species level identifications, and therefore possibly the most important assessment in terms of operational usability. Unfortunately the documentation of how BOLD-IDS works is poor, and very little information is provided in its description (Ratnasingham & Hebert, 2007). From what information is known, BOLD-IDS aligns sequences using a hidden Markov model of COI, and carries out a "linear search", probably similar to those that are used to generate standard genetic distances. The method provides an identification if all sequences within 1% of the query are congruent.

Two additional distance based measures were chosen, being the "best close match" (BCM) method of (Meier *et al.*, 2006) and the *k*-nearest neighbour (*k*-NN) approach of Austerlitz *et al.* (2009). Both of these methods are similar, operating on a match of the query to a single sequence in the dataset, although they are different enough to deserve comparison (see Section 3.2.4.1). Austerlitz *et al.* (2009) reported *k*-NN as well performing in their simulated and real data tests, while Virgilio *et al.* (2010) reported BCM as one of the most effective methods among their comparisons.

The BCM and BOLD approaches both rely on a molecular divergence threshold to estimate group membership and guard against providing an identification for a query without a conspecific represented in the database (a false negative, type II error). The use of a universal threshold (e.g. 1%, as used by BOLD), has been questioned repeatedly due to rate variation issues in COI (Section 1.3.3; Hickerson *et al.*, 2006; Meier *et al.*, 2006; Meyer & Paulay, 2005; Rubinoff *et al.*, 2006), and it is clear that no single threshold is likely to suit all species. However, error can be minimised across a dataset for different threshold values (Meyer & Paulay, 2005).

3.1.1.3 General mixed Yule-coalescent (GMYC)

Lastly, a tree-based discrete-data method incorporating an estimation of group membership will be tested: the general mixed Yule-coalescent model (GMYC) of Pons *et al.* (2006) and Monaghan *et al.* (2009). As described in Chapter 1, using an ultrametric phylogenetic tree as input, the GMYC calculates likelihood of species-like clusters based on branching rates over time and incorporating variable coalescent depths. The method has many desirable properties using sophisticated likelihood and coalescent modelling, and has yet to be used for specimen identification purposes in DNA barcoding (Zaldívar-Riverón *et al.*, 2011, used it to estimate biodiversity). This study provides a test to demonstrate the method's potential for biosecurity.

3.1.2 GenBank data

As outlined in Section 2.1.3, GenBank contains a considerable amount of potentially useful information, and can be affected by poorly curated data. The problem of how this may impact identification success in the present study will be addressed by conducting separate analyses for: new data generated in Chapter 2, the GenBank data cited in Chapter 2, and both these datasets combined.

3.1.3 Singletons

A particular challenge to biosecurity is the steady change in the number and identity of species that are traded. Any useful identification method must be robust to these changes; i.e. sequences from new species in the trade should not be erroneously matched to species with barcodes in the database, while a good identification technique should maintain accurate identification of species that are already represented. The extent to which uncommon, singleton specimens affect identification success

rates is rarely explored, and is a problem for DNA barcode identification systems (Lim *et al.*, 2012). As few taxon-specific barcoding projects (i.e. databases) can be considered complete (Lim *et al.*, 2012), the aim here is to examine how the identification criteria are affected by singletons.

It is therefore important for analyses to distinguish between two identification scenarios. First, a query specimen belongs to a species that has already been barcoded and whose DNA barcode is maintained in a DNA barcoding database. Once sequenced, the best identification result for such a specimen is a "correct identification". Second, the query specimen belongs to a species that remains to be barcoded (it is a singleton). The best result here is "no identification", since the specimen has no conspecific barcode match in the database. The best overall identification technique is one that maximises identification success for scenario one, and yields a "no identification" result under scenario two. In light of this, the results with both singleton species included (scenario two) and excluded (scenario one) will be reported. When the analyses are carried out, however, singletons should remain in the datasets as possible matches for non-singletons.

3.1.4 Objectives

The aim of this part of the study is to test how likelihood of identification success—assigning the correct species name to a query barcode sequence—is affected by experimental (sampling of GenBank data, presence of singletons), and analytical factors (identification method). Improved techniques to carry out comparative analyses of identification success for DNA barcode data will be presented, and appropriate ways to address problems arising from these issues will also be discussed. A large part of this work will also be to implement the range of identification methods using a free, open-source software environment.

3.2 Materials and methods

3.2.1 Data collection

The data used to test the suitability of COI barcodes as a species identification tool were those presented in Chapter 2. This included DNA barcodes generated as part of this research, as well as those acquired from GenBank and BOLD. A summary of the

data used is presented in Table 2.1. Including GenBank data, a total of 1,240 COI sequences were used.

3.2.2 Identification methods

Unless otherwise stated, all analyses were conducted using the SPIDER package for R (Brown *et al.*, 2012; Paradis *et al.*, 2004). Many of the functions in this package were written specifically for this part of the study, in an attempt to address the lack of extensible, open-source, and cross-platform software suitable for analysing barcode data. A tutorial of how to conduct these analyses is presented in Brown *et al.* (2012), and also in the online Appendix Section B.5. Three tree-based analyses were used as well as three distance-based measures, and these are described in further detail below.

The protocol used to test each methods was that of simulating a real identification problem for a biosecurity official by treating each individual as an identification query. In effect, this means that each sequence is considered an unknown while the remaining sequences in the dataset constitute the DNA barcoding database that is used for identification. This is referred to as "leave-one-out" by some authors (e.g. Austerlitz *et al.*, 2009; Zhang *et al.*, 2012). Identification rates for these queries were divided into four categories: "correct" or "incorrect", and "no identification" or "ambiguous" if applicable to the method.

3.2.3 Tree based analyses

3.2.3.1 NJ monophyly

A tree-based test of species monophyly was conducted, with this measurement reporting the exclusivity of the genetic clusters in the NJ phenograms. As in Section 2.2.6, a genetic distance matrix and NJ phenogram was generated. The procedure implemented in SPIDER (function: *monophyly*) returns each species as either monophyletic (correct identification), non-monophyletic (incorrect identification) or as a singleton (incorrect identification, as no possible match available). This per-species measure was then scaled to include the number of individuals in each species.

3.2.3.2 NJ bootstrap

A bootstrap test of node support was also incorporated, with correct identifications scored if taxa were monophyletic (as above), and had bootstrap values greater than 70% (Hillis & Bull, 1993). This was carried out using the *monophylyBoot* function of SPIDER; 1,000 replications and codon resample constraints (block = 3 option) were used for the bootstrap analysis.

3.2.3.3 GMYC

For the GMYC analyses, following Monaghan *et al.* (2009), data were first reduced to haplotypes using ALTER (Glez-Peña *et al.*, 2010), with gaps treated as missing data (ambiguous bases were first transformed to gap characters). Next, ultrametric chronograms were generated in BEAST v1.6.1 (Drummond *et al.*, 2006; Drummond & Rambaut, 2007) under the following settings: site models as suggested by the BIC in jModelTest (Guindon & Gascuel, 2003; Posada, 2008); strict molecular clock; $1/x$ Yule tree prior; two independent MCMC chains with random starting topologies; chain length 20 million; total 20,000 trees; burn-in 10%; all other settings and priors default. The GMYC model was fitted in the SPLITS package for R (Monaghan *et al.*, 2009), using the single threshold method under default settings. An individual was scored as a correct identification if it formed a GMYC cluster with at least one other conspecific individual. An incorrect identification was made when an individual clustered with members of other species, and a "no identification" was made when an individual formed a single entity (did not cluster with anything else). Exploratory results (data not shown) suggested that more sophisticated BEAST and GMYC analyses using relaxed clocks, codon partitioned site models, outgroups, and multiple threshold GMYC resulted in a poorer fit to the morphologically identified species names, as did a full dataset (sequences not collapsed into haplotypes).

3.2.4 Distance based analyses

3.2.4.1 *k*-nearest neighbour

The first distance-based analysis comprised the k-nearest neighbour (k-NN) approach, using a K2P distance matrix (Austerlitz *et al.*, 2009). The k-NN analyses was implemented in R, using a script from Austerlitz *et al.* (2009), and provided by Olivier David (a co-author of that article). The method is now implemented in SPIDER with the *nearNeighbour* function. A k-nearest neighbour ($k = 1$) conspecific with the

query returns a correct identification, otherwise an incorrect identification; singletons (where applicable) are reported as an incorrect identification (as no possible match available), and ties were broken by majority, followed by random assignment.

3.2.4.2 Best close match

The "best close match" (BCM) method presented by Meier *et al.* (2006) is provided in the SPIDER function *bestCloseMatch*. BCM is similar to *k*-NN, using a single best match criterion, but matches must be within a pre-specified threshold value (e.g. 1%, but see below) otherwise a no identification result is returned (Meier *et al.*, 2006). In contrast to *k*-NN, ties are reported as ambiguous rather than broken by majority.

3.2.4.3 Approximating BOLD

The third distance technique is one approximating the threshold method used by the BOLD-IDS identification engine (Ratnasingham & Hebert, 2007), and is named *threshID* in SPIDER. It was not possible to actually use BOLD-IDS itself, due to the custom datasets used, and the requirement for the comparisons between methods to be equal. Therefore, when the BOLD method is referred to in this context, it applies to the interpretation used here. BOLD-IDS will return a positive identification if a query shares a $> 99\%$ similar unambiguous match with a reference specimen (Ratnasingham & Hebert, 2007). A correct identification was returned if all matches within 1% of the query were conspecific, an incorrect identification resulted when all matches within the threshold were different species, while an ambiguous identification result was given when multiple species, including the correct species, were present within the threshold. This method is similar to BCM, but operates upon all matches within the threshold, rather than just the nearest neighbour match.

3.2.4.4 Distance threshold revision

A range of threshold percent values were tested for their effect on both the false positive (type I) and false negative (type II) error rates. Categorisation of these error rates follows Meyer & Paulay (2005): "False positives are the identification of spurious novel taxa (splitting) within a species whose intraspecific variation extends deeper than the threshold value; false negatives are inaccurate identification (lumping) within a cluster of taxa whose interspecific divergences are shallower than the proposed value" (p. 2230). The optimum threshold is found where cumulative

errors are minimised. True positives were recorded when only conspecific matches were delivered within the threshold percent of the query. False negatives occurred when more than one species was recorded within the threshold, and a false positive was returned when there were no matches within the threshold value although conspecific species were available in the dataset. This analysis was carried out using the *threshOpt* function in SPIDER. A modification of the BOLD and BCM analyses was incorporated, using the revised threshold values generated during this procedure.

3.2.5 Singletons

To understand the effects of singletons on identification success rates, analyses were carried out as described above; results were reported with and without the singletons. This means that singletons still remained in the datasets as possible matches for non-singletons. This was carried out using the *rmSingletons* function in SPIDER.

3.3 Results

A breakdown of identification success rate for each method and for each dataset used is presented in Table 3.1 and Figure 3.1. When comparing across methods (Table 3.1), success rates for the data generated in this study were generally high ($> 93\%$) when singletons were excluded from the results. The only exception was the NJ bootstrap analysis (89.7%). When GenBank data were added (combined dataset), correct identification rates dropped between 4% and 15% depending on identification technique. If singleton species were included in the results, the reduction in success rate was between 2.7% and 2.9% for the data generated in this study, and 5.2% and 7.4% when GenBank data were combined; when just the GenBank data were considered, success rates decreased between 13.6% and 20.8% depending on the method. When thresholds were optimised, values were reported at 1.4% for the barcodes in this study, and 0.8% when combined with GenBank (Figure 3.2).

The method with the highest proportion of correct identifications with both singletons included and excluded, and across all data partitions, was k-NN. The method with the lowest rate of correct identification for both the data from this study, and the combined dataset, was NJ bootstrap (singletons included and excluded). For the GenBank dataset, the method with the lowest correct identification rate with singletons excluded was the GMYC, and for singletons included, were both the GMYC and BOLD methods (Table 3.1).

Table 3.1. Identification percent success rates for each of the analytical methods across three data partitions (with singletons both included and excluded from results), plus optimum threshold values from cumulative error estimation.

Measure	Singletons	This study (%)	GenBank/BOLD (%)	Combined (%)
NJ mono.	excl.	96.7 (3.3)	83.5 (16.5)	84.7 (15.3)
	incl.	93.8 (6.2)	64.9 (35.1)	78.1 (21.9)
NJ mono. boot.	excl.	**89.7** (10.3)	78.7 (21.3)	**74.7** (25.3)
	incl.	**87.0** (13.0)	61.2 (38.8)	**68.9** (31.1)
k-NN ($k = 1$)	excl.	**98.9** (1.1)	**93.6** (6.4)	**94.8** (5.2)
	incl.	**96.0** (3.9)	**72.8** (27.2)	**87.4** (12.6)
GMYC	excl.	94.2 (3.6, 2.1)	**72.1** (17.3, 10.5)	82.2 (12.5, 5.3)
	incl.	91.4 (3.5, 5.0)	**58.5** (14.1, 27.4)	77.0 (11.7, 11.3)
BOLD: 1% thresh.	excl.	93.2 (0.0, 3.2, 3.6)	75.3 (2.5, 12.8, 9.4)	82.9 (1.5, 6.6, 8.9)
	incl.	90.4 (0.0, 6.0, 3.6)	**58.5** (5.3, 28.8, 7.3)	76.5 (2.8, 12.5, 8.2)
BOLD: opt. thresh.	excl.	93.9 (0.0, 2.4, 3.6)	75.3 (2.5, 12.8, 9.4)	83.4 (1.7, 6.9, 8.0)
	incl.	91.2 (0.0, 5.3, 3.5)	**58.5** (5.3, 28.8, 7.3)	76.9 (2.9, 12.0, 7.3)
BCM: 1% thresh.	excl.	94.8 (0.2, 3.2, 1.8)	77.6 (3.4, 12.8, 6.2)	86.7 (2.4, 6.6, 4.2)
	incl.	92.0 (0.1, 6.0, 1.8)	60.3 (6.0, 28.8, 4.8)	79.9 (3.7, 12.5, 3.9)
BCM: opt. thresh.	excl.	95.6 (0.2, 2.4, 1.8)	77.6 (3.4, 12.8, 6.2)	86.5 (2.4, 6.9, 4.2)
	incl.	92.8 (0.1, 5.3, 1.8)	60.3 (6.0, 28.8, 4.8)	79.8 (3.5, 12.9, 3.9)
Opt. thresh. value		1.4	1.0	0.8

Values in parentheses show failure rate broken down into "misidentification", "no identification" and "ambiguous" (BCM and BOLD only) respectively. "Combined" refers to data generated in this study combined with collected GenBank/BOLD data. Methods with highest and lowest rates of correct identification are presented in bold font. Abbreviations: BCM = best close match; boot. = bootstrap (> 70%); excl. = excluded; incl. = included; mono. = monophyly; opt. = optimum; thresh. = threshold.

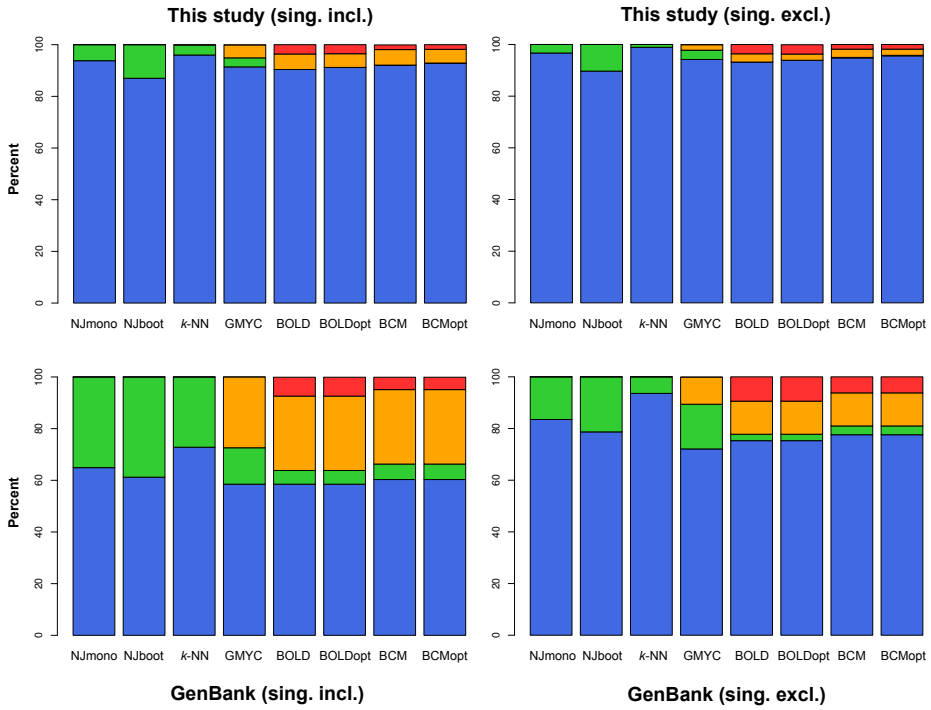

Figure 3.1. Identification success for data derived from this study and downloaded from GenBank/BOLD, with both singletons included and excluded. Key: blue = correct identification; green = misidentification; orange = no identification; red = ambiguous. Abbreviations: NJmono = neighbour-joining monophyly; NJboot = neighbour-joining monophyly with ≥ 70% bootstrap support; *k*-NN = *k* nearest neighbour; GMYC = general mixed Yule coalescent; BOLD = "BOLD method (1% threshold)"; BOLDopt = BOLD method with optimised threshold (Table 3.1); BCM = best close match (1% threshold); BCMopt = best close match with optimised threshold (Table 3.1); sing. excl. = singletons excluded from results; sing. incl. = singletons included in results.

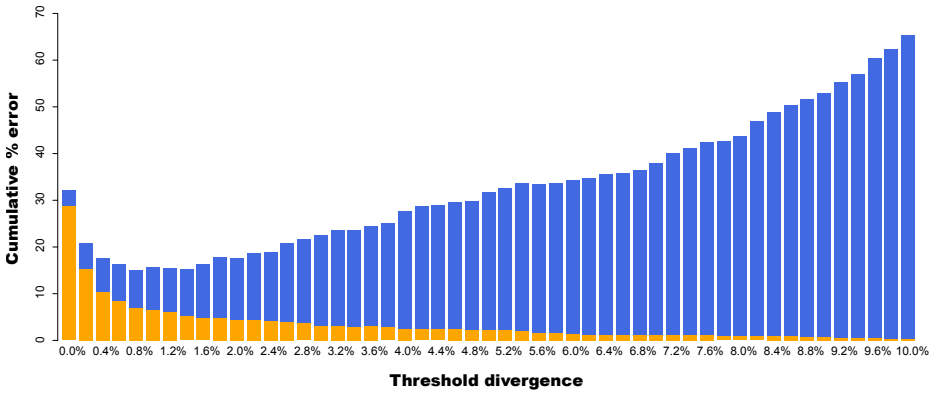

Figure 3.2. Example of cumulative error and threshold optimisation for the combined dataset. False positive (orange) and false negative (blue) identification error rates summed across a range of distance thresholds from 0–10% in 0.2% increments. Definition of errors follows Meyer and Paulay (Meyer & Paulay, 2005). Optimum threshold is 0.8%.

3.4 Discussion

3.4.1 Identification criteria

Many barcoding studies employ terminology describing, for example, species forming "cohesive clusters" differentiated from one another by greater interspecific than intraspecific divergence. This is known as the barcoding gap of Meyer & Paulay (2005). In this study, clustering was measured in terms of monophyly in NJ phenograms, a tree-based method which performed well on data generated here, but suffered when combined with GenBank information. This method requires strict monophyly of each species, resulting in a situation where the inclusion of a single misidentified specimen renders all queries in that species as misidentifications. Although alternative tree-based measures are available (e.g. Ross *et al.*, 2008), the use of NJ trees in general is questionable due their method of construction (Lowenstein *et al.*, 2009; Meier *et al.*, 2006) and topological uncertainty (Meier *et al.*, 2006; Will & Rubinoff, 2004). As discussed already (see Section 1.3.4), for a variety of reasons, "good species" may not always be monophyletic at mtDNA loci, so this method may fail to recognise species with either a history of introgression, or young species with large effective population sizes retaining ancestral polymorphisms (Austerlitz *et al.*, 2009; Elias *et al.*, 2007; Funk & Omland, 2003).

These problems are not resolved through the use of bootstrap values, as a significant reduction (up to 10%) was observed in identification success rate when node support was considered; recently divergent sister species on short branches were often not supported, even if they were monophyletic and diagnosable. DNA barcoding aims to maximise congruence between morphological identifications and sequence information while minimising misdiagnosis. However, this is seriously undermined when bootstrap support values are included. For the reasons stated above, NJ trees are best avoided as a sole identification method (Meier, 2008), although they can be a useful way to visualise and summarise patterns within barcode data. This is discussed further in relation to standard DNA barcoding practices in Section 7.2.3.

The BCM and *k*-NN methods do not require reciprocal monophyly of each species, but merely that the nearest neighbour (single closest match) is conspecific. Thus, even when conflicting GenBank data were included, identification success could still remain high. In cases of a tied closest match, the *k*-NN method ignores this uncertainty and will offer an identification based on majority, while the BCM method reports this as ambiguous. Similarly to NJ, practical difficulties can occur with *k*-NN

when identifying a divergent query from an unsampled species or population, as there is no option for a "no identification". This is a serious problem for undersampled datasets, but the BCM and BOLD methods are able to offer a "no identification" result by incorporating a heuristic measure of species membership (a threshold of 1% distance divergence).

Despite fundamental criticisms of threshold methods (Section 1.3.3), they at least provide an approximate criterion for separating intraspecific from interspecific variation (Meier, 2008). In assessing whether the threshold of 1% best-fitted data generated in this study, the analysis of cumulative error demonstrated that error was variable depending on the dataset. However, it did not grossly depart from BOLD's 1% threshold, perhaps justifying the use of this metric at least in the cases presented here. When the BOLD and BCM methods were modified to employ these revised thresholds, slight improvements were found in the identification success rates.

Using the BOLD method of identification, all matches within the threshold need to belong to conspecifics, rather than the single closest match (as in BCM and *k*-NN). Similarly to NJ monophyly, the BOLD technique is also confounded by even a single misidentified or haplotype sharing specimen in that 1% cluster, and will return an ambiguous result in this situation. This is advantageous when all sources of uncertainty need to be considered, although it can lower the number of successful identifications. As a biosecurity tool, it is worth noting that while the method used by BOLD performed well, identification rates can be improved further by adopting a method such as BCM with a revised, data-derived threshold.

The GMYC incorporates a measure of species membership, but rather than an arbitrary or generalised cut-off, the GMYC employs biological model specification, speciation patterns and coalescent theory in estimating species-like units. As a likelihood based approach, measures of probability and support can be incorporated. Results were highly congruent with the threshold analyses, suggesting the GMYC is picking up the same signal. However, optimising the method for all situations may take prior experience or significant trial and error. Another drawback is that the GMYC is not a particularly user friendly technique, requiring many steps and intensive computation, perhaps precluding its use in some border biosecurity applications where fast identifications may be required (Armstrong & Ball, 2005). Our analysis of 663 haplotypes took approximately five days on a dual processor desktop PC, and although unquantified here, the method also appears sensitive to initial tree-building methodologies.

3.4.2 GenBank data

GenBank certainly offers a formidable resource in terms of taxon coverage and extra information, providing sometimes expert-identified wild-caught specimens with published locality data. However, the absence in many cases of preserved vouchers and justified identifications can undermine the utility of in GenBank data for identification purposes (Harris, 2003; Meier *et al.*, 2006; Ward *et al.*, 2009). BOLD data are certainly better curated, and with higher quality standards, but are also likely to suffer from misidentified specimens to some degree (Meier *et al.*, 2006). Our results do show a decrease in identification success when GenBank data were used, and this was generally due to the higher proportion of singleton species and misidentified specimens, rather than conflicting genetic data *per se*. However, a large proportion of the sequences on BOLD for species in this study remain in private projects and were not available for comparison. Many of these were in fact observed to be conflicting (see Section 7.1 for a discussion of future implications relating to this).

Realistically, as long as the practitioner is aware of alternative explanations for patterns, and is also aware of the relative disadvantages with each analytical technique, there is every reason for incorporating these additional data, especially when a smaller dataset is unable to provide a match. No database is immune to errors, but in this study identifications are transparent, and characters, photographs and preserved vouchers can be scrutinised and corrected at any time via BOLD. Perhaps a two-step approach is required, where GenBank data are consulted if an identification cannot be made using the library generated here.

3.4.3 Singletons

Results were reported with both singleton species included and excluded (Table 3.1). The exclusion of singletons represents a scenario where a barcode database is complete and no new species are to be encountered. However, this is an unrealistic assumption, as the traded cyprinid species come from a much larger pool not currently available in the trade, and the number of singletons in the trade survey shows that it is likely that more singletons will be encountered in the future. These singleton species were usually rare/expensive species, contaminants, or bycatch. When singletons comprised a large proportion of the reference database (such as with the GenBank data), the correct identification rates were significantly reduced for all methods. However, GMYC, BOLD, and BCM were able to discriminate when a specimen could

not be assigned to species. In this respect, the NJ and *k*-NN methods were poorly performing because they are not sensitive to the presence of singletons in a data set; they will always misidentify a query when a match is not available in the database, and this problem may preclude their use until reference databases are complete.

3.5 Summary

This chapter provides an analysis of identification measures. The DNA barcode library generated in Chapter 2 was used to test how different identification methods and sampling strategies influence identification success. The commonly used method based on NJ trees and bootstrap values performed poorly, but alternative and less well known techniques with revised threshold values offered better results (e.g. BCM). The presence of singleton species affected success rates also, and highlighted the need for more complete sampling. GenBank data provided a large number of extra species to fill this gap, although it is not known how accurate the identifications of these specimens are as links to voucher material is often missing (Hanner, 2009; Ratnasingham & Hebert, 2007).

Chapter 4

An evaluation of nucleotide substitution models for specimen identification

4.1 Introduction

As discussed in Section 1.4 and Chapter 3, standard DNA barcoding procedures frequently require genetic distances, and this similarity metric often provides the basis for data summary and specimen identification (Hebert *et al.*, 2003a). Similarity is inferred though pairwise comparison between homologous sequences, and can be expressed as a single value: the number of substitutions per site in a given alignment. These distances are then used in the generation of identification success rates with, for example, nearest-neighbour thresholds or neighbour-joining phylograms. Due to this reliance on distance metrics, a robust and effective estimate of these distances is a prerequisite for non-expert end users of barcode data to have confidence in specimen identifications from public reference databases, such as BOLD (Ratnasingham & Hebert, 2007).

4.1.1 Model choice

In the context of phylogeny estimation, models play an important role in determining our interpretation of evolution. Relationships, branch lengths, and rates over time are all approximated in light of processes assumed by a model (Kelchner & Thomas, 2007), and investigations using simulated and real data have shown that model selection can influence both support values and tree topologies (Buckley & Cunningham, 2002; Cunningham *et al.*, 1998; Lemmon & Moriarty, 2004; Ripplinger & Sullivan, 2008). A model selection procedure aims to identify a model which can best represent mutational processes, while minimising the loss of predictive ability through overparameterisation (Sullivan & Joyce, 2005).

In terms of choosing between models, advances in information theory have allowed for more effective discrimination between competing schemes (Posada & Buckley, 2004). Implementation of information-theoretic approaches such as the Akaike Information Criterion (AIC) now allow for assessment of model fit, as well as taking into account increases in variance by penalising over-parameterisation and information loss (Bos & Posada, 2005; Posada & Buckley, 2004; Sullivan & Joyce, 2005). We are now also able to assess relative support for a given set of substitution models using AIC weights (Posada, 2008; Posada & Buckley, 2004). This approach is particularly useful given that an alternative model may be an equally good estimator as the model with the lowest AIC value (Kelchner & Thomas, 2007). These weights approximate probabilities for a given set of models, and evidence ratios between these weights offer a comparison of support for competing models (Anderson, 2008).

4.1.2 The K2P model

In terms of generating genetic distances, sequence similarity can be derived directly from observed data as raw *p* distances. However, unobserved substitutions at mutational hotspots such as third codon positions can lead to an underestimation of differences between lineages (Sullivan & Joyce, 2005). Mathematical models used in phylogenetics correct for this saturation by applying a more realistic scenario of nucleotide substitution than observed from raw data, and can vary considerably in complexity (Bos & Posada, 2005). In DNA barcoding studies, Kimura's two-parameter model (Kimura, 1980), hereafter referred to as the K2P model, is the *de facto* standard metric for computing these distances (Ward, 2009). The K2P model provides a substitution framework with a free parameter for both transitions and transversions, accounting for the likely higher substitution rate of transitions in mitochondrial DNA (Kimura, 1980; Wakeley, 1996). Base frequencies are assumed to be equal under this model, although departures from this assumption are common in real datasets and different nucleotide compositions may influence particular types of substitution rate (Galtier & Gouy, 1995; Tamura, 1992; Ward *et al.*, 2005).

The use of the K2P model in DNA barcoding began with Hebert *et al.* (2003a), who stated: "For the species level analysis, nucleotide-sequence divergences were calculated using the Kimura-two-parameter (K2P) model, the best metric when distances are low (Nei & Kumar 2000) as in this study" (p. 315). Hebert *et al.* were presumably referring to the following passage in Nei & Kumar (2000): "Even the p distance becomes very similar to other distance measures when $p \leq 0.1$. Therefore

when one is studying closely related sequences, there is no need to use complex distance measures. In this case, it is better to use a simpler one, because it has smaller variance" (p. 40–41; also see p. 112). This point made by Nei & Kumar is important because at a fundamental level, and despite the widespread use of the K2P model in DNA barcoding, it remains to be demonstrated whether model corrected distances are justified over using the uncorrected *p* distances (i.e. can the raw data serve adequately for the purpose required?). Although it has been noted that barcode variation within species is generally low (Hebert *et al.*, 2010; Ward, 2009), it is not clear if simple measures could systematically bias results by underestimating change (Sullivan & Joyce, 2005). In terms of specimen identification, an underestimate of genetic distance may increase the number of false negative "lumping" errors, while overestimating change may increase false positive "splitting" errors (Meyer & Paulay, 2005). This is linked to the principal of the barcoding gap, which relies on individuals within a species being more similar to one another than to the closest individual of another species (Meier *et al.*, 2008; Meyer & Paulay, 2005). It may be that when simple measures such as *p* distances are used, this gap is decreased, hindering identification success. For an effective specimen identification system it is important, therefore, to fully understand how measures of inferred similarity (model corrected distances) or observed similarity (uncorrected distances) could affect results.

4.1.3 Objectives

Two recently published studies have investigated the application of substitution models in DNA barcoding, although they offer fundamentally different conclusions. Fregin *et al.* (2012), based on their analysis of 120 cytochrome *b* sequences from 61 acrocephalid bird species, recommended "Only distances based on the optimal substitution model should be used". In contrast, Srivathsan & Meier (2012) looked at 5,283 published COI sequences from 200 genera, and showed that "the use of uncorrected distances yields higher or similar identification success rates" [compared to K2P correction]. These contradictory findings suggest the question of model specification deserves further attention.

Given the availability of model selection software such as jModelTest (Guindon & Gascuel, 2003; Posada, 2008), it seems an appropriate time to re-examine how sensitive DNA barcode analyses are to alternative models, and ask whether the indiscriminate use of the K2P model is really justified. Using an explicit test of DNA

barcode data under justifiable model selection criteria, this chapter aims to specifically address the following: (1) is the K2P a well fitting model at the species level; (2) how different are distances generated under a better model to those generated under the K2P model; (3) can applying different models change identification success rates and estimations of the barcoding gap; (4) does model correction in general, perform better than using no model; and (5) how did Fregin *et al.* (2012) and Srivathsan & Meier (2012) reach such conflicting conclusions?

4.2 Materials and methods

4.2.1 Data acquisition

Fourteen datasets were obtained in FASTA format from project pages on BOLD. These datasets comprised large studies of relatively well known taxonomic groups including butterflies (Dincă *et al.*, 2011; Hajibabaei *et al.*, 2006a; Lukhtanov *et al.*, 2009), birds (Johnsen *et al.*, 2010; Kerr *et al.*, 2009a,b, 2007), fishes (Hubert *et al.*, 2008; Rasmussen *et al.*, 2009; Steinke *et al.*, 2009a,b; Ward *et al.*, 2005; Wong *et al.*, 2009), and bats (Francis *et al.*, 2010). Well known faunas were chosen to minimise discrepancies between the molecular data and taxonomy. BOLD sequence identifiers (taxon names) were trimmed using regular expressions to include only GenBank accession number and taxonomic identification (species name). Alignment was carried out by BOLD, followed by visual editing using translated amino acids in MEGA4 (Tamura *et al.*, 2007).

4.2.2 Species-level model selection

To test whether the K2P is a well fitting model at the species level, each dataset was split into species using the APE package (Paradis *et al.*, 2004) for R (R Development Core Team, 2010), with species delimited by their unique binomials. The individual species data were exported in NEXUS format, and species with less than five individuals were excluded in order to represent a dataset of at least an average intraspecific sample size (Ward *et al.*, 2009). Using nested UNIX shell scripts, the program jModelTest was run as a batch process for each species in each dataset, producing a corresponding jModelTest output file. All eleven substitution schemes were tested (Posada, 2008), along with base frequency and rate variation options (total 44 models). An invariant sites parameter was not included, as species comprising a

single haplotype could not be optimised under this setting in jModelTest. The model frequencies and AIC weights for the best and K2P models were extracted from the jModelTest output files using shell commands.

4.2.3 Difference between K2P and best model

To test how different intraspecific K2P distances are from best-model distances, firstly batch processes in PAUP* (Swofford, 2003) were used to calculate pairwise comparisons under standard K2P distance settings (distance = K2P). Next, estimations for the best model were generated as maximum likelihood (ML) distances (distance = ml), with likelihood settings derived from jModelTest's PAUP* block output. Shell scripting was used to manipulate corresponding likelihood settings from the jModelTest output into the NEXUS file for each species, before initiating PAUP* as a concatenated batch process. K2P distances were then subtracted from best-model estimates for each pairwise comparison. For this analysis using PAUP*, the pairwise deletion option for missing data was used (missdist = ignore), and undefined distances were set to "NA" (undefined=asterisk); all other settings were default. Except for K2P (= K80), abbreviated nomenclature of models follows Posada (2008).

4.2.4 Identification success

To test the influence of model selection on identification success rate, both intraspecific and interspecific values were required. Distances were generated from the undivided datasets which also included the previously excluded species with less than five individuals. To illustrate the effects of different substitution schemes, a selection of standard "off the shelf" models in PAUP* were used, offering a variety of parameterisations from simple to complex: JC, F81, K2P, TrN, HKY, HKY+Γ and GTR+Γ. Gamma shape values were derived from jModelTest. Identification success rates were measured using the "best close match" (BCM) criterion of Meier *et al.* (2006), and was applied as is described in Section 3.2.4. As highlighted in Chapter 3, the BCM method has several desirable properties, such as being able to make correct identifications for non-monophyletic species, and so was chosen as the appropriate measure of identification to be used in this case. The threshold was initially set at the 1% value, as used by the BOLD identification engine (Ratnasingham & Hebert, 2007). Because threshold values are likely to be contingent upon the models they

are generated under, we also optimised new thresholds for each model and dataset. This optimisation procedure minimises false positive (no matches within x of query) and false negative (more than one species match within x of query) errors for a range of threshold values (0.2%–5.0% in 0.2% increments). To assess the effect of model selection on magnitude of the barcoding gap, both maximum intraspecific and minimum interspecific distances were calculated (Meier *et al.*, 2008), with the barcoding gap expressed as minimum interspecific distance divided by maximum intraspecific distance; singletons were not considered for intraspecific variation, and intraspecific values of zero were replaced with a value of 0.001536098 (corresponding to a single nucleotide change over 651 bp). Analyses were carried out in R using the DNA barcoding package SPIDER (Brown *et al.*, 2012; Paradis *et al.*, 2004).

4.3 Results

4.3.1 Species-level model selection

From the fourteen datasets 1,446 species were extracted with ≥ 5 individuals, resulting in 14,472 DNA barcodes; the mean number of barcodes per species was ten (Table 4.1). For the individual species tested by jModelTest ($n = 1,446$), the model most frequently selected as best (zero AIC Δ value) was the HKY ($n = 579$), followed by F81 ($n = 312$) and TrN ($n = 264$). Overall, twenty models were selected by the AIC, and the K2P model was never selected as best model (Figure 4.1). Models with a gamma shape parameter were selected on 7.95% of occasions. The AIC weight (w) of the best model ranged between 0.08 and 0.64 (mean $w = 0.21$). As an alternative model, the AIC weight for the K2P was no greater than 0.019 (mean $w = 0.000134$). The mean evidence ratio (E) for the best model vs. K2P model weight was $E = 1.9 \times 10^{33}$ (range = 10.0 to 2.8×10^{36}). A representation of the relative model weights is shown in Figure 4.2.

4.3.2 Difference between K2P and best model

In calculating distances within species, a total of 191,402 pairwise comparisons were made. When the K2P distance was subtracted from the best-model distance, 31.2% of the total comprised zero change, and 39.6% were greater than zero and less than 0.1%; 8.12% showed a difference greater than 1%, and 15.6% were negative (K2P distance larger than best-model distance). Average differences were 0.64% (mean)

Table 4.1. Summary and citations for datasets used in this study, with numbers of individuals per species remaining after filtering for ≥ 5 individuals.

Dataset citation	Taxon	No. spp. ≥ 5 indiv.	No. indiv.	Seqs. per sp.
Dincă *et al.* (2011)	Romanian butterflies	144	1,273	8.8
Francis *et al.* (2010)	Southeast Asian bats	88	1,736	19.7
Hajibabaei *et al.* (2006a)	Tropical Lepidoptera	65	723	11.1
Hubert *et al.* (2008)	Canadian freshwater fishes	132	1,203	9.1
Johnsen *et al.* (2010)	Scandinavian birds	31	173	5.6
Kerr *et al.* (2007)	North American birds	230	2,386	10.4
Kerr *et al.* (2009b)	Argentinian birds	106	687	6.5
Kerr *et al.* (2009a)	Palearctic birds	148	1,063	7.2
Lukhtanov *et al.* (2009)	Central Asian butterflies	34	192	5.6
Rasmussen *et al.* (2009)	North American salmonids	8	934	116.8
Steinke *et al.* (2009b)	Ornamental marine fishes	162	1,169	7.2
Steinke *et al.* (2009a)	Pacific Canadian fishes	107	1,029	9.6
Ward *et al.* (2005)	Australian marine fishes	148	921	6.2
Wong *et al.* (2009)	Commercial sharks	43	983	22.9
Total		1,446	14,472	10.0 (avg.)

Abbreviations: avg. = mean; indiv. = individuals; spp./sp. = species; seqs. = sequences.

and 0.00012% (median); range was −0.068% to 136.7%. A density plot illustrating the differences between the K2P model and best-model distances for each dataset is presented in Figure 4.3.

4.3.3 Identification success

A total of 21,514 DNA barcodes were used to measure identification success (including species represented by < 5 individuals). Under the 1% BOLD threshold, differences in identification success for all models varied by no greater than 0.04%; the two models with gamma shape parameters (HKY+Γ and GTR+Γ) had the lowest correct identification rates of 91.81% (Table 4.2). Optimised threshold values varied according to dataset (range 0.2% to 1.2%), although not by model, except for the GTR+Γ threshold for Dincă *et al.* (2011) (Table 4.3). Identification success varied by up to 0.28% under optimised thresholds, with *p* distance having the highest value and the GTR+Γ model with the lowest (Table 4.2). Ambiguous identification tended to decrease with model complexity, along with an increase in incorrect and unidentifiable individuals (Table 4.2). In terms of the distribution of the barcoding

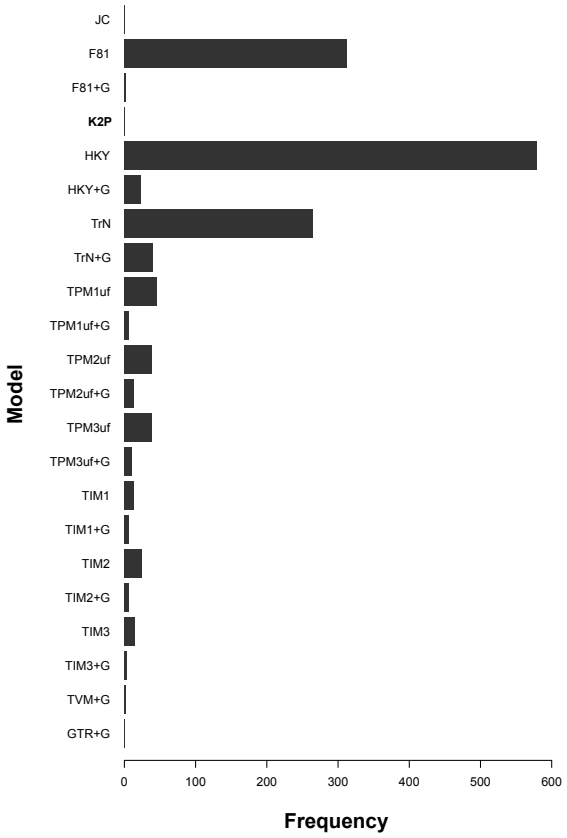

Figure 4.1. Frequency of per-species models selected by jModelTest under the AIC. The K2P model is highlighted in bold (frequency = 0). Except for K2P model, abbreviated nomenclature of models follows Posada (2008). Summary of the properties of these models can also be found in Posada (2008).

gap under different models, for schemes without a gamma parameter, median values remained generally similar with smallest interspecific distances between 12.33× and 13.17× maximum intraspecific distances; the models with a gamma parameter had higher median (16.02× to 16.59×) and also higher maximum values (Figure 4.4). No barcode gap was found for between 8.72% (p distance) and 8.50% (HKY+Γ) of individuals. Overall, the effect of model selection on all distances (both intraspecific and interspecific) is represented in Figure 4.5.

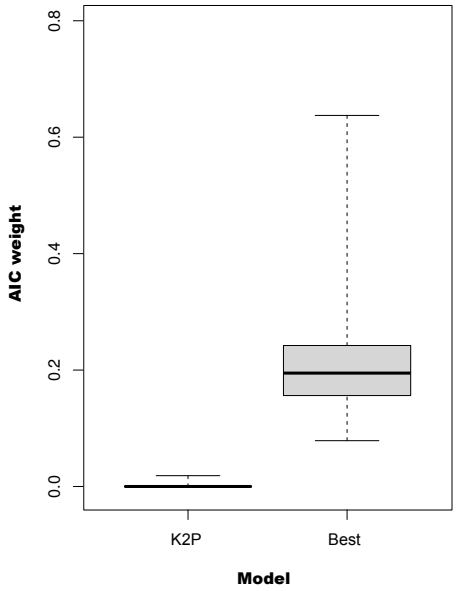

Figure 4.2. Distribution of AIC weights for best and K2P models. Whiskers extend to full range of data; boxes represent quartiles; black lines show median values.

4.4 Discussion

4.4.1 K2P model selection

Although the species level analyses show that the K2P was never selected as the best model, picking a model with the lowest AIC value may ignore credible alternative models that are also good approximators (Alfaro & Huelsenbeck, 2006; Anderson, 2008; Kelchner & Thomas, 2007). Therefore, it could have been possible that the K2P model was a reasonable alternative model. However, when AIC weights and evidence ratios between models were considered to assess support, it was found that the K2P was without exception a poorly approximating model at the species level; the lowest evidence ratio was 10:1 against the K2P. It is likely that the assumption of equal base frequencies led to the rejection of the K2P model in most cases, thus favouring the otherwise similar F81 and HKY models with unequal frequencies (Figure 4.1). In general, substitution schemes tended to be relatively simple at the species level, with either equal rates (F81), or separate transition/transversion rates (HKY) selected. In terms of the suitability of the AIC for answering these questions, other model selection criteria such as likelihood ratio tests or the Bayesian Information Criterion (BIC) could have been considered here, but these measures are considered to be

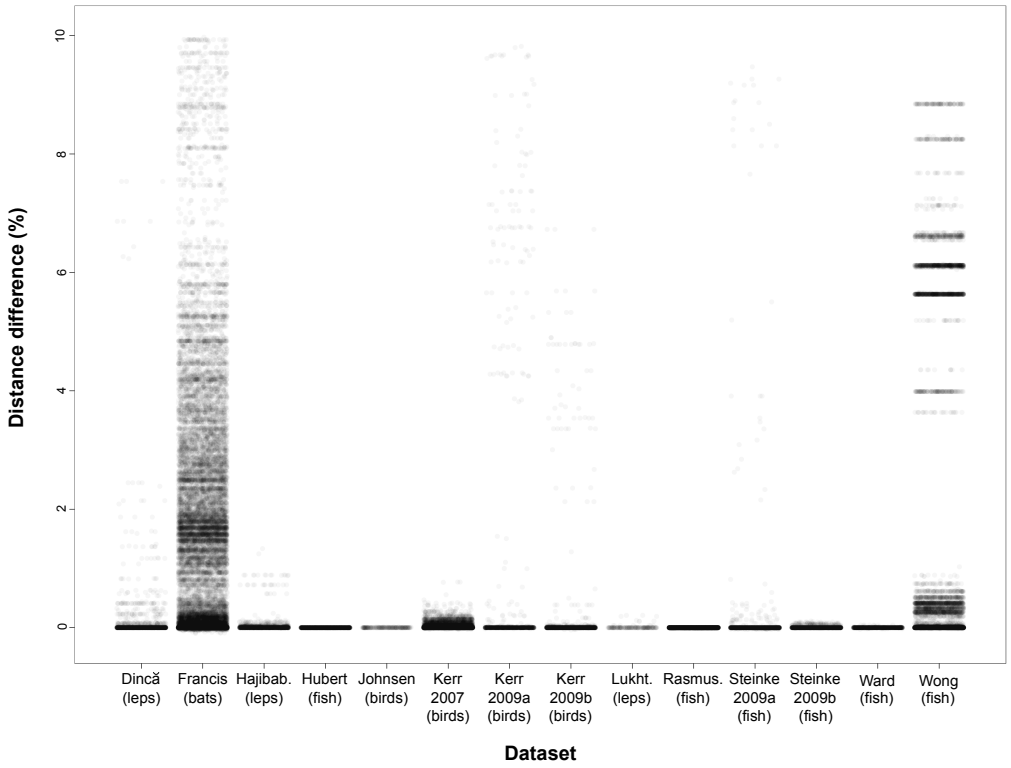

Figure 4.3. Jittered density plot showing percent difference between best AIC model and K2P model distances for each of 14 datasets. The *y*-axis limit was set to 10% to assist presentation. The plot was created in R using ggplot2 (Wickham, 2009).

based upon weak philosophical foundations, and the latter has a tendency to give high weights to poorly fitting models (Anderson, 2008; Posada & Buckley, 2004).

4.4.2 Difference between K2P model and best model

Overall there was little difference between intraspecific distances optimised under best model or K2P model parameters. The majority (86.3%) of the difference was either zero or minor ($< \pm 0.1\%$). The Francis *et al.* (2010) bat dataset had the largest differences (Figure 4.3). When this dataset was excluded, 93.9% of differences in distance were less than $\pm 0.1\%$. At least a third of the bat species analysed in this study had multiple divergences of over 2% K2P distance (Francis *et al.*, 2010). This study group reflects a high proportion of underestimated diversity, and this discrepancy between current taxonomy and DNA data indicates that the species-level units from this study were probably not comparable with the other datasets used. Conversely for the other datasets, species level diversity may have been artificially reduced,

Table 4.2. Identification success rates using the best close match criterion of Meier *et al.* (2006) across a selection of models for $n = 21,514$ individuals. Threshold values were determined from BOLD's 1% (open values), or were optimised according to error minimisation (values in parentheses); refer to Table 4.3 for optimised threshold values.

Dist. measure	Ambig. (%)	Correct (%)	Incorrect (%)	No ident. (%)
p distance	2.35 (2.31)	91.84 (90.81)	0.91 (0.75)	4.90 (6.13)
JC	2.34 (2.31)	91.84 (90.77)	0.91 (0.75)	4.91 (6.17)
F81	2.33 (2.31)	91.85 (90.77)	0.92 (0.75)	4.91 (6.17)
K2P	2.34 (2.31)	91.84 (90.76)	0.91 (0.75)	4.91 (6.18)
TrN	2.30 (2.29)	91.85 (90.76)	0.94 (0.78)	4.91 (6.18)
HKY	2.32 (2.31)	91.85 (90.76)	0.92 (0.76)	4.91 (6.18)
HKY+Γ	2.31 (2.29)	91.81 (90.75)	0.93 (0.77)	4.95 (6.20)
GTR+Γ	2.30 (2.29)	91.81 (90.53)	0.94 (0.77)	4.95 (6.41)

Abbreviations: ambig. = ambiguous; dist. = distance; ident. = identification.

as it was not clear from the methods sections of the publications cited (Table 4.1) whether code numbers or designations such as cf. were appended to species names during the morphological identification process, or were post-hoc assignments based on barcode divergences. As these would be considered different species in the analysis, an indication of how this may have affected results is necessary; of all 14,472 individuals, only 7% failed to satisfy a regular expression conforming to a correctly constructed binomial (‘`[A-Z][a-z]*_[a-z]*`’). However, regardless as to the degree of match between barcodes and taxonomic names, optimising intraspecific distances under a more statistically justifiable model than the K2P did not substantially change them in the majority of cases (Figure 4.3).

4.4.3 Identification success under different models

Although most changes in distance observed among models were small, when strict thresholds are used as identification criteria (e.g. by BOLD), in theory even relatively minor differences in distance could change the assignment of an unknown specimen. However, there was only a negligible decrease in identification success rate when more complex models were employed (Table 4.2), and although the BOLD threshold value of 1% was generated from data under the K2P model, when revised thresholds optimised under different models were provided, the identification success rates continued to remain robust to model selection. This is likely due to the observation

Table 4.3. Optimised distance thresholds for each dataset under a selection of models. Thresholds were optimised for a range of values (0.2% to 5.0%) under a procedure that minimises false positive and false negative error rates (Meyer & Paulay, 2005). The threshold varying by model is highlighted in bold.

Dataset	*p* dist. (%)	JC (%)	F81 (%)	K2P (%)	TrN (%)	HKY (%)	HKY+Γ (%)	GTR+Γ (%)
Dincă *et al.* (2011)	0.4	0.4	0.4	0.4	0.4	0.4	0.4	**0.2**
Francis *et al.* (2010)	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
Hajibabaei *et al.* (2006a)	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Hubert *et al.* (2008)	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
Johnsen *et al.* (2010)	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Kerr *et al.* (2007)	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Kerr *et al.* (2009b)	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Kerr *et al.* (2009a)	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Lukhtanov *et al.* (2009)	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
Rasmussen *et al.* (2009)	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
Steinke *et al.* (2009b)	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
Steinke *et al.* (2009a)	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Ward *et al.* (2005)	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Wong *et al.* (2009)	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Mean	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6

Abbreviations: dist. = distance.

that distance values pertinent to specimen identification (i.e. largest intraspecific and smallest interspecific), were generally low enough not to be significantly affected by model correction (Figure 4.3, Figure 4.5). Overall, genetic distances generated under models without a gamma shape parameter scarcely deviated from estimations made by the K2P model at *p* distances of $< 10\%$, although when a gamma shape parameter was introduced distances had an increased proportion of correction at this level (Figure 4.3). As an indication of how correction may influence a typical dataset, Ward (2009) reported mean interspecific K2P distances of 5.5% for congeneric bird species, while these results for a wider variety of taxa (Table 4.1) report a mean K2P distance of 6.9% for all nearest non-conspecific values, and a mean maximum intraspecific value of 1.0%.

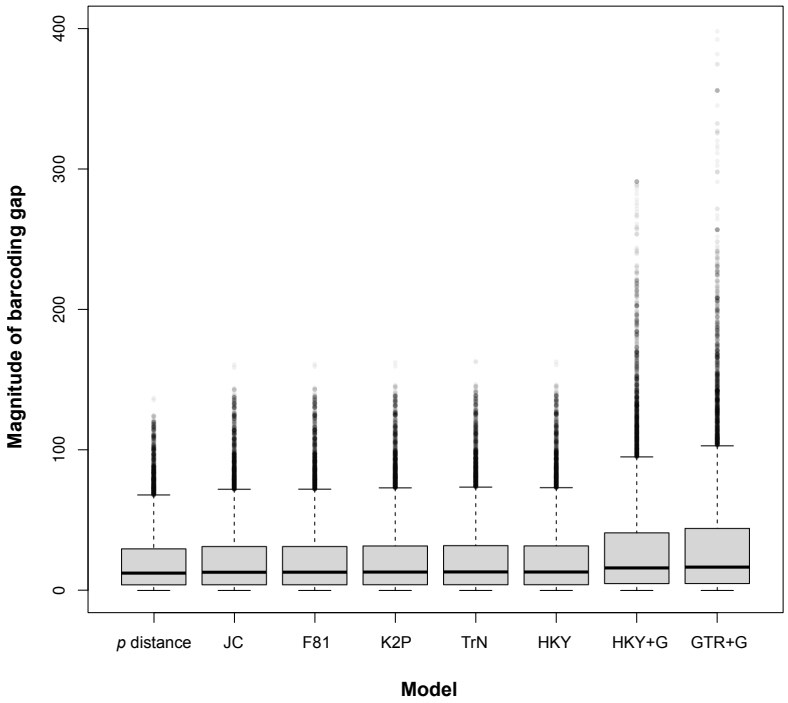

Figure 4.4. Distribution of variation in the magnitude of the barcoding gap according to model for $n = 20,643$ individuals. The barcoding gap is expressed as interspecific divergence as a multiple of intraspecific divergence, and was calculated by dividing each minimum interspecific value by the corresponding maximum intraspecific value. Singletons were not considered for intraspecific variation. Whiskers extend to 1.5× interquartile range, black lines show median values, and points represent outlying data.

4.4.4 Discrepancies between previous studies

Regarding the discrepancy between conclusions presented by Fregin *et al.* (2012) and Srivathsan & Meier (2012), the results of this study were found to be entirely congruent with those of Srivathsan & Meier (2012), in that substitution models have little effect on specimen identification. This study found a slight degree of systematic bias, with more complex models having marginally lower ambiguous identification error rates (interspecific distances underestimated), although this was countered by a larger proportion of incorrect and unidentifiable specimens (intraspecific distances overestimated). When taking this bias into account, the results shown here demonstrate that for identification purposes, *p* distances perform as well, or marginally better (optimised thresholds), than more complex models due to the higher false positive error rates of the latter (Table 4.2). Similarly, increasing model complexity produced an increase in the magnitude of the barcoding gap (Figure 4.4). However, this was not translated into an increase in the number of

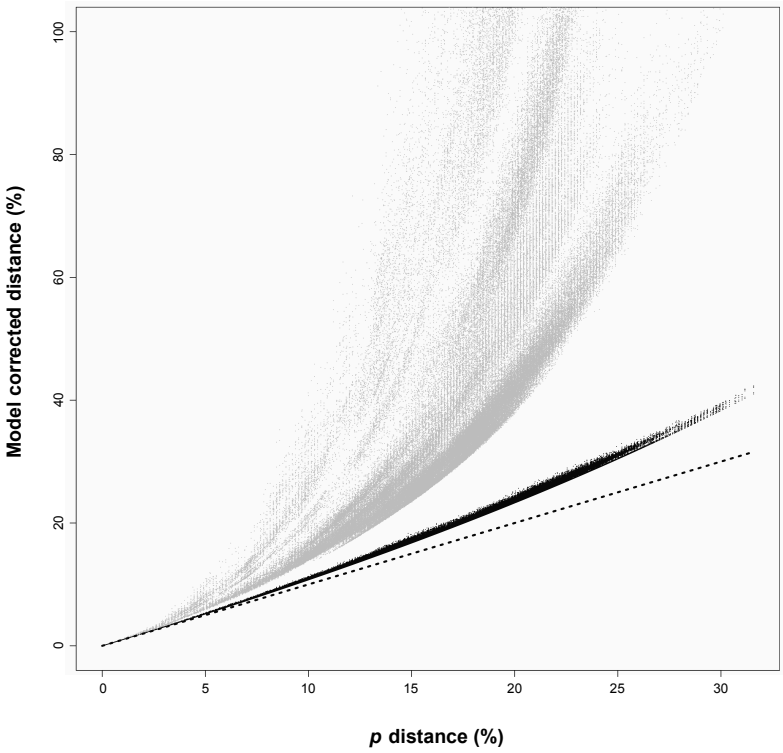

Figure 4.5. Scatter plot of a representative random sample ($n = 100,000$) of intraspecific and interspecific distances as a function of increasing p distance. Models with a gamma shape parameter (HKY+Γ and GTR+Γ) are shown by grey points, p distance by the dotted line, and distances derived under the JC, F81, K2P, TrN and HKY models by black points.

individuals for which a gap was present. Increasing parameterisation further, with the inclusion of an invariant sites model (GTR+I+Γ), resulted in another increase in the magnitude of the barcoding gap, and again generated a reduction in identification success (data not shown). Given the assertion of Nei & Kumar (2000) that "when one is studying closely related sequences, there is no need to use complex distance measures", it should be asked again why models are used in DNA barcoding? Thus, it appears that observed similarity is an acceptable way to identify specimens, unless a user is particularly interested in minimising one error rate over another for a specific application.

Despite their call for better fitting models to be used in studies using genetic distances, a reanalysis of the data presented by Fregin *et al.* (2012) showed no differences according to model in either identification success rate or proportion of specimens lacking a barcode gap (TrN+Γ and p distances; their Supplementary Table 1). It is not clear to whom their advice is aimed, because their conclusions appear to blur the distinctions between specimen identification and species discovery—

assigning unknowns to a pre-identified reference library vs. species delimitation and description (Padial *et al.*, 2010; Vogler & Monaghan, 2007). Although the same data can be used for both purposes, the objectives remain fundamentally different and each require distinct experimental procedures (Padial *et al.*, 2010, and also see Section 7.2 for further discussion). There appears to be no standard practice regarding model correction for taxonomic questions, and different substitution frameworks are often employed among studies, frequently without a model selection procedure or justification (for references see Fregin *et al.*, 2012). When making taxonomic decisions, understanding evolutionary process is arguably more important than for DNA barcoding, and may be especially critical in circumstances such as supporting a new species status for a divergent taxon. When framed in this context, a greater emphasis on model choice must indeed be recommended, which is therefore in agreement with the conclusions of Fregin *et al.* (2012).

4.5 Summary

In conclusion, model selection should remain an important consideration in many disciplines, and DNA barcoding should be no different. Practitioners of DNA barcoding may feel reassured that identification rates were not significantly affected by model selection. However, they should also be aware that a model selection process can increasingly influence conclusions when larger distances are being considered. In taxonomic studies where these conclusions are important, statistical uncertainty in distance estimation could certainly be better explored with information-theoretic techniques such as multi-model inference and model averaging.

Chapter 5

An evaluation of nuclear genetic information in detecting interspecific hybrids and assessing cryptic species

5.1 Introduction

One of the aims of DNA barcoding is to provide a universal system of identification, using a standardised mitochondrial DNA reference system (Hebert *et al.*, 2003a). It has been pointed out that there are situations where mitochondrial DNA may be inappropriate or may lack properties desirable to make suitable biological inferences (Section 1.3). In particular, these are the detection of interspecific hybrids (Aliabadian *et al.*, 2009; Dasmahapatra & Mallet, 2006), and the accurate delimitation of morphologically cryptic lineages among species (Dasmahapatra *et al.*, 2010). The use of nuclear genetic information is in theory able to address these problems. Nuclear loci are increasingly used to validate mitochondrial results and also provide an independent, additional source of data for use in identification, systematic, or taxonomic studies (Vogler & Monaghan, 2007). In the case of aquarium fishes, a nuclear marker may also offer advantages in detecting natural introgression patterns, or interspecific hybridisation events that may have occurred during indiscriminate or deliberate breeding at ornamental fish farms.

5.1.1 Interspecific hybrids

As outlined in Section 1.3.4.3, introgression has been shown to be a relatively frequently occurring phenomenon in wild populations of animals. However, in the case of ornamental fishes, identifying captive bred and mass-produced domesticated organisms presents unique problems for both morphological and molecular identification procedures. Loss of diagnostic phenotypic/genotypic characters may occur in ornamental fishes due to the processes of artificial selection and interspecific

hybridisation for retail purposes. Interspecific hybrid organisms may be of biosecurity concern (either or both of the parental species), and specimens of mixed genealogy may be unpredictable in both phenotype and genotype (Mallet, 2005), making them additionally challenging to identify. Interspecific hybrids have long been used in aquaculture to transfer desirable traits such as increased growth rate or environmental tolerances (Bartley *et al.*, 2001). As hormone breeding technologies become more accessible to breeders, the aquarium industry is now producing increasing numbers of novel hybrid organisms for the trade such as loaches and *Synodontis* catfishes (Clarke, 2008; Ng, 2010). These hybrids may be selected for aesthetic reasons, growth rate, or even to be fraudulently passed off as species with a high market value (Ng, 2010). There is also the possibility of accidental, non-deliberate breeding of hybrids at farms.

5.1.1.1 Identifying hybrids with mtDNA

Due to their frequently intermediate phenotypes, hybrids can be difficult to identify using morphological characters. However, DNA barcoding is well suited to identifying specimens with an atypical phenotype created by artificial selection. However, matrilineal inheritance of mtDNA means any hybrid "unknown" will be incorrectly identified as the maternal species only, ignoring its history of introgression (Avise, 2001). Therefore, hybrid consignments may be inadvertently granted access into New Zealand and other countries based upon positive barcode identification of the maternal species. Valuable information could be lost by using the standard COI approach alone, and misleading conclusions could be reached regarding the identification of query specimens. This may have implications for biosecurity risk assessments, with life history data and nomenclature becoming associated with the maternal species only. Hybrids could also have important biological traits (e.g. temperature tolerances or pathogen resistance) associated with one, both, or neither of the parent species (Reyer, 2008; Seehausen, 2004). Testing hypotheses of hybridisation in the ornamental fish trade could quantify the margins of error when making identifications in hybrid-risk groups.

5.1.1.2 Identifying hybrids with allozymes

The use of nuclear allozyme loci was popular in early studies employing molecular techniques for detecting and understanding hybrid organisms using heritable genetic markers (e.g. Avise & Saunders, 1984). Allozymes are different alleles of the same

enzyme, coded at the same locus. Differing biochemical properties of the protein molecules allow the discrimination and genotying of interspecific variation via a gel electrophoretic assay (Alarcón & Alvarez, 1999; Scribner *et al.*, 2001). The method is both cost effective and fast (van der Bank *et al.*, 2001). However, it requires knowledge and/or fresh tissue samples of both the potential parental species to be effective in detecting a hybrid organism in a biosecurity situation, something which is not always feasible due to the sporadic availability of many species in the trade.

5.1.1.3 Identifying hybrids with microsatellites

Most studies of naturally occurring introgression use allele frequency data from microsatellite markers (Sanz *et al.*, 2009), and this can be combined with mitochondrial or other organellar DNA (Aliabadian *et al.*, 2009; Avise, 2001). For a rough estimate of hybridisation (i.e. F_1), Boecklen & Howard (1997) recommend 4–5 markers, while significantly more complicated situations of advanced backcrossing require over 70. Vähä & Primmer (2006) recommend similar numbers, with 12–24 for F_1, and > 48 for detecting backcrossing. Generating and testing protocols for this number of markers takes significant time and effort, and importantly, they need to be generated specifically for each taxon. Despite offering fine-scale information, this type of method cannot be applied universally to any species in the way that DNA barcoding can, so therefore the use of microsatellite markers is limited for biosecurity applications.

5.1.1.4 Identifying hybrids with nDNA sequence data

Nuclear sequence data can be used in a phylogenetic context to identify hybrids, as there will be incongruence between gene trees (Sota & Vogler, 2001). Unfortunately, this requires nuclear and mitochondrial sequence data from both parental species. However, hybrid individuals will frequently have higher levels of heterozygosity than non-hybrids (Sonnenberg *et al.*, 2007), as diploid organisms will carry divergent copies of the same gene from each parent on separate chromosomes. Therefore, a stand-alone test for hybridisation would simply require an nDNA sequence from a single gene to flag the possibility of a hybrid by way of level of heterozygosity, which could then be investigated with other means. Although hybrids between recently diverged sister species would be difficult to detect with this method, reports suggest that in order to create new and “interesting” varieties for sale (Ng, 2010), many of the aquarium hybrids are produced from phylogenetically quite distinct parentage

(sometimes different genera or families). Therefore, cases such as these would be likely to show high levels of heterozygosity.

5.1.2 Cryptic and unrecognised diversity

5.1.2.1 Definitions

Cryptic species are defined as "two or more distinct species that are erroneously classified (and hidden) under one species name" (Bickford *et al.*, 2007). They are thought to be widespread throughout metazoan taxa, and across biogeographic realms (Hebert *et al.*, 2004; Lohman *et al.*, 2010; Pfenninger & Schwenk, 2007). The classification of multiple species as a single species, is usually due to a lack of morphological distinction as reported in the taxonomic description. Some cryptic species are truly morphologically cryptic—at least as far as the currently employed morphological methods allow us to investigate—and can only be detected with genetic data. However, others may have morphological differences which become apparent when the characters are reassessed (Smith *et al.*, 2007); here these are termed "pseudocryptic species". Another scenario is where a taxon is already recognised as being different (usually with morphological data), and simply remains undescribed; this is termed "unrecognised diversity".

Morphological similarity can persist for long periods of time, with tens of millions of years of morphological stasis having been documented in the African osteoglossomorph fish *Pantodon* (Lavoué *et al.*, 2010). Also, in insects, many previously assumed generalist species are actually a complex of host specifics (Smith *et al.*, 2006). The important crop pest *Bemesia tabaci*, for example, is thought to comprise a complex of genetically distinct, but morphologically conservative lineages (Boykin *et al.*, 2012).

5.1.2.2 Cryptic species, biosecurity and DNA barcoding

The presence of cryptic species, or species complexes with poorly resolved taxonomy can be a problem for identification, as a seemingly well-sampled barcode library may be lacking important reference specimens from these lineages; estimating sampling breadth using taxonomic names may be an underestimate of the underlying mtDNA diversity. When no reference material exists, the presence of cryptic species can therefore increase the potential for unknowns to fail to be identified by a DNA barcode library. When only a single taxonomic name is given to a species complex, it also raises problems for biosecurity management (Boykin *et al.*, 2012). The boundaries

for evolutionary significant units (ESUs) within a species complex may be fuzzy, and intra-group misidentifications may be common. Therefore, because some of these units can have a higher biosecurity risk than others, it is essential to be able to effectively reference these to ensure information is consistent on databases and between biosecurity organisations.

5.1.2.3 DNA barcoding and species concepts

Given the focus of the thesis on the taxonomic rank of species as a basis for correct identification, it seems appropriate to briefly discuss species concepts with reference to DNA barcoding and cryptic species. As stated by Schindel & Miller (2005), there are two distinct aims of DNA barcoding: specimen identification, and species discovery (this dichotomy is discussed in greater detail in Section 7.2). In terms of both aims, DNA barcoding[1] can be considered independent of the "problem" of species concepts (for a review of species concepts, see de Queiroz, 2007). DNA barcoding for specimen identification relies upon matching genetic data to *a priori* described taxonomic names via the generation of a reference library of associated voucher material, pre-identified using morphological characters. Consequently, the problem of species concepts and delimitation is addressed by the original taxonomic description of the species. In this context, DNA barcoding is simply concerned with techniques maximising the congruence between the predefined names and the DNA data (Chapter 3).

In situations where "species" are not associated with names—they are part of an undocumented fauna or cryptic complex of species—DNA barcoding can play a part in initially recognising and documenting these lineages. In this respect, the application of DNA barcoding is as a "species discovery" or biodiversity triage tool (Schindel & Miller, 2005). This process can offer information about population structure, speciation events and potential conservation status (Francis *et al.*, 2010), and is therefore useful for rapid biodiversity assessments as well as for ecological or biosecurity applications (Boykin *et al.*, 2012).

Species delimitation methods such as the general mixed Yule coalescent (Monaghan *et al.*, 2009) and the Automatic Barcode Gap Discovery tool (Puillandre *et al.*, 2012), are able to assess species diversity directly from molecular data, and independently of prior taxonomic knowledge. It is important to note, however, that in the context of species discovery, divergent mtDNA groups derived from methods

[1]Note the emphasis on DNA "barcoding" rather than DNA "barcodes".

such as these, or even just a simple monophyletic group above an arbitrary percent divergence threshold, do not require a concept of species either (although this is perhaps arguable). The methods operate by detecting biological pattern, consistent with theoretical expectations and broad empirical observations across multiple, previously defined species from independent studies. In other words, they report species-like groups using heuristic methods, which are typical of expectations as observed from other data. In COI, for example, if intraspecific variation greater than 3% is rare in well circumscribed taxa, then this level of divergence could be more consistent with interspecific variation. This is not however, a formal species hypothesis in a taxonomic sense, although the same underlying data could be used as a next stage in forming part of an integrated taxonomic process (Padial *et al.*, 2010). It is important to note here that basing taxon descriptions on molecular data, and in particular using statistical species-delimitation methods can be difficult, unless also framed in the context of diagnostic characters consistent with relevant nomenclatural codes (Bauer *et al.*, 2011; Lowenstein *et al.*, 2009).

Confusion can also arise between the form of molecular parataxonomy as described above, and with formal DNA taxonomy (cf. Tautz *et al.*, 2003), which is more explicit in promoting a central rather than auxiliary role for DNA in descriptive taxonomic practice (Vogler & Monaghan, 2007). In this respect, DNA taxonomy certainly requires a species concept, or in reference to de Queiroz (2007), an operational criterion for a species hypothesis.

5.1.2.4 Detecting cryptic species

For some applications such as community ecology, crude measures of biodiversity from mtDNA may be all that are required (Valentini *et al.*, 2009). However, for more rigorous applications, heuristic hypotheses from DNA barcoding methods may need to be tested with further data (Smith *et al.*, 2007). Therefore, some authors have questioned the validity of putative cryptic taxa as reported by divergences in mtDNA analyses (Brower, 2006; Dasmahapatra *et al.*, 2010; Dasmahapatra & Mallet, 2006; Elias *et al.*, 2007); they insist that COI is insufficient to robustly recognise a biparental lineage, and that candidate species be additionally supported with independent datasets, thus increasing the degree of corroborative evidence.

With the tendency of DNA barcoding studies to discover putatively cryptic taxa (Zemlak *et al.*, 2009), it is likely that previously unrecognised lineages or candidate species are uncovered in this study. Nuclear markers are an important tool in this

process and can assist in the critical assessment of these lineage divergences, with concordant patterns from both genomes adding extra support to hypotheses of speciation within morphologically constrained lineages.

Biosecurity decisions are better informed with a good knowledge of the molecular diversity (Boykin *et al.*, 2012). The purpose in this chapter is to assess how valuable nuclear gene information can be in supporting relationships within putatively cryptic species, and for investigating unrecognised diversity in general (undescribed, but morphologically distinct species).

5.1.3 Nuclear marker selection

The *a priori* choice of an appropriate nuclear marker is difficult. The nuclear genes sequenced for fishes tend to be those used for phylogenetic studies, and as a result are more directed toward resolving relationships at a deeper level than those between closely related species (e.g. Li *et al.*, 2007). Phylogeographic studies, on the other hand, investigate a more appropriate evolutionary level and could be a better source of loci. Historically, most have used mtDNA and microsatellites (Zink & Barrowclough, 2008). Nuclear sequence data are becoming increasingly employed in phylogeography (Edwards & Bensch, 2009; Hare, 2001). However, few genes have been identified so far as suitable in fishes, and *de novo* generation of potential loci is complicated and time consuming (Lee & Edwards, 2008). Fortunately, nuclear-gene DNA barcoding has to some degree been investigated; Sevilla *et al.* (2007) assessed nuclear rhodopsin (RHO/Rhod/RH1/RH)—a marker having been observed to show variation at the species level for molecular systematic questions (Fang *et al.*, 2009)—and incorporated it into their multi-locus fish identification tool, while Sonnenberg *et al.* (2007) used the D1–D2 region of LSU 28S rRNA to distinguish closely related fish species.

5.1.4 Objectives

Here, the aim is to answer two different problems associated with DNA barcoding—detection of interspecific hybrids and cryptic species—with the use of the same tool: DNA sequence data from nuclear loci. A range of potential nuclear markers will be assessed for suitability, and then nuclear barcodes will be generated from a suitable candidate to test how they compare to COI barcodes in detecting species level variation for the same taxa. One of these nuclear markers will then be used to firstly

identify hybrid aquarium species both independently using sequence heterozygosity, and in conjunction with COI data. Secondly, patterns of putatively cryptic speciation or unrecognised diversity will be investigated with nDNA to assess support for hypotheses raised from the COI data.

5.2 Materials and methods

5.2.1 Nuclear marker selection

A three-step screening procedure was used to identify potentially useful genes, and is outlined as follows.

5.2.1.1 Genomic screening

Firstly, a broad range of candidate nuclear loci was selected by reviewing recently published phylogenies of fishes, or studies looking specifically at marker development or specimen identification. Due to the wide range of taxa that have been studied, it was not possible to make a universal comparison across genes using GenBank data from these studies. Instead, the Ensembl Genome Browser (http://www.ensembl.org/) was searched for each gene using the *Danio rerio* database. Orthologous gene sequences were then downloaded for the other four model teleost fishes (*Gasterosteus aculeatus*, *Oryzias latipes*, *Takifugu rubripes* and *Tetraodon nigroviridis*). This protocol allowed a crude screening of the more variable loci across a large part of the Acanthopterygii and Ostariophysi, with the assumption being that genes variable across different orders of fishes may correspond to show variability at the species level, and therefore warrant further investigation. To estimate diversity, pairwise *p* distances were calculated for each gene using MEGA4 (Tamura *et al.*, 2007).

5.2.1.2 Intrageneric diversity

Next, a subset of five genes was selected to be tested empirically for intrageneric diversity (using uncorrected *p* distances as above) on a selection of *Danio* species (*D. rerio*, *D.* aff. *kyathit*, *D. kyathit*, *D. dangila*, *D. albolineatus* and *D. margaritatus*). For promising loci that did not have published or working primers, new primers were designed from the Ensembl alignments using PRIMER3 with the default settings (Rozen & Skaletsky, 2000).

5.2.1.3 Comparison with COI

Finally, a single marker was selected for testing across a wider range of species within the Cyprinidae, and to be compared to information from the COI barcode region (as generated in Chapter 2). A subset of 200 individuals was amplified for both markers, comprising 82 species (1–10 individuals per species). Barbs (*Puntius*) and danios (Danionini) were targeted, along with other taxa showing putative interspecific COI divergences. Patterns in agreement between matched nuclear and COI subsets were investigated using the NJ monophyly and *k*-NN methods (as presented in Chapter 3).

5.2.2 PCR protocols for nuclear genes

Nuclear data for the five shortlisted genes (Table 5.2) were generated with the following lab protocol. DNA extractions were as outlined in Section 2.2.4.1. Optimised PCR reactions were carried out using a GeneAmp 9700 thermocycler (APPLIED BIOSYSTEMS) in 10 µl reactions of:[2] 1.7 µl ultrapure water; 1.0 µl Expand High Fidelity 10× PCR buffer (ROCHE DIAGNOSTICS); 2.0 µl Q-Solution (QIAGEN); 0.2 µl $MgCl_2$ (25.0 mM); 2.0 µl dNTPs (1.0 mM); 1.0 µl forward and reverse primer (2.0 µM); 1.0 µl DNA template; 0.1 µl Expand High Fidelity polymerase (ROCHE DIAGNOSTICS). Thermocycler settings for amplification were as follows: 4 min at 94.0°C; 40 cycles of 20 s at 94.0°C, 30 s at 52.0–56.0°C and 60 s at 72.0°C; 7 min at 72.0°C; ∞ at 4.0°C. Primer pairs used are given in Table 5.2. Sequencing protocol was as for the COI data presented in Section 2.2.4.2.

5.2.3 Breeding interspecific hybrids

To compare how effectively sequence data can identify introgression, experimental hybrids were bred in the laboratory under natural aquarium conditions. Two species (*Danio rerio* and *D*. aff. *kyathit*) were selected as candidates for hybridisation as they are similar in appearance, relatively closely related (Fang *et al.*, 2009; Tang *et al.*, 2010), easy to breed (Cottle, 2010), and readily available in the pet trade. *Danio rerio* was chosen as the maternal species. Breeding procedures followed Cottle (2010), and comprised keeping males and females in separate tanks for conditioning (until females were gravid), followed by adding a single female and male into an empty tank in the evening. The spawning tank was decorated with Java moss (*Taxiphyllum*

[2]Final concentrations of reagents are as follows: 1× buffer; 2.0 mM MgCl2 ; 0.2 mM dNTPs; 0.2 µM per primer; 0.35 U polymerase.

barbieri), and fitted with an air powered box filter, and importantly, a raised wire mesh across the base to prevent adults eating the eggs after spawning (aquarium set-up is detailed further in Section 6.2.1). The following morning the tank was checked and if spawning was successful, the adults were removed along with the mesh. Fry were fed on liquidised propriety flake food and microworms (*Panagrellus redivivus*). Permission to carry out the hybridisation experiment was approved by Lincoln University Animal Ethics Committee (code #294; May 29, 2009).

5.2.4 Detecting hybrids

5.2.4.1 Heterozygosity

The proportion of heterozygosity in an individual may indicate recent hybridisation (Sonnenberg *et al.*, 2007). The aim here was to investigate the amount of heterozygosity present in the lab bred hybrid compared to that of the putative non-hybrid cyprinid fishes collected as part of this study, and from fishes more generally. When assessing heterozygosity in the data generated in this study, the polymorphic positions were scored by visually assessing each chromatogram following Sonnenberg *et al.* (2007). Double peaks should be present in both forward and reverse chromatograms, and with a secondary peak height of at least $^1/_3$ of total peak height.

To assess the level of heterozygosity of putative non-hybrids in an overall sample, GenBank was searched on the 28th July 2011 for all rhodopsin (RHO) sequences from teleost fishes using the term "Teleostei AND (rhodopsin Rhod gene)". A total of 1,530 sequences were downloaded. Ambiguous sites were inferred from the sequence data using the standard IUPAC ambiguity code (Cornish-Bowden, 1985), and counted in R using *grep* and the *seqStat* command of SPIDER (Brown *et al.*, 2012; Paradis *et al.*, 2004). The "N" code (all bases) was excluded.

5.2.4.2 Identifying parental species

To test if nuclear sequences can be used to identify both parent species of a hybrid, a composite nuclear DNA sequence was generated *in silico*. The COI data was used to reveal the maternal species, so a putative paternal nuclear sequence can be calculated by resolving the ambiguities in the hybrid sequence using the information from the maternal species' nuclear sequence. For example, at a given position, if the maternal species (as identified by COI) has a cytosine (C), and the hybrid has a Y (C or T), then the putative paternal sequence was scored as a thymine (T). If ambiguities were

also present in the maternal nuclear sequence, these remained as ambiguous in the composite sequence. The composite paternal sequence was then identified against the nuclear RHO reference library using the BCM method of identification (see Section 3.2.4.2); the threshold was optimised for the RHO data using the *threshOpt* function of SPIDER. This method was tested with both the lab bred *Danio* hybrids and a putative hybrid *Puntius* purchased in the aquarium trade (RC0171).

In addition to the hybrid *Puntius*, tissues were available from both museum specimens and the ornamental trade for some putative hybrid catfishes, identified as such morphologically; this included a clariid catfish (RC0739; BMNH:2008.9.17.1-2), a pimelodid catfish (RC0374), and 16 mochokid catfishes (*Synodontis* spp.). To make a maternal identification, mitochondrial DNA was used, but few COI data were available for these groups in BOLD or GenBank. Instead, as cytochrome *b* data were available for a large number of species, the specimens here were sequenced for the mitochondrial cytochrome *b* gene using the primers Glu-2 and Pro-R1 (Hardman & Page, 2003). PCR was carried out with a Veriti thermocycler (APPLIED BIOSYSTEMS) in 10 μl reactions with the following reagents: 1.0 μl ultrapure water; 5.0 μl GoTaq Green Master Mix (PROMEGA); 1.5 μl forward and reverse primer (2.0 μM)[3]; and 1.0 μl DNA template. Thermocycler settings comprised: 2 min at 94.0°C; 40 cycles of 20 s at 94.0°C, 30 s at 60°C and 60 s at 72.0°C; 7 min at 72.0°C; ∞ at 4.0°C. The hybrids were also sequenced for RHO using methods outlined previously, to detect polymorphisms.

5.2.5 Cryptic and unrecognised diversity

Using the COI data generated in Chapter 2, divergent lineages consistent with interspecific variation (e.g. $> 3\%$) were found to be present within several common aquarium species. When a sufficient number of specimens were available (≥ 5) for aquarium species showing clear COI clusters, patterns were tested against the nuclear data. Four methods were used in assessing support for unrecognised or cryptic species: mean intergroup K2P distances; a character based approach using diagnostic, fixed character states between lineages[4]; bootstrap estimates of NJ clade support (settings as described in Section 3.2.3.2); and Rosenberg's *P*, a statistical

[3]Final concentration of each primer 0.3 μM.

[4]These have been referred to as "pure, simple characteristic attributes", or CAs (Lowenstein *et al.*, 2009; Sarkar *et al.*, 2008)

measure testing the probability of reciprocal monophyly over random branching processes (Rosenberg, 2007).

5.3 Results

5.3.1 Nuclear marker selection

5.3.1.1 Step one: 22 loci

A total of 22 candidate loci were selected from the review of the phylogenetic literature. Names, lengths, Ensembl references, and citations are reported in Table 5.1. The diversity of these genes across the five model organisms is presented in Figure 5.1, where they are ranked according to median levels of divergence. Of these 22 loci, the IRBP, RAG1(exon2), and MLL loci were chosen as sub-candidates due to their greater comparative variability when ranked by median divergence (Figure 5.1). Although the PRLR gene was also highly ranked, the alignment was highly divergent and the homology was questionable. The RAG2 locus was also favourably positioned as a variable nuclear region, although previous studies have suggested limited divergence at the species level (Hardman, 2004). Despite appearing relatively conserved at the ordinal level, the rhodopsin (RHO) gene has been proposed as a nuclear fish barcode (Sevilla *et al.*, 2007), and therefore warranted comparison with other loci identified in this study. Likewise, despite the relatively low divergence for LSU 28S, it has been reported to distinguish closely related species of fish (Sonnenberg *et al.*, 2007), and was therefore also chosen.

5.3.1.2 Step two: five loci

As described above, five loci in total (IRBP, RAG1exon2, MLL, RHO, LSU 28S) were chosen as sub-candidates to be tested on the selected *Danio* spp. (as outlined in Section 5.2.1). A total of 30 sequences were generated from the six *Danio* species with these nuclear genes. Primers and citations are presented in Table 5.2. GenBank accession numbers for the sequences generated here are presented in Table 5.3. The nuclear rhodopsin gene (RHO) was chosen as the marker with most potential for within species variation, showing the largest maximum, median and minimum pairwise distances of all comparison nuclear loci (Figure 5.2).

Table 5.1. Names of 22 candidate nuclear loci, with length (bp), citation, and Ensembl reference data (for *Danio rerio* sequences). Nomenclature follows literature cited.

Gene	Base pairs	Citation	*D. rerio* Ensembl gene ref.
BMP4	863	(Cooper *et al.*, 2009)	ENSDARG00000019995
EGR1	1071	(Chen *et al.*, 2008)	ENSDARG00000037421
EGR2B	1134	(Chen *et al.*, 2008)	ENSDARG00000042826
EGR3	1071	(Chen *et al.*, 2008)	ENSDARG00000089156
ENC1	810	(Li *et al.*, 2007)	ENSDARG00000035398
GLYT	870	(Li *et al.*, 2007)	ENSDARG00000010941
IRBP	1236	(Chen *et al.*, 2008)	ENSDARG00000059163
LSU 28S	1152	(Sonnenberg *et al.*, 2007)	EF417169 (GenBank)
MLL	2624	(Dettai & Lecointre, 2005)	ENSDARG00000004537
MYH6	732	(Li *et al.*, 2007)	ENSDARG00000090637
PLAGL2	672	(Li *et al.*, 2007)	ENSDARG00000076657
PRLR	1193	(Townsend *et al.*, 2008)	ENSDARG00000016570
PTR	705	(Li *et al.*, 2007)	ENSDARG00000008249
RAG2	1628	(Cooper *et al.*, 2009)	ENSDARG00000052121
RAG1 exon2	1140	This study	ENSDARG00000052122
RAG1 exon3	1749	(López *et al.*, 2004)	ENSDARG00000052122
RHO	1065	(Chen *et al.*, 2003)	ENSDARG00000002193
RYR3	822	(Li *et al.*, 2007)	ENSDARG00000071331
SH3PX3	705	(Li *et al.*, 2007)	ENSDARG00000014954
SREB2	987	(Li *et al.*, 2007)	ENSDARG00000068701
TBR1	660	(Li *et al.*, 2007)	ENSDARG00000004712
ZIC1	858	(Li *et al.*, 2007)	ENSDARG00000015567

Notes: LSU 28S is not available on Ensembl, so GenBank reference is included.
Abbreviations: ref. = reference.

5.3.1.3 Step three: one locus

A total of 200 RHO sequences were generated for 82 species of cyprinid fish (1–10 individuals per species), and are presented in FASTA format (online Appendix Section B.2), and uploaded to BOLD. The RHO fragment corresponded to an 858 bp length (sites 58–915) of the *Astyanax mexicanus* rhodopsin gene: GenBank accession U12328 (Sevilla *et al.*, 2007; Yokoyama *et al.*, 1995).

When comparing suitability of COI and RHO as a species level marker in the reduced, matched datasets, the NJ monophyly analysis yielded 98.6% identification success rate for COI, and 87.8% for RHO. The rates for the nearest neighbour analyses (k-NN) were 99.0% for COI, and 92.2% for RHO. The two genes representing two different genomes produced consistent results. However, the nuclear data performed slightly poorer at discriminating some closely related species. An NJ phenogram

Table 5.2. Primer names, sequences, and citations for five candidate nuclear loci.

Gene	Direction	Reference	Primer name	Primer sequence 5′–3′
RAG1 (exon2)	Forward	This study	RAG1ex2F	GGTGGATGTGACAACCGATA
RAG1 (exon2)	Reverse	This study	RAG1ex2R	ACGGGTCAGTGACAACAGGT
RHO	Forward	(Chen *et al.*, 2008)	RH28F	TACGTGCCTATGTCCAAYGC
RHO	Reverse	(Chen *et al.*, 2003)	RH1039R	TGCTTGTTCATGCAGATGTAGA
IRBP	Forward	(Chen *et al.*, 2008)	IRBP109F	AACTACTGCTCRCCAGAAAARC
IRBP	Reverse	(Chen *et al.*, 2008)	IRBP1001R	GGAAATGCATAGTTGTCTGCAA
MLL	Forward	This study	MLLcypF	GGCCCAGAGAAATTGATTGT
MLL	Reverse	This study	MLLcypR	ACTGGAAGGGACCGACACTA
LSU	Forward	(Sonnenberg *et al.*, 2007)	LSU D1-D2 fw1	AGCGGAGGAAAAGAAACTA
LSU	Reverse	(Sonnenberg *et al.*, 2007)	LSU D1-D2 fw1	TACTAGAAGGTTCGATTAGTC

Table 5.3. GenBank accession numbers for sequences generated from five candidate nuclear loci.

Danio species Specimen	*D. rerio* RC0394	*D.* aff. *kyathit* RC0405	*D. kyathit* RC0129	*D. dangila* RC0345	*D. albolineatus* RC0076	*D. margaritatus* RC0107
RAG1 (exon2)	JQ624037	JQ624038	JQ624035	JQ624036	JQ624040*	JQ624034
RHO	JQ614147	JQ614118	JQ614139	JQ614131	JQ614121	JQ614141
IRBP	JQ624025	JQ624026	JQ624023	JQ624024	JQ624021	JQ624022
MLL	JQ624031	JQ624032	JQ624029	-	JQ624030	JQ624028
LSU	EF417169	JQ624047†	JQ624045	JQ624046	JQ624043	JQ624044

Notes: (*) *Danio albolineatus* sequence from RC0445; (†) *D.* aff. *kyathit* sequence from RC0120.

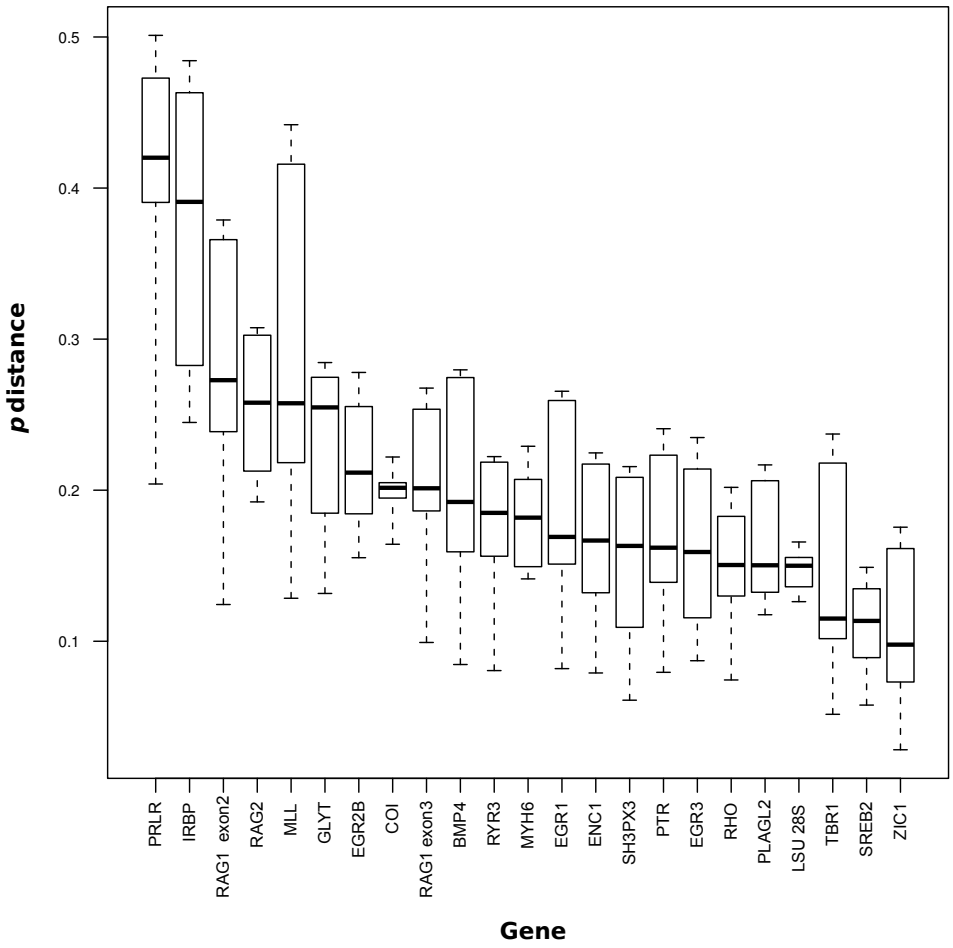

Figure 5.1. Uncorrected pairwise *p* distance ranges for 22 homologous candidate nuclear loci (and COI) between the following model organisms: *Danio rerio*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Takifugu rubripes* and *Tetraodon nigroviridis*. Whiskers extend to full range of data; boxes represent quartiles; black lines show median values.

of RHO data is presented in the online Appendix Section B.4, with links to the specimen pages on the BOLD Web site. Taxa unable to be resolved by RHO, but resolved for COI, include some members of the *Puntius conchonius* group including *P. padamya*, *P. tiantian* and *P. manipurensis*. *Danio albolineatus* and *D. roseus* were also unresolved, as were *Microdevario kubotai* and *M. nana*, plus *Devario* cf. *browni* and other associated undescribed/unidentified *Devario* species.

5.3.2 Interspecific hybrids

Interspecific hybrids (*Danio rerio* × *D.* aff. *kyathit*) were bred successfully under aquarium conditions. This hybrid had an identical COI sequence to *Danio rerio* RC0067 (BOLD process ID RCYY001-10), and the overall phenotype of the hybrid

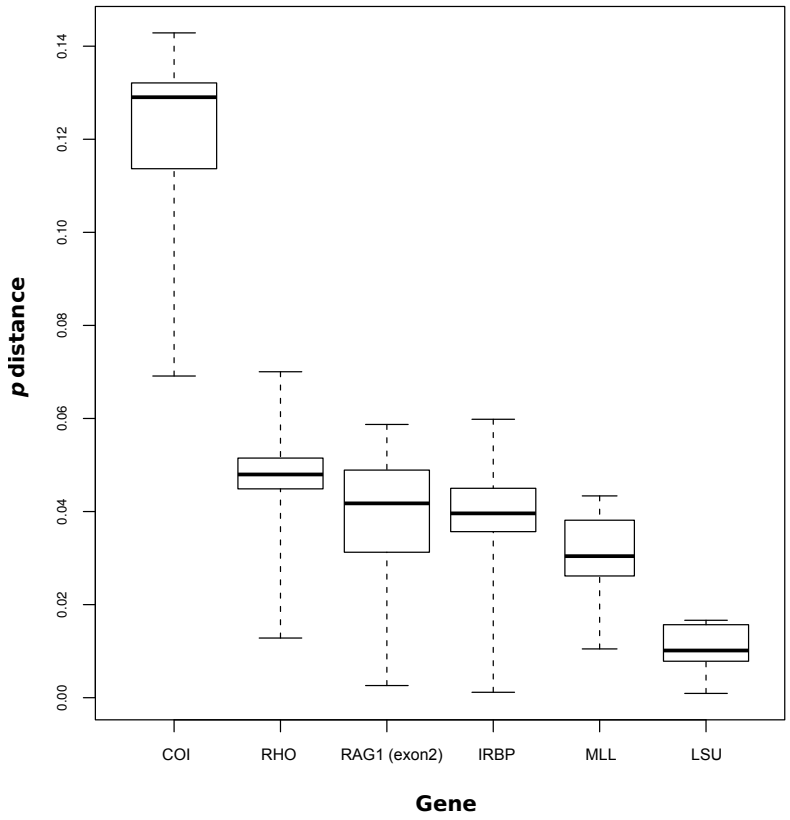

Figure 5.2. Intrageneric uncorrected pairwise *p* distance ranges between candidate nuclear loci from the following *Danio* species: *Danio* aff. *kyathit*, *D. albolineatus*, *D. dangila*, *D. kyathit*, *D. margaritatus* and *D. rerio*. Whiskers extend to full range of data; boxes represent quartiles; black lines show median values.

is shown in Figure 5.3. This hybrid was then sequenced for four of the short-listed nuclear genes (LSU 28S was not used at this stage due to sequencing problems). Heterozygosity was substantially higher in hybrid over non-hybrid parental species for all nuclear genes (Table 5.4), with the RHO gene showing the most polymorphic positions in the hybrid (32), compared to the other nuclear genes. Figure 5.4 shows a section of a trace file chromatogram for the hybrid *Danio*, with corresponding double peaks in both forward and reverse reads.

For the 200 RHO sequences of putative non-hybrid cyprinid fishes generated in this study, 95% had ≤ 4 heterozygous positions (median $= 0$; mean $= 0.99$; max. $=$ 17). Of these, seven individuals from six species (*Puntius conchonius*, *P. fasciatus*, *P. orphoides*, *P. oligolepis*, *P.* aff. *gelius* and *P. jerdoni*) had > 5 heterozygous positions. However, this had not been flagged as potential hybrids using morphological data. Three individuals from two species had > 10 (*P. oligolepis* and *P. jerdoni*). For the 1,530 RHO sequences downloaded from GenBank, 96% had ≤ 1 polymorphic sites

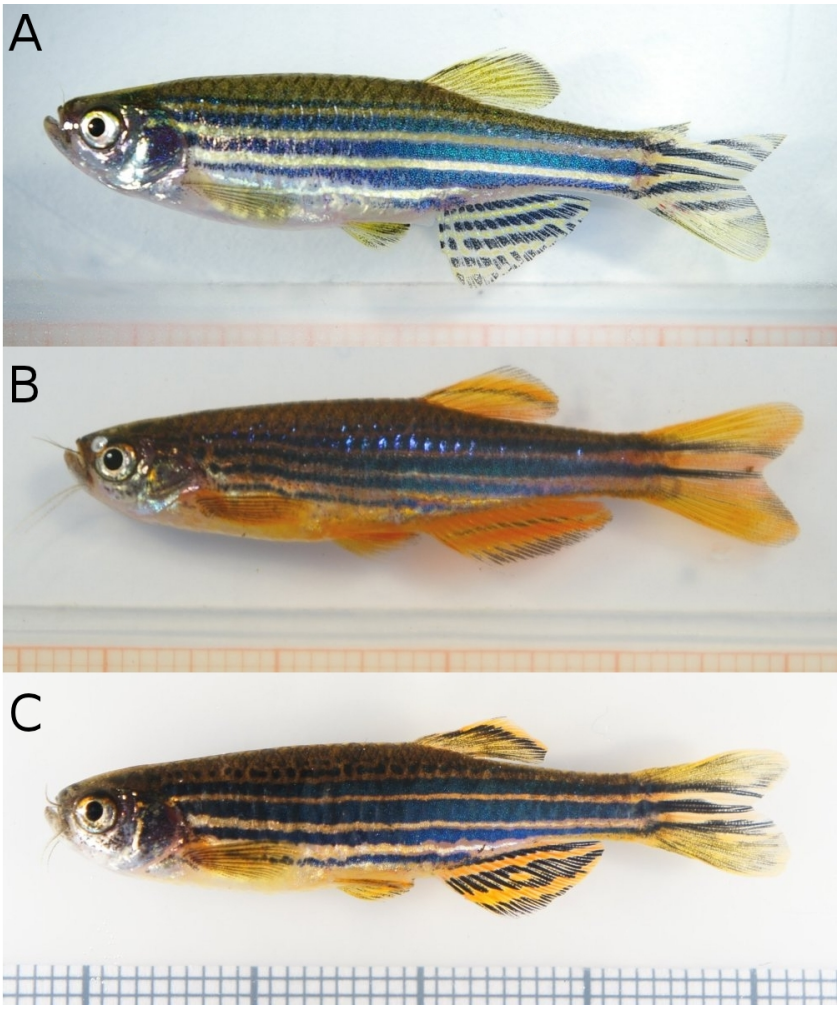

Figure 5.3. Phenotype of laboratory bred *Danio rerio* × *D*. aff. *kyathit* (C), parental species phenotype of *Danio rerio* RC0067 (A), and *D*. aff. *kyathit* RC0120 (B).

(median = 0; mean = 1.6; max = 35). The GenBank sequences varied in length from 336 to 1062 bp (mean = 561 bp).

Using the *Danio rerio* RHO sequence (RC0394) as the maternal species for the lab bred hybrid, a composite paternal sequence was generated. This sequence was identified as *Danio* aff. *kyathit* (the correct paternal species) using the BCM method. The sequence had an uncorrected *p* distance of 0.23% from the closest *D.* aff. *kyathit*, and clustered closest to this species in an NJ phenogram (not shown). The optimised threshold for minimising error of identification was 0.34% for the RHO data.

For the hybrid *Puntius* purchased in the aquarium trade, 14 polymorphic sites were observed in the RHO data (GenBank accession JQ614265). However, the maternal species could not be identified using the current COI library, being over

Table 5.4. Number of heterozygous nucleotide positions at four nuclear loci in a hybrid *Danio* (*D. rerio* × *D.* aff. *kyathit*) and specimens of its non-hybrid parental species. GenBank accession numbers for the hybrid are also presented.

Gene	Size (bp)	*Danio rerio* (RC0394)	*D.* aff. *kyathit* (RC0405)	Hybrid (RC0455)	GenBank accession
RAG1 (exon2)	768	2	1	24	JQ624039
RHO	858	0	0	32	JQ624041
IRBP	859	4	0	28	JQ624027
MLL	765	0	1	17	JQ624033

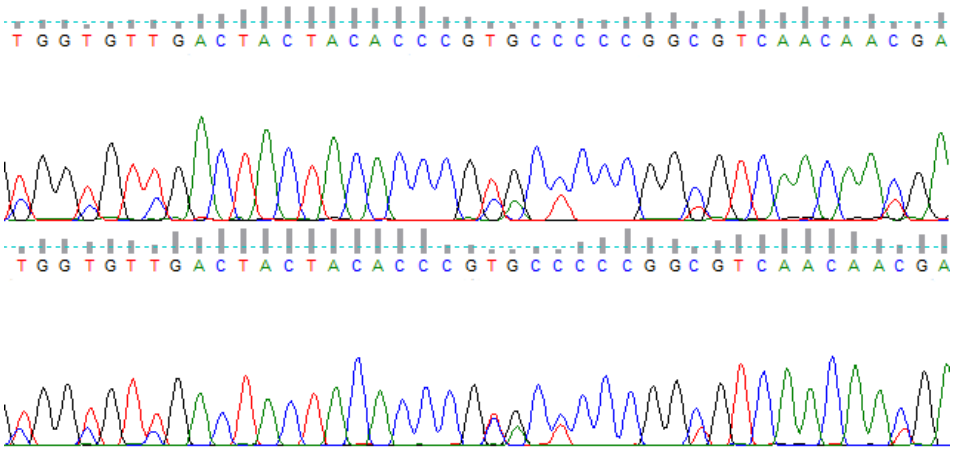

Figure 5.4. Chromatogram trace files for interspecific hybrid RC0455 (laboratory bred *Danio rerio* × *D.* aff. *kyathit*), showing multiple heterozygous positions in both forward (top) and reverse (bottom) reads). Note the low quality scores around the polymorphisms.

3% different from the closest match (*P. arulius*), and well above the 1.4% threshold for this dataset (Table 3.1). The composite sequence approach (using subtraction) was attempted using the closest available sequence of *P. arulius*. The resulting RHO composite could not be satisfactorily identified either, being 0.47% different from the nearest match of *P. denisonii* (threshold 0.34%). However, in the NJ phenogram (not shown) the sequence was nested within the *P. denisonii* cluster, and this species was identified as a potential parent during the morphological identification process having a distinctive red longitudinal stripe, which is present in few *Puntius* species.

Of the catfishes, the hybrid clariid RC0739 sequenced for RHO, was found to have 11 polymorphisms. Due to conflicting GenBank data (multiple species names with identical haplotypes), a species level identification could not be made using

cyt *b* downloaded from GenBank, or via a BLAST search. However, the specimen nested within the cluster of *Heterobranchus* (NJ phenogram not shown). Data for this specimen were uploaded to GenBank: JQ624018 (RHO); JQ624019 (cyt *b*). The pimelodid catfish hybrid (RC0374) also had a large number of polymorphisms at 19. This specimen was again unable to be identified to species from cyt *b* data in GenBank, and clustered within a poorly resolved group comprising several species of *Pseudoplatystoma* (NJ phenogram not shown). Data for this specimen (RC0374) were uploaded to GenBank: JQ624042 (RHO); JQ624020 (cyt *b*). The 16 hybrid *Synodontis* catfish specimens sequenced for cyt *b* formed seven distinct NJ clusters (phenogram not shown), four of which were close to species represented in the GenBank data. These specimens did not amplify well for RHO, unfortunately, with the sequences being of poor quality (different primer pairs and combinations were also tried). There also did not appear to be a large number of polymorphic sites in this *Synodontis* RHO data.

5.3.3 Cryptic and unrecognised diversity

Aquarium species identified as having significant "within species" variation for COI are reported asn NJ phenogram in Figure 5.5; they included: *Danio choprae*, *D. dangila*, *D. kyathit, Devario devario, Epalzeorhynchos kalopterus, Microdevario kubotai, Microrasbora rubescens, Puntius assimilis*, *P. denisonii*, *P. fasciatus*, *P. gelius*, *P. lateristriga*, *P. stoliczkanus*, *Rasbora dorsiocellata*, *R. einthovenii*, *R. heteromorpha*, *R. maculata*, *R. pauciperforata* and *Sundadanio axelrodi*. Some were expected, based on the morphological examination process, to be unrecognised diversity (noted by "sp.", "cf." or "aff."), and some were divergent in the absence of apparent morphological differences (i.e. so-called cryptic species).

For 11 of the species, greater than five individuals were available for comparisons between both loci to assess whether the COI relationships were supported with nuclear RHO data. Where COI splits were large, the RHO distances were also large, albeit on average 9.9× smaller (range 3.8–22.7×). Discrete character states were observed for all species in both genes, were again fewer at the nuclear locus, and also corresponded to lower bootstrap support. Rosenberg's *P* statistic of reciprocal monophyly showed significance for all but two comparisons with COI, and all but four comparisons with RHO. A full summary is presented in Table 5.5.

Table 5.5. Exploring unrecognised diversity: undescribed and putative cryptic species were assessed with COI and nuclear RHO data in the context of their closest known congener or conspecifics.

Putative cryptic or unrecognised taxon	Taxon comparison	$n =$	Mean K2P % COI/RHO	No. CAs COI/RHO	Bootstrap % COI/RHO	Rosenberg's *P* COI/RHO
Danio aff. *choprae*	*D. choprae*	6	7.4 / 0.5	23 / 2	100 / 92.7*	Y / N*
Danio aff. *dangila*	*D. dangila*	7	9.0 / 1.3	21 / 10	100 / 89.9	Y / Y
Danio aff. *kyathit*	*D. kyathit*	6	7.0 / 1.1	40 / 7	100 / 100	Y / Y
Danio sp. "hikari"	*D.* cf. *kerri*	6	8.6 / 0.6	48 / 5	100 / 97.1	Y / Y
Devario sp. "purple cypris"	*D. auropurpureus*	6	8.1 / 0.6	47 / 5	100 / 99.8	Y / Y
Microrasbora cf. *rubescens*	*M. rubescens*	5	3.7 / 0.5	23 / 3	100 / 95.3	N / N
Puntius aff. *gelius*	*P. gelius*	7	17.2 / 4.1	76 / 27	100 / 100	Y / Y
Puntius denisonii	intraspecific	5	7.8 / 0.4	40 / 3	100 / 95.7	N† / N
Rasbora aff. *dorsiocellata* ‡	*R. dorsiocellata*	6	10.9 / 1.5	46 / 8	100 / 82.5	Y / Y
Rasbora cf. *heteromorpha*	*R. heteromorpha*	7	2.2 / 0.2	11 / 1	100 / 18.1	Y / N
Sundadanio cf. *axelrodi*	intraspecific	10	13.8 / 2.3	42 / 9	100 / 99.6	Y / Y

Notes: (*) renders *Danio choprae* paraphyletic; (†) *P* monophyly significant to the α 10^{-4} level with combined COI data (15 specimens); (‡) species likely described during thesis preparation as *Brevibora cheeya* (Liao & Tan, 2011). Abbreviations: CA = pure, simple characteristic attribute (i.e. discrete diagnostic character state); Y = Rosenberg's *P*, significant to $\alpha = 0.05$; N = not significant.

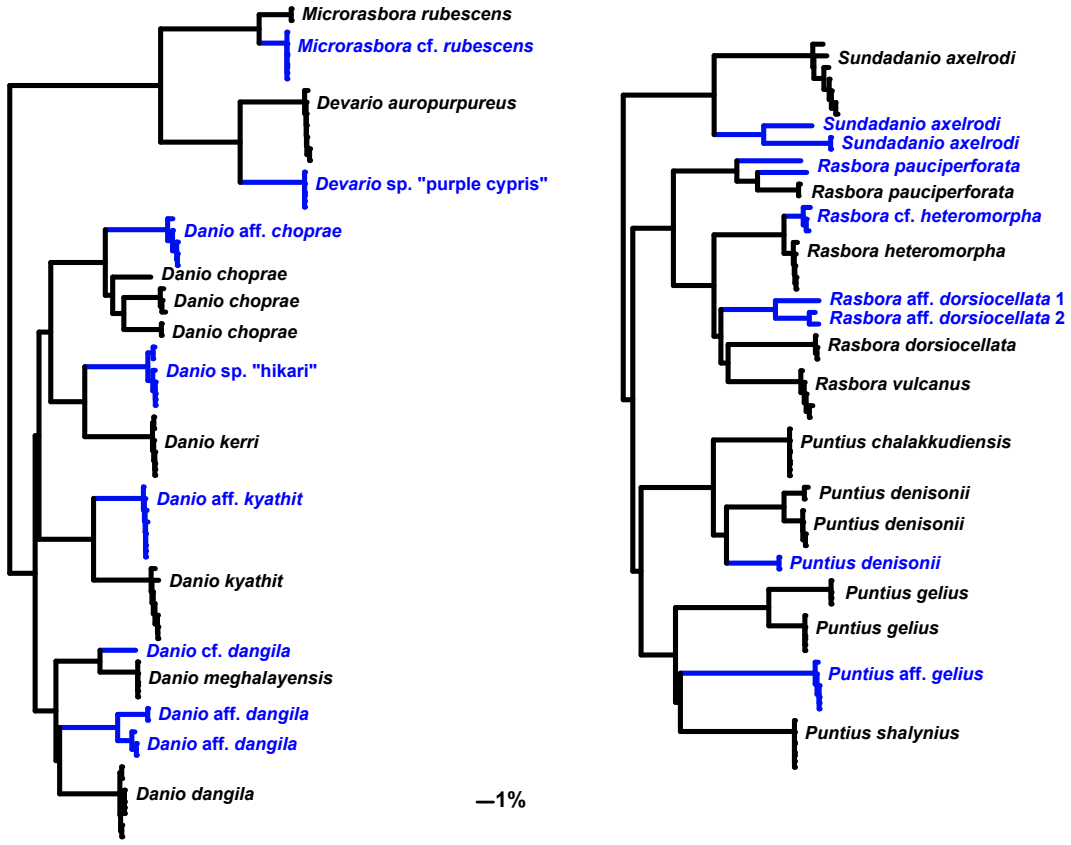

Figure 5.5. Cryptic and unrecognised species. An NJ phenogram showing deep COI barcode divergences in selected ornamental species. Taxa of interest are highlighted in blue.

5.4 Discussion

5.4.1 Nuclear marker selection

The relationship between genomic diversity across orders as an indicator of that within species is not necessarily a justified one, as selection or homoplasy may provide substantial sources of bias. As an example, COI is highly variable at the species level, but Figure 5.1 shows that its maximum variation is quite limited—this is likely due to the functional constraints of the mitochondrial protein. Despite this, as a crude way to screen for fast or slowly evolving loci, looking at genomic diversity may help in uncovering potentially useful markers for further testing. Among the nuclear genes tested for diversity within the *Danio* genus, and with the exception of LSU 28S, the chosen loci showed similar levels of diversity (Figure 5.3). As proposed by Sonnenberg *et al.* (2007), LSU 28S appeared a promising marker for species

level inference. However, as well as the low levels of variability, tests using this marker on *Danio* and *Puntius* indicate numerous indels, considerable ambiguity in alignment, and difficulty in both amplification and sequencing (slippage due to long mononucleotide stretches). For these reasons, this marker was abandoned as a tool that could be fit for purpose in a biosecurity diagnostics context. The protein coding nuclear loci offered a considerably easier laboratory procedure, although do not benefit from the homogenisation by concerted evolution as seen in the rRNA genes (Elder & Turner, 1995), and may display some allelic variation (Chen *et al.*, 2008). The rhodopsin gene was finally selected to investigate variation at the species level, due its variability (Figure 5.2, Table 5.4), wide use in phylogenetics (e.g. Fang *et al.*, 2009), and the availability of published primer sets (e.g. Chen *et al.*, 2003; Sevilla *et al.*, 2007).

When tested on 200 specimens of cyprinid fish, RHO was found to separate species well, broadly agree with morphological assignments, and support COI. Its resolution, however, was not as fine as that of COI, failing to discriminate among some closely related groups. It could not be therefore recommended as a single locus identification system, but does offer a suitable method of verifying mitochondrial results in terms of hybridisation and unrecognised diversity (see below).

Among these protein-coding nuclear genes, several potential pitfalls may occur. Many cyprinid fishes have undergone historical whole-genome duplication events, and are therefore polyploid and highly diverse in terms of alleles, even before hybridisation (Chen *et al.*, 2008). Furthermore, it is questionable whether some of these nuclear loci represent neutral markers (see Galtier *et al.*, 2009), as for example, substantial adaptation to local spectral environments has been documented in the RHO gene—a vision pigment—for a *Pomatoschistus* goby (Larmuseau *et al.*, 2009). This may call into question the utility of the gene for accurately recovering phylogenetic relationships or even offering species level identifications; does sequence similarity between two groups reflect convergent adaptation, conspecificity, or lack of variation and incomplete lineage sorting?

5.4.2 Interspecific hybrids

The breeding of aquarium hybrids in a controlled environment provided an important opportunity to test how effectively screening with an nDNA marker can detect interspecific hybridisation events. When both mtDNA and nDNA data were available for the maternal species, it was possible to accurately predict the paternal species of

the hybrid using the polymorphisms in the RHO data, as was the case with the lab bred hybrid, and to some degree the hybrid *Puntius* from the trade. For taxa where these extra data were not available (hybrid catfishes), the high level of heterozygosity in the nDNA was able to independently suggest potential for hybrid origin.

Separating the hybrid and non-hybrid individuals with nDNA data required a difference in the proportion of heterozygosity. The background level of heterozygosity for RHO in putatively natural populations is estimated here to be low, with most (95%) of the cyprinid fishes surveyed having less than four polymorphic sites across 858 bases. The data taken from GenBank proved to be even less heterozygous (96% with < 1 polymorphism). However, it is almost a certainty that the bulk of this data were not investigated as thoroughly for polymorphisms as those presented here, and were scored using the automated base calling in programs such as SEQUENCHER. Many of the GenBank sequences were also shorter than those used here, so fewer polymorphic sites are to be expected.

The lab produced hybrid had a considerably higher levels of heterozygosity at 32 positions, than these putative background levels, as did the hybrid *Puntius* purchased in the aquarium trade (14 positions). The two catfish (clariid and pimelodid) species sourced, also showed high levels (11 and 19 respectively). Therefore, an individual with an arbitrary level of heterozygosity of over ten bases in 858 appears indicative of a hybrid, and less than five bases, of a non-hybrid. However, some specimens with intermediate to large values were reported, and did not appear to be hybrids. It is possible that these high values were caused by large intrapopulation variation (potentially due to adaptive selection), polyploidy, or interspecific hybridisation that was not detected by examining the morphology of the fishes.

The *Synodontis* catfishes are well known subjects of hybridisation in the aquarium trade (Ng, 2010). However, the RHO protocol used here failed to yield consistently clean PCR products or sequence data. From those that were sequenced, the amount of polymorphism appeared to be low (frequently < 5). This may have been a consequence of the primers binding to only one allele, the RHO gene being insufficiently variable in this group, or that these putative hybrids were not in fact hybrids. Regardless, using the measure of heterozygosity as presented here to detect hybrids may not be effective in all cases, especially where primers are poorly fitting.

Whether the method can be applied to a wider variety of groups remains to be tested more thoroughly, and is dependent upon getting tissue samples of specimens with known hybrid and non-hybrid pedigrees. It is also unlikely that the method will be sufficiently sensitive to detect hybridisation among natural populations of closely

related species in hybrid zones for example, as this would require a considerably more sophisticated approach using multiple microsatellite markers (see Section 5.1.1). Fortunately, many of the hybrids created for the aquarium trade are selected for novel phenotypes, and therefore more distantly related species are deliberately chosen. A crude test for heterozygosity should therefore in theory be able to detect the more egregious examples of the practices undertaken by ornamental fish breeders. However, it is unknown how heterozygosity is affected by the further breeding of hybrid and backcrossed generations past F_1, something which may well be taking place in the trade.

5.4.3 Cryptic and unrecognised diversity

In terms of unrecognised diversity and potentially cryptic species, significant within-species COI diversity was observed in several common ornamental species, and cases of otherwise unreported morphological variation was also recognised. For an exemplar group of aquarium species, and where sufficient numbers of individuals were available, additional support for these divergent COI lineages was assessed with the nuclear RHO marker using statistical and character-based analyses, successfully demonstrating evidence in both genomes. The RHO supported most of the relationships proposed by COI, indicating that both genes are effective and complimentary tools in assisting in species delimitation for poorly known taxa.

Implications for conservation and sustainable management of fisheries are apparent here; *Puntius denisonii*—a species at risk of over-exploitation (Raghavan *et al.*, 2007)—was found to possibly comprise at least two structured and morphologically cryptic lineages. As highlighted by Rosenberg's P, sample sizes were relatively small, and this may indicate where further sampling would be beneficial.

Supporting methods using nuclear data attempt to build on the solely mitochondrial approach by providing congruence with an external dataset (Dasmahapatra *et al.*, 2010; Dasmahapatra & Mallet, 2006; Elias *et al.*, 2007). Of course, if taxonomic work is also undertaken, then specimens with known locality data should be sourced. However, the hypotheses generated here certainly warrant further investigation into species limits of these particular taxa, and this process provides useful reference points for closer examination. Until this work is carried out, data are made available in the BOLD database, and identifications of fishes in the ornamental trade will have to be made using tag names.

5.5 Summary

In this chapter, the benefits of incorporating nDNA data into a DNA barcoding approach are apparent. The ability of a simple nDNA test to detect fishes of interspecific hybrid origin was assessed, and which worked as predicted for controlled, lab bred hybrids, plus some examples from the trade. Identification of both parental species was even possible when sufficient reference data were available. Unfortunately, other hybrids purchased from the aquarium trade were unable to be identified as such, indicating a universal and simple method to detect fish hybrids through nDNA sequencing requires further work (possibly with allozymes). Taxonomically unrecognised lineages as well as morphologically cryptic ones were deemed biologically plausible with the support of data from the nuclear genome. This assists in verifying the authenticity of patterns in the mtDNA data, and can provide additional hypotheses for taxonomic investigation.

Chapter 6

An evaluation of environmental DNA for biosecurity applications

6.1 Introduction

Environmental DNA (eDNA) can now be accessed from a diverse range of substrates, opening up new areas of biodiversity research in terms of both microbiological and macrobiological samples (Thomsen *et al.*, 2012; Venter *et al.*, 2004). In aquatic ecosystems, assessment of species' distribution can now be made using eDNA present in water, an approach allowing the detection and monitoring of invasive species (Ficetola *et al.*, 2008; Jerde *et al.*, 2011), rare and secretive species (Goldberg *et al.*, 2011), or community composition as a whole (Minamoto *et al.*, 2012). In terms of invasive species monitoring, Ficetola *et al.* (2008) reliably detected the presence of invasive bullfrogs in both controlled conditions and in natural ponds, while Jerde *et al.* (2011) delimited invasion fronts of two Asian carp species in the Laurentian Great Lake system of the United States. Despite the relatively recent introduction of the technique, eDNA analyses are quickly becoming recognised as an important tool for invasion biologists and ecosystem managers (Darling & Mahon, 2011).

6.1.1 Border quarantine

Immediately upon import at the border, ornamental fishes in many countries are subjected to a period of quarantine (Ploeg *et al.*, 2009). This is particularly the case for Australia and New Zealand, where fish imports are restricted, and shipments are monitored for exotic pathogens (MAF Biosecurity New Zealand, 2011; McDowall, 2004; Whittington & Chong, 2007). Freshwater fishes imported into New Zealand are currently quarantined at transitional facilities for no fewer than four weeks, in order to allow manifestation of infection or mortality (MAF Biosecurity New Zealand, 2011). The quarantine stage therefore also offers an opportunity to identify the

shipped species, and monitor the imports for the presence of clandestine hitchhikers (i.e. contaminant or bycatch species).

The benefits of molecular over morphological approaches for border biosecurity identification of specimens have been acknowledged elsewhere (Chapter 1; Armstrong & Ball, 2005; McDowall, 2004). However, there are also several benefits of using eDNA over tissue sampling of imported fishes (i.e. standard DNA barcoding). First of all, tissue sampling procedures are invasive in terms of damage to the organism tested. Fin clips or swabs can be taken, but may leave the fish susceptible to infections through breaking the skin, or the removing the protective mucous layer (Le Vin *et al.*, 2011). On the other hand, destructively sampling entire individuals may not be possible if the fish is valuable or only a single example is available.

Using eDNA, we have the ability to detect presence of a target species among multiple individuals of a shipment, rather than that of the single specimen chosen for testing; this may be important in terms of identifying mixed consignments. Environmental DNA techniques therefore have the potential to assess abundance and composition of fishes in a shipment. Because water will to some degree hold a "molecular memory" of the species present in it, eDNA protocols can therefore track the historical presence of a species in a water sample. This may be of benefit if a particular high-risk taxon in terms of pathogen vectoring potential has been in recent close contact with an otherwise low-risk species at a wholesaler or transshipper. This would perhaps justify added precautions to be taken in terms of disease risk and quarantine.

6.1.2 Transport of live fishes

Internationally, live ornamental fishes are transported by air freight. This entails securely packing the fishes to enable their survival for a minimum of approximately 48 hours (Ploeg *et al.*, 2009). Packaging requirements depend on various factors such as the sensitivity, size, and value of the species concerned. However, densities are usually maintained at the highest possible, to maximise cost-effective shipping (Cole *et al.*, 1999). Fishes are typically placed in plastic bags with 20–35% water, inflated with oxygen, sealed, and then shipped in polystyrene boxes. Bag size varies, but large bags (37.5 cm × 37.5 cm × 55 cm) will contain up to seven litres of water and between 10 and 500 fish depending on their size (Cole *et al.*, 1999); individual fishes are bagged in smaller volumes. In contrast to the low concentrations of eDNA from samples of natural water bodies, due to the high packing densities of traded

fishes, retrieving eDNA in this situation may in some respects be less complicated (notwithstanding the potential for PCR inhibition due to fish metabolites in the transport water).

6.1.3 eDNA targets

Mitochondrial DNA is the preferred target for aquatic eDNA studies, although microsatellites have been genotyped from degraded substrates such as faecal matter (Taberlet *et al.*, 1996). Mitochondrial DNA offers a higher copy number than nDNA, and therefore better amplification likelihood essential when dealing with potentially degraded samples (Valentini *et al.*, 2009; Willerslev & Cooper, 2005). This is also due partly to "cellular location, chromatin structure and transcriptional activity" (Foran, 2006). As a result, most studies of aquatic eDNA focus on short amplicons of mtDNA between 80 and 300 bp (Ficetola *et al.*, 2008; Jerde *et al.*, 2011; Thomsen *et al.*, 2012). Fortunately, the high variability of the standard DNA barcode marker COI, can allow species discrimination using mini-barcode fragments much smaller than the standard ~650 bp (Hajibabaei *et al.*, 2006b; Shokralla *et al.*, 2011).

The choice of which mini-barcode regions best differentiate taxa is important, but rarely explored. Roe & Sperling (2007) in their analysis of COI and COII, found significant substitutional heterogeneity through these genes and across taxa; they found no one region was best in all cases. Ideally, however, the most informative regions should be chosen for a specific study taxon, although to some degree the choice is limited by the availability of suitable priming sites (Ficetola *et al.*, 2010). Sliding window analyses can therefore be used as a tool to evaluate variability though a gene alignment and find informative regions flanked by less variable priming locations, or, for species specific applications, to locate diagnostic sites for probe design (Boyer *et al.*, 2012). A sliding window method "extracts all possible windows of a chosen size in a DNA alignment" and performs various analyses on these subsets of the full alignment (Boyer *et al.*, 2012). Alternatively, for larger scale meta-barcoding projects (cf. Andersen *et al.*, 2012; Valentini *et al.*, 2009), use of software such as ecoPrimers (Riaz *et al.*, 2011) can now utilise huge genomic datasets to automate and optimise selection of primer sets for informative short length markers.

6.1.4 Environmental persistence of eDNA

DNA molecules have been shown to persist in the environment for some considerable time—many hundreds of thousands of years if preserved in favourable conditions (Pääbo *et al.*, 2004; Willerslev & Cooper, 2005). DNA is shed by organisms via their faeces, urine and epidermal cells (Thomsen *et al.*, 2012), and can survive in an extracellular state for some time. The persistence of eDNA can be expressed as the presence of viable nucleic acids in the environment at a given rate of degradation, after the removal of its source (i.e. living tissues), while its detection depends on the concentration in the sample and the sensitivity of the test (Darling & Mahon, 2011; Dejean *et al.*, 2011). The aquatic environment is not one suited to the long term preservation of DNA, and most studies acknowledge that the observation of eDNA reflects only the relatively contemporary presence of the target (Thomsen *et al.*, 2012). Numerous mechanisms accelerate eDNA decomposition, and are outlined by Hofreiter *et al.* (2001) and Pääbo *et al.* (2004). They include: endogenous nucleases, microorganisms, oxidation, radiation, and hydrolysis, with these being influenced in turn by factors such as temperature, pH or light (Thomsen *et al.*, 2012). Dejean *et al.* (2011) experimentally demonstrated the decrease in detection ability of eDNA in freshwater, with detection possible up unto approximately 30 days under their controlled conditions.

6.1.5 Techniques for eDNA extraction

Compared to tissue sampling, successfully retrieving viable nucleic acids dissolved at low concentrations in water presents challenges. Two techniques are currently available to achieve this: filtration and precipitation. Filtration by vacuum can pass large volumes of water though a micropore filter (0.5–1.5 μm), before extractions are carried out on the filter material (Goldberg *et al.*, 2011; Jerde *et al.*, 2011). Alternatively, dissolved DNA can be precipitated out of water directly by adding an ethanol and sodium acetate solution before centrifugation at high speeds to concentrate the DNA (Ficetola *et al.*, 2008; Minamoto *et al.*, 2012). Although filtration is unlikely to recover DNA as efficiently as precipitation, due to the limitations in the volumes that can be centrifuged, filtration remains the primary option where very low concentrations of eDNA are expected, and water sample volumes are required to be measured in litres rather than millilitres (Thomsen *et al.*, 2012).

6.1.6 Objectives

The primary objective of this study is to create a proof-of-concept for the amplification and subsequent identification of ornamental fishes using eDNA in aquarium water. Secondly, a standardised protocol will be outlined to further develop the method to encompass more species. The sliding window method of marker evaluation and design will be assessed, and technical aspects of eDNA detection will also be tested, particularly in reference to relaxing some of the published requirements in terms of water volume and PCR repetition.

6.2 Materials and methods

6.2.1 Fish husbandry

To test a mini-barcode eDNA approach, experimental fishes were maintained in stock aquariums. Fishes chosen were the hybrids of *Danio rerio* and *D.* aff. *kyathit*, as bred in Chapter 5. They are maternally *D. rerio* and have the mitochondrial DNA of this species (haplotype of RC0067, BOLD process ID RCYY001-10), and are from here on referred to as *D. rerio*. The experimental fishes were kept in 50 cm × 25 cm × 25 cm aquariums (∼30 litre). Tanks were individually filtered with an EHEIM internal power filter, and supplementary aeration was provided via an airstone. Tank decoration comprised either a bare or inert sand substrate, along with Java moss (*Taxiphyllum barbieri*). Fishes were fed twice daily with proprietary flake food (TETRA brand). Temperature was ambient lab temperature at approximately 18–24°C). A 75% water change was carried out weekly with untreated tapwater at approximately tank temperature; Lincoln University tapwater is not chlorinated.

6.2.2 Primer design using sliding windows

The COI DNA barcode reference library as generated in Chapter 2 was chosen as the base for mining a short length *Danio rerio* specific marker[1]. The alignment of COI sequences for all *Danio* species was analysed for suitable mini-barcodes using the *slideAnalyses* (sliding window) function of the DNA barcoding package SPIDER (Brown *et al.*, 2012). The sliding window function takes a fixed length section of DNA (e.g. 100 bp), and from the first base, moves down the entire alignment at set

[1]During initial tests, attempts were made to amplify full length DNA barcodes from water samples, but these proved unsuccessful (data not shown).

intervals (e.g. every one or three bases). For each window, a series of calculations are made on the information content or discriminatory power. For this analysis the following measures were used: species monophyly, proportion of species with non-zero distances to nearest non-conspecifics (i.e. proportion of species that do not have identical sequence to a different species), mean K2P distance for all distance comparisons, and the number of diagnostic sites for each species, i.e. pure, simple characteristic attributes (Sarkar *et al.*, 2008). The resulting plots can then be viewed, and primers designed using information from the output. Design of final primer pair is described in Results (Section 6.3.2).

6.2.3 Primer specificity

6.2.3.1 *In vitro* PCR

The *in vitro* analysis comprised testing for PCR amplification success of the mini-barcode primers against previously extracted tissue samples of all sampled *Danio* spp., plus representatives of closely related genera (e.g. *Devario, Microrasbora, Microdevario*). Tissue extractions had been stored in elution buffer at −20°C, and were between 18 and 38 months old (see Section 2.2.4.1 for protocol). A list of species is presented in Table 6.2; at least two specimens of each species were tested, comprising different haplotypes where possible. As a control for DNA degradation since extraction, full length DNA barcodes were also amplified in parallel on the same tissue extractions.

Optimised PCR reactions were carried out using a Veriti thermocycler (APPLIED BIOSYSTEMS) in 10 µl reactions with the following reagents: 2.5 µl ultrapure water; 5.0 µl GoTaq Green Master Mix (PROMEGA); 1.0 µl forward and reverse primer (2.0 µM)[2]; and 0.5 µl DNA template. The primer pair used for the mini-barcode amplicon are presented in Section 6.3.2. Primers used to amplify the full DNA barcode were either LCO1490A and HCO2198A (Tang *et al.*, 2010), or FishF1 and FishR1 (Ward *et al.*, 2005). A negative (water) and positive (*D. rerio* template) PCR control was also used for both the mini and full barcode amplification reactions. Thermocycler settings for the mini-barcode reaction comprised: 2 min at 94.0°C; 35 cycles of 15 s at 94.0°C, 30 s at 61.0°C and 30 s at 72.0°C; 7 min at 72.0°C; ∞ at 4.0°C. Thermocycler settings for the full barcode comprised: 2 min at 94.0°C; 35 cycles of 15 s at 94.0°C, 30 s at 48–52°C and 45 s at 72.0°C; 7 min at 72.0°C; ∞ at 4.0°C.

[2]Final concentration of each primer 0.2 µM.

PCR products were visualised over ultraviolet light on a 4% agarose gel, stained with RedSafe (ChemBio), according to the manufacturer's protocol. Electrophoresis was run for 15 min (170 v, 50 mA) in a sodium hydroxide and borate buffer (pH 8.5); 6 µl of PCR product was added directly to the well.

6.2.3.2 *In silico* PCR

To test if organisms other than the immediately related ones (i.e. those tested in the *in vitro* experiment) are likely to amplify with the mini-barcode primers, an *In silico* search was made using the program MFEprimer (Qu *et al.*, 2009). MFEprimer is able to evaluate the "specificity of PCR primers based on multiple factors, including sequence similarity, stability at the 3′ end of the primer, melting temperature, GC content and number of binding sites between the primer and DNA templates" (Qu *et al.*, 2009). All COI sequences were downloaded from the GenBank nucleotide database (date: 02/02/2012), under the search term "COI" (total 810,305 sequences). A local installation of MFEprimer was run under both default settings (word size 11, and *e* value 1,000), and more stringent settings (word size 7, and *e* value 10,000).

Primer specificity was also tested against a larger set of published data in GenBank (i.e. targets other than COI, as well as COI), using the Primer-Blast tool available online at http://www.ncbi.nlm.nih.gov/tools/primer-blast/ (Altschul *et al.*, 1990; Rozen & Skaletsky, 2000). Template DNA was entered as the target *Danio rerio* sequence, and primers used were as presented in Table 6.1. The reference database selected was set to "nr" (all nucleotide records in GenBank), misprimed product size deviation was set to 100 bp to minimise hits on products that will be identifiable by significant length variation, and all other settings remained as default. Total allowed mismatches with at least one primer were set from between one to nine.

6.2.4 eDNA detection

6.2.4.1 Experimental treatments

Environmental DNA experiments were carried out in 20 litre containers, each with an airstone—from a single air pump supply—to ensure animal welfare during the experiments. Water used for each experiment was tapwater at the same temperature as the stock aquariums. Fishes were caught from the stock tanks with a sterilised net, and transferred to the container minimising dripping water. Fish were left in

the container overnight in a dark room for 16 hours. The air pump was turned off 10 minutes prior to the collection of water, to allow any detritus to settle. When water was collected, the fish remained in the water; samples were collected from the surface in clean, 50 ml FALCON tubes.

Two density treatments were used: (A) a single fish in four litres of water (∼0.24 g fish per litre); and (B) a single fish in 12 litres of water (∼0.08 g fish per litre). Each treatment was repeated four times in sets of four and included one negative control container on each occasion (total 12 repetitions with fish, and four without fish); i.e. for every three replicates, each container was in turn used as a negative control (no fish added). Average fish mass was estimated by placing 25 fish in a water-filled beaker on a zeroed digital balance, and a mean taken (0.95 g). The experiment otherwise proceeded as outlined below.

All equipment was sterilised after each experiment for a minimum of three hours with 1.25% sodium hypochlorite solution (one part 5% bleach solution to three parts water) (Champlot *et al.*, 2010; Kemp & Smith, 2005). As both fishes and DNA molecules are sensitive to chlorine (Brungs, 1973; Champlot *et al.*, 2010; Kemp & Smith, 2005), after rinsing with freshwater three times, any remaining chlorine was neutralised with SEACHEM PRIME at quadruple the recommended dosage (to account for the increased chlorine content of the diluted bleach solution). Containers were rinsed again with tap water.

Three 15 ml water samples were taken from each container, and immediately added to a premixed FALCON tube containing 33 ml of pure ethanol and 1.5 ml of 3 M sodium acetate (pH 5.2) at −20°C following Valiere & Taberlet (2000), and Ficetola *et al.* (2008). They were incubated at −20°C overnight, and then centrifuged for 1 hour at $10,000 \times g$ and 6°C in an Eppendorf 5810R centrifuge (cf. Minamoto *et al.*, 2012). The supernatant was then poured off and the tube placed horizontally to air dry for approximately three hours at room temperature. The DNA pellet was then subjected to a spin column extraction using the Quick-gDNA spin-column kit (ZYMO RESEARCH CORPORATION). The Genomic Lysis Buffer (250 µl) was added directly to the FALCON tube, vortexed for 20 seconds and then the three samples from each fish container were pooled into a single spin column. The extraction followed the manufacturer's protocol, but was scaled to use a 50% volume of pre-elution reagents. Fish experiments and DNA extractions were carried out in dedicated rooms, free of PCR product contamination. An outline of experimental procedure for a single replication of water sampling from one container is shown in Figure 6.1.

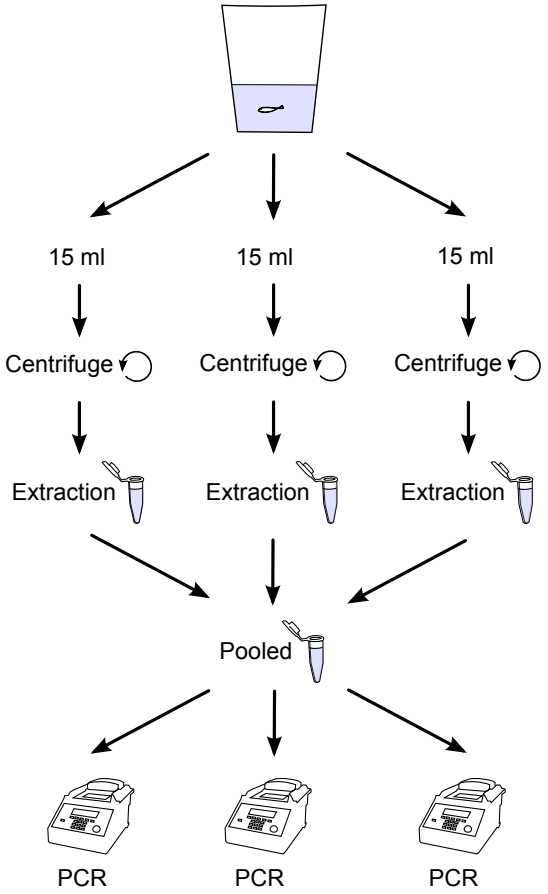

Figure 6.1. Flow diagram illustrating the protocol for a single experimental replication of eDNA extraction from water. © Rupert A. Collins, 2012.

PCR protocols remained as for the specificity experiment (Section 6.2.3.1), but the number of thermocycles was increased to 45, and the proportion of some reagents was changed: 1.0 µl DNA template, 1.5 µl of forward and reverse primer[3], and 1.0 µl ultrapure water. Again, a negative (water) and positive (eDNA extraction template of *D. rerio*) PCR control was used. Following the multi-tubes approach (Taberlet *et al.*, 1996), to reduce stochastic variation in amplification success from low DNA concentrations (i.e. that a failure to amplify is not due to chance), three PCRs were carried out on each of the DNA extractions from the pooled samples (Jerde *et al.*, 2011). Gel electrophoresis was carried out as above. A positive identification comprised a single band at the expected length (∼100 bp) in at least one of the three PCRs for each extraction. From both of the two density treatments, four positive PCR

[3]Final concentration of each primer 0.3µM.

products were chosen at random to be bidirectionally Sanger sequenced (protocol as Section 2.2.4.2).

6.2.4.2 Operational testing

To test the technique in an operational, biosecurity context, water samples from a shipment of the target species (*Danio rerio*) were taken at a MAF Biosecurity New Zealand transitional (quarantine) facility. The fishes were identified visually by officials, and six 15 ml water samples were taken from the shipment bag. Two replicates were carried out, using as above, 3× 15 ml shipment water per sample (plus a negative extraction control). DNA precipitation, extraction and PCR procedures were also as outlined above, but the DNA precipitation and extraction steps were performed at a separate laboratory to the PCR stage. From the resulting PCRs, a single random product was Sanger sequenced (protocol as Section 2.2.4.2).

6.2.4.3 Relaxed protocol

A further experiment was carried out to test whether these published protocols could be relaxed, and DNA recovered in less time using smaller volumes of reagents, fewer tubes, fewer PCRs, and more portable equipment. The protocol outline above was scaled down into a 1.7 ml EPPENDORF tube, containing 1,000 µl ethanol, 454.5 µl tank water and 45.5 µl of sodium acetate. Samples were incubated at -20°C for only one hour, and centrifuged ($10,000 \times g$) at room temperature on a bench-top EPPENDORF centrifuge (5415D). Water was taken from the *Danio* stock aquarium, with a density of 30 fish in 30 litres of water. DNA extractions and PCR reactions were performed as above, and carried out for both pooled samples (three water samples resulted in one DNA extraction) and not-pooled samples (one water sample resulted in one DNA extraction). The not-pooled experiment was repeated 12 times, with four negative controls from a biologically mature aquarium (fishes, plants, algae, molluscs etc), without the target *Danio* species. The pooled experiment was carried out five times with two of the same negative controls. Three PCR reactions were carried out on each extraction to test if a single PCR would be reliable.

6.3 Results

6.3.1 Sliding window analysis

When the sliding window was set to 100 bp, there was considerable variation in the information content across the COI barcode marker for the *Danio* species analysed (Figure 6.2). Mean genetic K2P distance varied from 7.9% to 18.1% through the windows. The proportion of species with a non-conspecific nearest-neighbour distance of zero varied from 5.5% to 22.0%. The proportion of monophyletic species varied between 47.4% and 73.7%. The optimum window, in terms of information content, started at base pair 531, where the proportion of monophyletic species was maximised, and the proportion of zero non-conspecific nearest-neighbour distances was minimised.

Information content does not, however, always equal suitable priming sites for species specific markers. Assessment of diagnostic nucleotides for *Danio rerio* shows that no species specific nucleotides are present in any windows past 300 bp, despite the higher information content and species discrimination power of that region (Figure 6.3). The highest frequency of diagnostic nucleotides is within the first 100 bases of the barcode marker. Primer design was therefore targeted in this area.

6.3.2 Primer design

Primers for the *Danio rerio* specific eDNA fragment were named eDR3fwd and eDR3rev, and are presented in Table 6.1. Primers were designed manually, and checked for Tm (melting temperature) and GC base content using PRIMER3 with default parameters (Rozen & Skaletsky, 2000). The amplicon comprised a total of 95 base pairs, and starts at position 6,456 through position 6,551 of the *Danio rerio* mitochondrial genome (Broughton *et al.*, 2001).

Table 6.1. Mini-barcode primers generated in this study for species-specific detection of *Danio rerio* using environmental mitochondrial DNA from the COI locus. Resulting amplicon length 95 bp.

Primer name	Direction	Primer sequence 5′–3′	Length (bp)	Tm (°C)	GC (%)
eDR3fwd	Forward	ATCATAAAGACATTGGCACCCTG	23	62.28	43.48
eDR3rev	Reverse	GCTAAGTTCAGCTCGGATTAAG	22	57.52	45.45

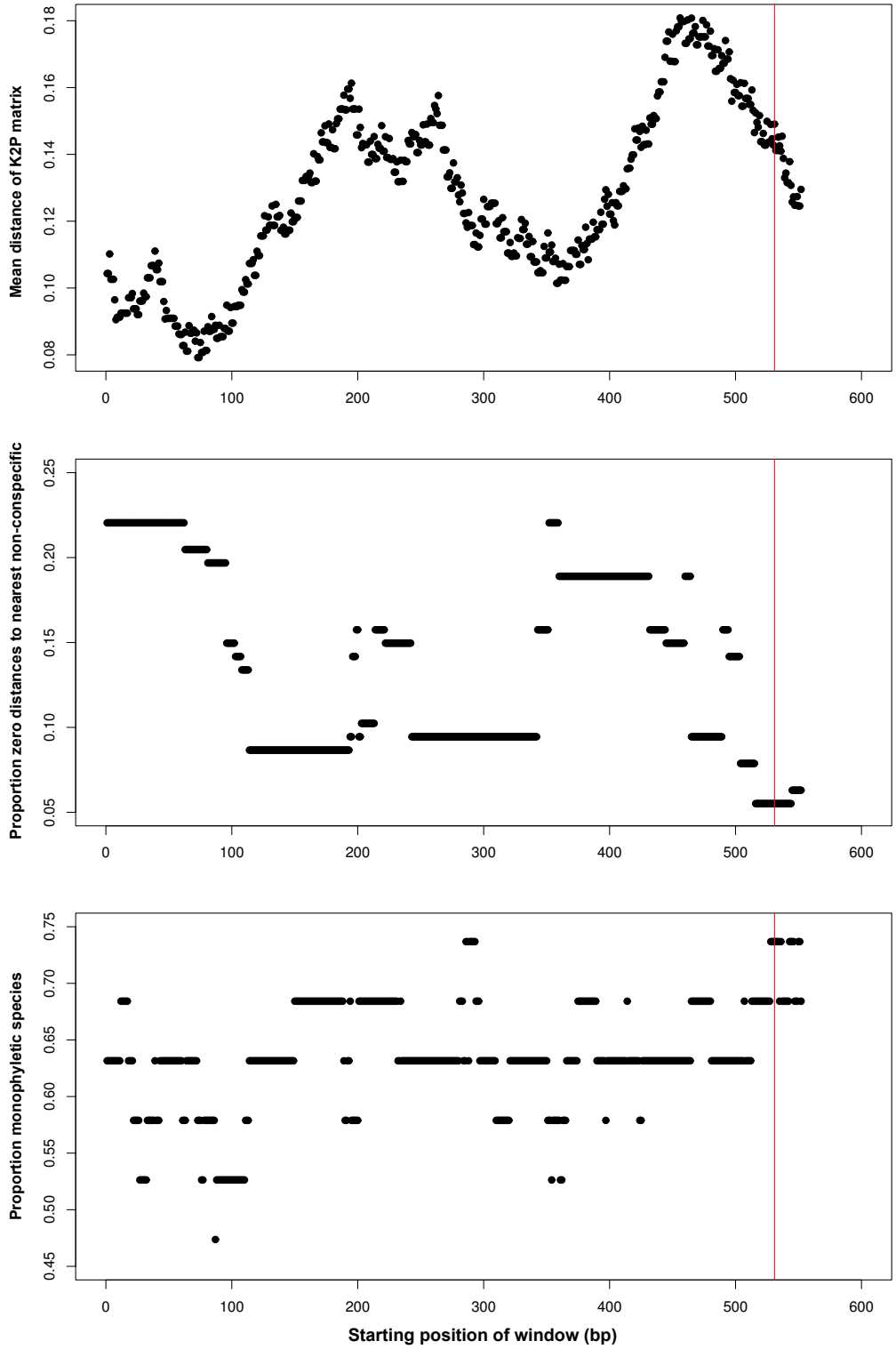

Figure 6.2. Three measures of mini-barcode discriminatory power (mean genetic distance, distance to nearest non-conspecific neighbour, and species monophyly) for a 100 base pair sliding window across the COI barcode marker for the genus *Danio*. Red line illustrates best window for discrimination at position 531.

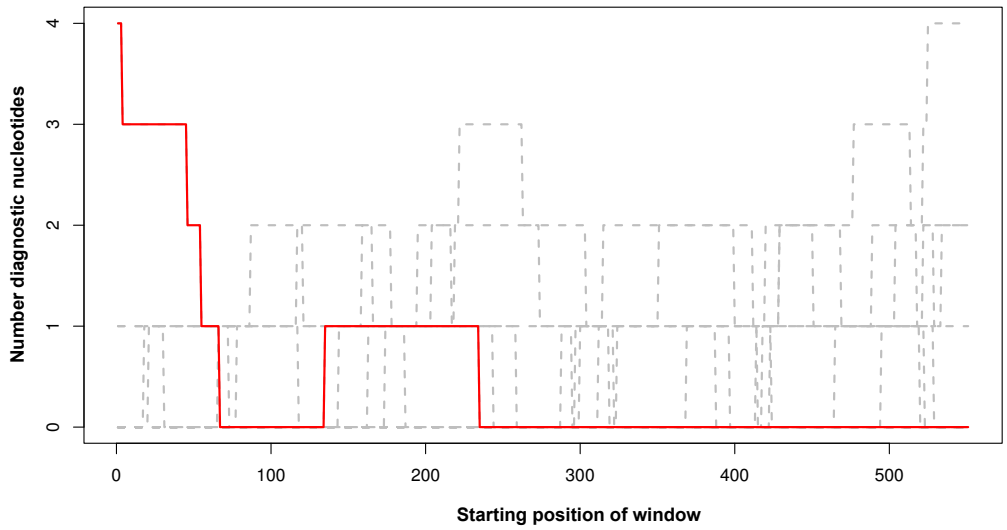

Figure 6.3. 100 base pair sliding window plot of nucleotide diagnostic positions across the COI barcode marker for the genus *Danio*. Red line is *D. rerio*; dashed grey lines are all other species.

6.3.3 Primer specificity

The *in silico* tests of primer specificity using the MFEPRIMER program under default settings made three matches from the local COI database that could potentially produce a PCR product; all three of these were from the target species *Danio rerio*. Under the more stringent settings, two additional matches were found; these were from a South American bird (*Jacamerops aureus*), and a bacterium (*Bacillus pseudofirmus*). The latter was a bacterial genome sequence that satisfied the "COI" search term, but had a PCR product length of 2,304 bp.

The test of specificity using PRIMER-BLAST showed the number of species hits increased as more mismatches were permitted to unintended targets (Figure 6.4). For specified mismatches of no less than four, two of the 129 BLAST hits did not have a mismatch on the terminal 3′ base of either of the primers. This number increased to three for mismatches greater than five. These three species comprised a salamander (*Batrachuperus pinchonii*), and two birds (*Orthotomus sutorius* and *Tolmomyias assimilis*).

For *in vitro* tests of primer specificity, full length DNA barcodes were amplified from all 46 specimens tested from 25 *Danio* and closely related species (Table 6.2). The mini-barcode eDNA primers amplified three individuals tested (RC0679, RC0067 and RC0394). These all corresponded to specimens identified as either *D. rerio* or *D.*

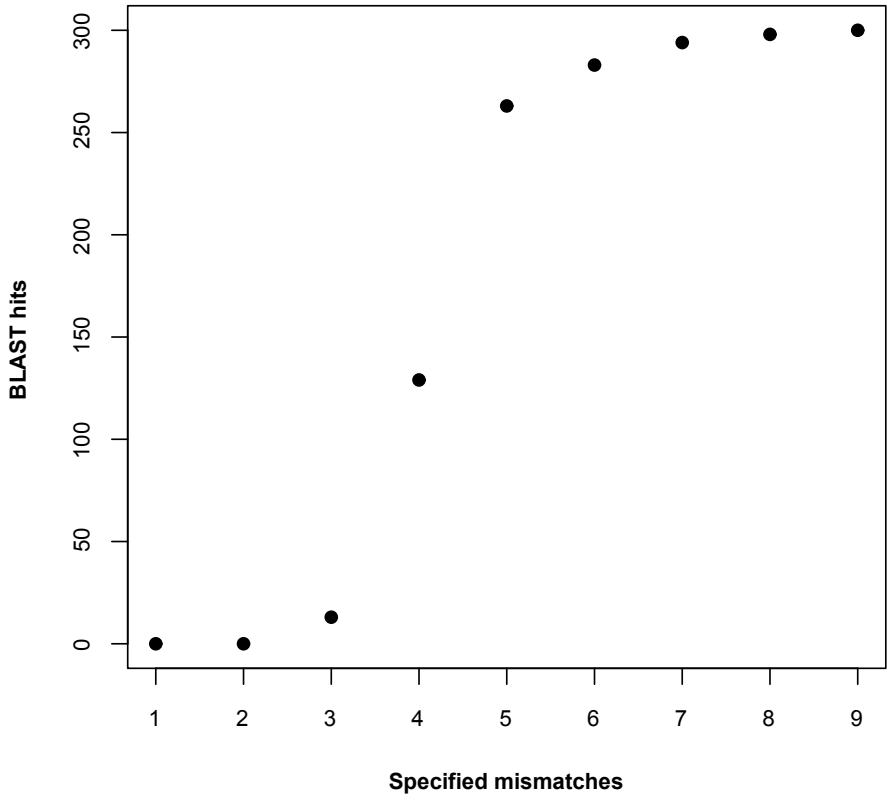

Figure 6.4. PRIMER-BLAST results for eDR3 *Danio rerio* specific primers according to 1–9 specified mismatches within each of the primer pairs. Only hits from unintended (non-*rerio*) targets are shown.

cf. *rerio* (= *D. rerio*). No species other than *D. rerio* were amplified. Figure 6.5 shows an example agarose gel, with only *D. rerio* being amplified.

6.3.4 eDNA detection

6.3.4.1 Experimental treatments

For density treatment A (single fish in four litres of water) a total of 48 PCRs were carried out, with three PCR reactions for each replicate (container with/without fish); 12 PCRs were the negative experimental control (no fish in container). All PCR reactions (three per replicate) were positive for *Danio rerio* (amplicon present of expected length). None of the negative experimental controls showed a band of expected length. Both the positive and negative PCR controls were positive and negative respectively. Results for density treatment B (single fish in twelve litres of water) were identical to treatment one. The subsample of four PCR products for

Table 6.2. PCR specificity reported for 46 specimens of 25 species from the genus *Danio* and other closely related taxa.

Species	Code	BOLD process ID	Barcode PCR	eDNA PCR
Chela dadyburjori	RC0333	RCYY262-11	✓	–
Danio aesculapii	RC0111	RCYY082-11	✓	–
Danio aesculapii	RC0706	RCYY518-11	✓	–
Danio aff. *choprae*	RC0523	RCYY376-11	✓	–
Danio aff. *choprae*	RC0525	RCYY378-11	✓	–
Danio aff. *dangila*	RC0564	RCYY409-11	✓	–
Danio aff. *dangila*	RC0561	RCYY406-11	✓	–
Danio aff. *kyathit*	RC0065	RCYY049-11	✓	–
Danio aff. *kyathit*	RC0121	RCYY092-11	✓	–
Danio albolineatus	RC0076	RCYY057-11	✓	–
Danio albolineatus	RC0445	RCYY327-11	✓	–
Danio cf. *dangila*	RC0343	RCYY272-11	✓	–
Danio cf. *kerri*	RC0267	RCYY224-11	✓	–
Danio cf. *kerri*	RC0270	RCYY227-11	✓	–
Danio cf. *rerio*	RC0679	RCYY501-11	✓	✓
Danio choprae	RC0060	RCYY045-11	✓	–
Danio choprae	RC0164	RCYY129-11	✓	–
Danio choprae	RC0446	RCYY328-11	✓	–
Danio dangila	RC0123	RCYY094-11	✓	–
Danio dangila	RC0345	RCYY274-11	✓	–
Danio erythromicron	RC0599	RCYY433-11	✓	–
Danio erythromicron	RC0705	RCYY517-11	✓	–
Danio feegradei	RC0246	RCYY204-11	✓	–
Danio feegradei	RC0249	RCYY207-11	✓	–
Danio kyathit	RC0090	RCYY066-11	✓	–
Danio kyathit	RC0129	RCYY098-11	✓	–
Danio margaritatus	RC0107	RCYY081-11	✓	–
Danio margaritatus	RC0139	RCYY108-11	✓	–
Danio meghalayensis	RC0567	RCYY412-11	✓	–
Danio meghalayensis	RC0568	RCYY413-11	✓	–
Danio nigrofasciatus	RC0081	RCYY060-11	✓	–
Danio nigrofasciatus	RC0242	RCYY200-11	✓	–
Danio rerio	RC0067	RCYY001-10	✓	✓
Danio rerio	RC0394	RCYY315-11	✓	✓
Danio roseus	RC0126	RCYY095-11	✓	–
Danio roseus	RC0547	RCYY396-11	✓	–
Danio sp. “hikari”	RC0264	RCYY221-11	✓	–
Danio sp. “hikari”	RC0266	RCYY223-11	✓	–
Danio tinwini	RC0062	RCYY046-11	✓	–
Danio tinwini	RC0158	RCYY123-11	✓	–
Devario malabaricus	RC0462	RCYY333-11	✓	–
Devario sondhii	RC0113	RCYY084-11	✓	–
Devario sp. "giraffe"	RC0687	RCYY508-11	✓	–
Esomus metallicus	RC0655	RCYY478-11	✓	–
Microdevario kubotai	RC0492	RCYY354-11	✓	–
Microrasbora rubescens	RC0662	RCYY485-11	✓	–

Notes: ✓= successful PCR amplification (band of expected length apparent).

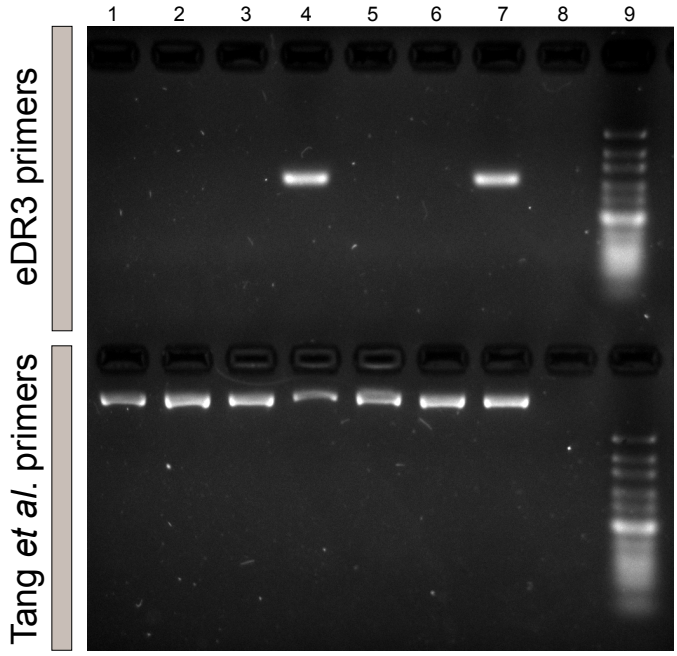

Figure 6.5. A 4% agarose gel showing *Danio rerio* specificity of the eDR3 primers. Top lanes four and seven are tissue extractions of *D. rerio*, and were amplified successfully using the mini-barcode envDR3 primer pair; no other *Danio* species amplified. Bottom lanes are successful PCRs for the same tissue extractions using the full DNA barcode region: primer pair LCO1490A and HCO2198A (Tang *et al.*, 2010). Lane eight was the negative PCR control. Strongest band in the DNA ladder is at 50 bp, while the longest band is at 300 bp.

which sequences were obtained showed clean chromatograms identical to the *D. rerio* mitochondrial genome (NC_002333).

6.3.4.2 Operational testing

The two sets of water samples taken from a shipment bag of *Danio rerio* at the quarantine facility both tested positive for this species in all six PCR reactions. The sequenced PCR product was, again, unambiguously *D. rerio*. The extraction and PCR controls were both negative.

6.3.4.3 Relaxed protocol

For the experiments where protocols were relaxed, three PCRs were also carried out for each replicate. For the experiment where extractions were not pooled, of the 12 replicates, three were positive for a minimum of one PCR reaction out of the three. For the five replicates of the pooled extractions, all five were positive for at least one PCR out of three.

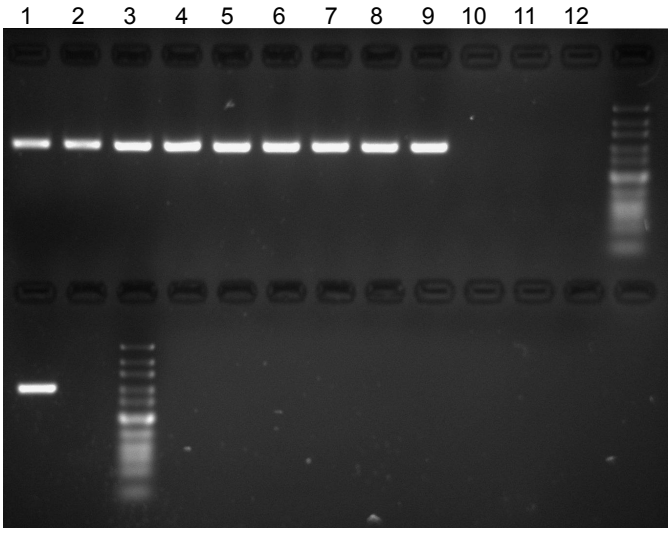

Figure 6.6. A 4% agarose gel showing a single experiment for treatment A (one fish in four litres of water). In the top row of lanes, three PCR reactions were carried out on each of the four containers, and show a positive PCR result of a clean amplicon at the expected length (95 bp) for containers with fish (lanes 1–9). Lanes 10–12 are negative experimental control with no fish present in the container. The bottom row of lanes show a positive PCR control in lane 1 and a negative PCR control in lane 2. Strongest band in the DNA ladder is at 50 bp. The longest band in DNA ladder is at 300 bp.

6.4 Discussion

6.4.1 Primer design and specificity

The sliding window analysis was found to be a useful tool in identifying target regions of DNA alignments for the development of species specific primers. The primers designed here were specific to the target for all *in vitro* PCR reactions of closely related species, and the positive tissue-sample controls showed that stored DNA extractions had not deteriorated below a point where a standard DNA barcode could be amplified. As measured by the *in silico* experiment using both MFEPRIMER and PRIMER-BLAST, there appears to be a low likelihood of non-target amplification, with a small number of hits for well corresponding sequences. As stringency of the PRIMER-BLAST parameters was relaxed, however, the number of potential mis-amplifications increased, but almost all of these had terminal 3′ mismatches. Of course, this conclusion is entirely dependent on the breadth of sequence data present in GenBank, and bias here cannot therefore be entirely avoided.

6.4.2 eDNA detection and sources of error

In both experimental and operational experiments, it has been shown that eDNA can be extracted from aquarium water samples of varying fish densities, and be reliably used to detect the presence of the target species. These densities correspond to those well below the densities at which fishes are typically exported; amplification was successful at fish densities of both 0.08 g/L and 0.24 g/L, while an import of large danios could be up to 40 g/L (300 fish in seven litres water Cole *et al.*, 1999). The technique could therefore be sufficiently sensitive to detect single specimens within mixed shipments.

Due to the sensitive nature of PCR reactions using large numbers of cycles, eDNA monitoring for biosecurity will require a rigorous assay design to ensure confidence in the results (Darling & Mahon, 2011; Willerslev & Cooper, 2005). Tests must be robust to errors, and these errors need to be well understood if the method is to be endorsed for use in management situations where there are political, financial, and legal stakes (Darling & Mahon, 2011). It is also important to distinguish between false positive and false negative errors caused by either the process or the method used (see Fig. 1 of Darling & Mahon, 2011).

Assuming a null hypothesis (H_0) of the target species not being present, a false positive (type I) error will erroneously indicate presence where there is none. A false negative (type II) error will erroneously offer a test result of not present when the species is in fact present. There are trade-offs to be made between the different types of error, and the degree of false positive errors may be a result of the sensitivity of the test and a lack of specificity in the primers. Early detection and monitoring of threats is generally regarded as more cost effective than management of organisms post-invasion (Finnoff *et al.*, 2007; Leung *et al.*, 2002), despite the potential of increased false positives when using sensitive eDNA technologies (Darling & Mahon, 2011). Therefore, the ornamental fish quarantine stage should be regarded as a first line of defence, and certainly false negative results are considered more serious than false positives in terms of potential risk. However, excessive false positives may erode relationships with the aquarium trade.

6.4.2.1 False positive error

There are multiple sources of false positive errors. The most serious of these is perhaps laboratory contamination. Negative controls need to be carried at a high ratio to that of the tests; for ancient DNA (aDNA) work, it is recommended there be

a 1:5 ratio for DNA extractions, and a 1:1 ratio for PCR, due to the irregularity in detecting low level background contamination (Willerslev & Cooper, 2005). Results should always be repeated, lab surfaces and equipment kept decontaminated, and positive controls should also be avoided or used with care (Willerslev & Cooper, 2005). Probe design is also important in preventing false positives through non-target amplification. This can be overcome to some degree by the routine sequencing of PCR products, which would confirm any non-specific priming problems. This should be carried for around 5% of the samples (Darling & Mahon, 2011). *In silico* methods can also be used, as they have here, to assess the likelihood of primers exhibiting this behaviour (Ficetola *et al.*, 2010; Qu *et al.*, 2009).

A sensitive protocol may also detect the presence of target DNA in water when the target organisms are no longer present. This may well occur with imports of aquarium species, as the shipping water may have derived from a source containing target DNA, but the species shipped is a different one. DNA may persist in these kind of environments for up to 30 days (Dejean *et al.*, 2011), so differentiating these two scenarios is important, and while it may appear a problem, is perhaps also a considerable benefit for biosecurity. Knowing whether a shipment has been associated with water from a high risk species would be quite useful in terms of disease risk management. A quantification approach to compare densities of eDNA could be carried out by using either a meta-barcoding approach on for example a 454 pyrosequencing platform, or by using qPCR to allow quantification of DNA concentrations against a fixed standards.

6.4.2.2 False negative error

False negative results may occur when organisms are present in the water, but no eDNA is detected. This may be due to the method being insufficiently sensitive at that concentration of DNA, but improvements in assay sensitivity can be made by further optimising the extraction and PCR techniques (see Rohland & Hofreiter, 2007). Further work could be carried out in evaluating how environmental conditions of the water samples may affect degradation rate of the eDNA at varying concentrations. PCR inhibitors may also be present in the sample, and this could theoretically be possible for densely packed aquarium fish shipments, which may contain metabolites released by the fish in transit, or chemical additives used by fish exporters to remove these metabolites (Cole *et al.*, 1999).

6.4.3 Relaxing protocols

Because eDNA protocols typically require an intensive laboratory procedure, involving time, repetition, and large quantities of reagents, it may be difficult to incorporate into a routine and fast method for biosecurity. Therefore, it was tested whether protocols could be relaxed, both in terms of time, and the volumes of water and reagents required. It was found that when the protocol was scaled into a 1.7 ml EPPENDORF tube with a water sample of 454.5 ml, DNA could be repeatedly isolated from a moderate fish density (0.95 g/L), but only when three samples were pooled. When samples were not pooled, but extracted individually, the likelihood of a successful PCR amplification was lower due to stochastic effects at reduced DNA template concentrations (Willerslev & Cooper, 2005). Repeating the PCR up to nine times did frequently, however, increase the chance of a detection (data not shown), but this perhaps defeats the purpose of a relaxed protocol. When densities of fish are expected to be high, a scaled-down protocol can potentially be incorporated as part of a high throughput routine surveillance system. However, it must be noted that with such an approach, the risk of false negative results is likely to increase due to the likelihood of not recovering sufficient quantities of eDNA from the water.

6.5 Summary

The results here support the usefulness of eDNA as a biosecurity tool for ornamental fishes, and represents a framework for developing the procedure further. The availability of large volumes of COI data from databases such as BOLD, for example, can allow mining of useful new markers for single species or groups of species. As part of the standardised DNA barcode system, these mini-barcodes remain compatible with the voucher specimens and supplementary data associated with those records, adding confidence to identifications. Environmental DNA surveys offer advantages over traditional techniques such as visual examination and barcoding from tissue samples, as they are non-destructive and potentially more sensitive at low population densities of target organisms. Refinement and up-scaling of the method opens up prospects for long term monitoring of entire quarantine facilities or ornamental fish retailers using either meta-barcoding technologies, or mini-barcode microarray systems (Andersen *et al.*, 2012; Hajibabaei *et al.*, 2007).

Chapter 7

Summary and conclusions

Despite the challenge of getting accurate identifications for many of the species collected here, a large database of demonstrably identified fishes and associated barcodes was assembled. For biosecurity applications, relying upon the names provided by aquarium fish suppliers is likely to be highly inaccurate, and therefore DNA barcoding represents not only a defensible approach, but a significant move forward in providing identification tools for aquarium species in biosecurity situations.

For the small percentage of cases where DNA barcodes fail to offer unambiguous identifications, additional data such as Web-based images of live specimens, morphological characters, and nuclear loci can be called upon to resolve these problematic specimens. Benefits from barcoding extend beyond a simple quarantine tool, and provide a basis for the generation of accurate and consistent trade statistics, allowing auditing, record keeping and harmonisation between jurisdictions and agencies (Gerson *et al.*, 2008). Benefits within the ornamental fish industry are also apparent, with accurately identified livestock providing a value added product suitable for export in compliance with international certification or legal standards (Ploeg *et al.*, 2009). Any country vulnerable to aquatic invasions of ornamental species can benefit, with barcode databases offering free and instant access to information. Additional benefits to conservation efforts arise in documenting the ornamental pet trade, with examples such as stock management, traceability, and effective regulation/enforcement of endangered and CITES controlled species (Steinke *et al.*, 2009b).

Development of operational databases such as BOLD rely on solid taxonomic foundations (Dincă *et al.*, 2011; Meyer & Paulay, 2005; Padial *et al.*, 2010), and it is important to note that for identification purposes, molecular data do not circumvent morphology, but merely standardise its application via taxonomic assignments (assuming agreement between morphological and DNA data). In situations where current taxonomy is inadequate, studies such as these support taxonomy in generating new hypotheses as well as adding a suite of fine-scale characters and lab protocols, easily accessible via the Web (Padial *et al.*, 2010). Nuclear data are especially valuable in providing support to the conclusions made from COI data

(Chapter 5; Clare, 2011; Dasmahapatra *et al.*, 2010), can assist in distinguishing hybrids (Chapter 5), and can also be used in species delimitation efforts and interim parataxonomy for diverse complexes of closely-related cryptic-species important in biosecurity (Boykin *et al.*, 2012).

Although the success of DNA barcoding for practical applications depends most importantly upon the accuracy in taxonomic determination of voucher specimens, analytical/bioinformatic methods used to provide the subsequent molecular identifications will also impact how effective the reference libraries can be. A selection of identification criteria were tested in Chapter 3, and success rates were found to differ among methods, sometimes considerably. The "best close match" (BCM) method was justified to be the best when reference libraries are incomplete (as is commonly the case, especially with ornamental fishes). The structure and composition of the reference library was also found to affect identification success, with data from the GenBank repository providing useful extra information, but also a large number of unidentifiable singleton species. In Chapter 4 it was found that the K2P model is not well supported as an evolutionary model in DNA barcode datasets, but misspecification of nucleotide substitution models in estimating genetic distances had little effect on overall rates of specimen identification. These are important findings in terms of understanding appropriate applications and limitations of DNA barcoding in biosecurity.

As demonstrated in Chapter 6, DNA barcode databases can also be used as a data source for developing new techniques in biosecurity. Diagnostic methods are no longer limited to destructively sampling quarantined organisms, or even to the contemporary presence of an organism. Using targeted probes to detect extracellular environmental DNA, high risk species can be detected during routine surveillance of water associated with ornamental fish imports.

Despite the advances and advantages outlined above for using DNA barcodes for biosecurity, challenges remain in being able to make full and confident use of barcode reference libraries. These are outlined below, and are discussed in terms of database management, data analysis, and use within an operational environment.

7.1 Challenges for DNA barcode databases

7.1.1 Incomplete information

Of the main challenges to real-world use of DNA barcoding are the composite problems of incomplete information and conflicting information. It has been shown that where DNA barcode libraries are complete, then the barcodes generally perform well for identification (Chapter 3; Ekrem *et al.*, 2007). Problems occur where queries are not matched with a conspecific in the database (the singleton problem). Here a operator would need to decide if the degree of match will place it with a represented or unrepresented species. In the short term, optimised distance thresholds can be used to determine intra- versus interspecific variation, but more sophisticated techniques such as those using fuzzy-set-theory, for example, should eventually be adopted (e.g. Zhang *et al.*, 2012). Ultimately, however, the most effective approach is to actually sample these missing species (Ekrem *et al.*, 2007).

Unfortunately, the ability to build upon current reference libraries is significantly hampered due to difficulties in accessing specimens, and for the species that are available, problems exist in accessing taxonomic literature for their accurate identification (Section 2.4.1; Monbiot, 2011; Taylor, 2012). Despite the ongoing digitisation efforts of organisations such as the Biodiversity Heritage Library, many of the required publications are hidden in obscure, old journals, or the modern treatments are published in highly specialised journals that few institutional libraries have electronic or even hard-copy access to. Ornamental fishes have a range almost throughout the world's tropics and subtropics, so informative literature can rarely be obtained from a single museum library. As outlined in Chapter 2, considerable effort was undertaken here to obtain scientific literature for cyprinid fishes. Given these problems, the prospects for an organisation such as MAFBNZ to be able to extend this barcoding approach to all ornamental taxa exported to New Zealand are poor[1]. DNA barcoding, is however, a global effort, and other laboratories together with initiatives such as FISH-BOL may be able to take up a lot of this slack (but see below). Unfortunately, freshwater fishes in Africa, Asia, and South America have been very poorly sampled by FISH-BOL (Becker *et al.*, 2011), but these are precisely the regions where aquarium fishes are derived.

[1]It is important here to note an obvious point: the problem of accessing taxonomic literature may prove an equally significant problem for any biosecurity agency wishing to identify fishes using morphological or visual methods.

7.1.2 Conflicts due to misidentifications

Of the most serious limitations to barcoding as an applied resource for regulation and molecular diagnostics, is not necessarily biological problems associated with mitochondrial DNA (e.g. numts, heteroplasmy, symbionts, introgression, paraphyly), but rather human error and uncertainty in creating and curating reference libraries. Becker *et al.* (2011) identify this as the primary source of error in FISH-BOL data. Conflicting identifications can be made when multiple labs are working on the same taxa, and in the process of their morphological identifications are ascribing different taxonomic names to the same species. As a case in point, any biosecurity official wanting to identify tissue from a *Danio rerio* sample—this species comes in a multitude of selectively bred phenotypes under many different trade names—will be unable to, using the current BOLD system. The problem here is that when BOLD 3.0 is queried using a default database search with a *D. rerio* sequence (28/01/12; URL: http://v3.boldsystems.org/index.php/IDS_OpenIdEngine), the system reports that "A species level match could not be made, the queried specimen is likely to be one of the following: *Danio rerio*, *Danio* cf. *rerio*, *Danio* sp., *Brachydanio froskei*, *Brachydanio rerio*.". Given that as a model organism, and of all 40,000+ fish species, *D. rerio* is arguably the one most studied scientifically, this is perhaps surprising and worrying. So, based on this information, an operator would have to make the decision of either destroying the shipment, or taking the time to attempt to resolve the ambiguity, thereby defeating the point of a fast, universal, and reliable identification system.

Overall, prospects for a universal identification system do not appear to be any better. In an analysis of the BINs (Barcode Index Numbers)—BOLD's as yet unpublished interim taxonomic and identification system—for the sequences generated in this work (BOLD project RCYY), a total of 54 BINs contain data from other, external projects (13/02/12; URL: http://v3.boldsystems.org/). Of this total, 19 (35%) contain more than one species name, and BOLD would be therefore unable, again, to offer a species level identification. Most of these discrepancies appear to be misidentifications, and indicates the severity of the potential problem. It is important to note, again, that because many records remain in private BOLD projects, the conflicting data described above were not available for direct comparison in this study. Therefore, the relatively few conflicts observed between the data partitions in Chapter 2 and Chapter 3, may be misleading.

There are currently few safeguards against a BOLD contributor misidentifying a specimen, and once a name has been added into a database, it may be difficult

for a third party to demonstrate that it should be changed. An important asset to the standardised barcoding protocol is the maintenance of records, supporting information, and importantly vouchers—this is what sets BOLD apart from GenBank (Ratnasingham & Hebert, 2007). A new feature of BOLD 3.0 is a wiki-like framework for community based annotation of barcode data (Ratnasingham & Hebert, 2011). However, pre-emptive solutions are perhaps a better use of time. To this effect, a system of identification confidence has been proposed, which rates identifications according to the degree of expertise and effort made in their generation (Steinke & Hanner, 2011). This will encourage data managers to be increasingly diligent about how identifications are generated and justified. The importance of accurate identification is obvious (Bortolus, 2008), and providing a bibliography of reference material and morphological characters used for identification should be mandatory for publication; these additional data may be extremely valuable in correcting mistakes without recourse to the effort of loaning and re-examining voucher material.

An extension of this would be to question whether the identifications made in this study are correct? This is an important question regarding the reliability of using the library created here as an operational barcoding tool, and should certainly be tested empirically in collaboration with independent, expert taxonomic specialists.

7.2 Challenges for DNA barcode analyses

Despite the broad benefits that DNA barcoding can bring to non-systematic endeavours such as food product regulation, conservation, and investigating species interactions, many of the principles inherent to DNA barcoding are based on those of systematic biology; it is here that shortcomings of the experimental design and analytical procedures inherent in some of the DNA barcoding literature are apparent. Most of these concerns have been raised previously in the literature (see references below), but should nevertheless be reiterated due to the repercussions of biosecurity decisions, and the possibility of DNA barcode data becoming admissible evidence in wildlife crime cases (Alacs *et al.*, 2010; Linacre & Tobe, 2011).

The main concern is over the goal of DNA barcoding (DeSalle, 2006; Goldstein & DeSalle, 2011; Moritz & Cicero, 2004; Rubinoff *et al.*, 2006; Taylor & Harris, 2012). Here, it is acknowledged that DNA barcoding can comprise two distinct aims: (1) specimen identification, i.e. assigning taxonomic names to unknown specimens using a DNA reference library of morphologically pre-identified vouchers (Schindel &

Miller, 2005); and (2) species discovery, i.e. a triage tool for sorting new collections into species-like units (Schindel & Miller, 2005). These aims are uncontroversial, provided that they are clearly defined. However, several authors have raised repeated concerns regarding the blurring of these boundaries (e.g. DeSalle, 2006; DeSalle *et al.*, 2005; Goldstein & DeSalle, 2011; Meier, 2008; Vogler & Monaghan, 2007), and it seems impossible to separate these objectives in many examples from the barcoding literature. This provides the basis for many of the criticisms outlined below.

7.2.1 The use of the term "species identification"

The term "species identification" is ubiquitous in the DNA barcoding literature, but this terminology is misleading, and reflects a long-standing confusion between the two sub-disciplines of DNA barcoding (specimen identification vs. species discovery; see above). Here, "species identification" is interpreted as shorthand for: identification of biological material—a specimen—to the level of species. However, it can also be seen in terms of identifying groups of species-like units, i.e. species discovery and delimitation (as used in Ferguson, 2002). One way to minimise this confusion and to clarify the distinct role of each of the two separate objectives, is to use the terms "specimen identification" or "species discovery" in place of "species identification", as appropriate. This more objectively states what hypotheses are being tested, and better ensures that identification is not confused with delimitation. Both of these aims fall within the purview of DNA barcoding, but they should be clearly distinguished as they require different methodological and analytical approaches.

7.2.2 Failure to set clear hypotheses

Perhaps one of the most problematic areas in many barcoding studies is the lack of clearly stated, objective hypotheses. A "typical" barcoding study (e.g. "DNA barcoding the [insert taxon] of [insert geographic region]") aims to: (1) assemble a reference library with specimens identified to species using morphological characters; (2) test how effective this library is for identification purposes; and then (3) explore previously unrecognised diversity apparent in the DNA barcodes. However, it is in regard to these three steps that there is often confusion in how hypotheses are generated and tested. Too frequently, objectives 2 and 3 are conflated, and methodological approaches do not appear to reflect these different goals (Goldstein

& DeSalle, 2011; Meier, 2008). Analytical techniques presented in many studies do not explicitly set out to test identification success (objective 2) by simulating a quantified identification scenario. Rather, they tend to employ the same method (usually a neighbour-joining tree) to test both objectives 2 and 3, and usually present a descriptive rather than analytical summary of the data. If the data collected are intended to be used as an identification tool, then they should be tested as such. Studies should define each objective more clearly in the methods section of the work, and explicitly separating the experimental procedures used to achieve each aim.

7.2.3 Inappropriate use of neighbour-joining trees

Almost all DNA barcoding studies present a neighbour-joining (NJ) tree, and perhaps as a graphical summary of the data can be considered appropriate (but see Goldstein & DeSalle, 2011). However, problems occur when NJ trees are presented as the sole analytical method, and when identification rates from the NJ trees are not quantified (Little & Stevenson, 2007). It has been well documented, both empirically and theoretically, that NJ trees perform poorly for specimen identification purposes (Little, 2011; Meier *et al.*, 2006; Virgilio *et al.*, 2010; Zhang *et al.*, 2012). It is important to note at this point that problems with NJ trees are not resolved by using any other tree inference method such as maximum likelihood or parsimony. The problem is with relying on phylogeny—and specifically the strict monophyly of mtDNA lineages—as an identification criterion.

Few species concepts require reciprocal monophyly (Meier, 2008), and in any case, monophyly is often an unrealistic scenario in closely related groups (Funk & Omland, 2003; Zhang *et al.*, 2012). Tree-based methods offer no assessment of possible group membership in the presence of incomplete taxon sampling (but see Ross *et al.*, 2008), and frequently resolve closely related taxa incorrectly (Lowenstein *et al.*, 2010). Furthermore, when conspecifics are not present in the reference library, tree-based methods are unable to provide the desired "no identification" result, and in the case of recently diverged paraphyletic species, will often result in ambiguous or incorrect identifications.

Despite the popularity and intuitiveness of NJ trees, identification success generally improves when using more accurate techniques, which are usually based directly on the genetic distance matrix. The single "best close match" method has been shown to be reliable, predictable, computationally tractable, and able to make identifications even in the presence of paraphyly (Chapter 3; Meier *et al.*, 2006). Alternatively, many

other criteria are also available for measuring identification success (see Casiraghi *et al.*, 2010), and comparisons of performance between some of these have already been made (Austerlitz *et al.*, 2009; Little & Stevenson, 2007; Meier *et al.*, 2006; Ross *et al.*, 2008; Virgilio *et al.*, 2010; Zhang *et al.*, 2012). It is important to note, however, that a quantification of monophyly still remains a useful description of the data, and should still be used in conjunction with other methods.

Ultimately, phenetic (similarity) methods using genetic distances may be regarded as something of a stop-gap solution. In the near future, the problem of accurately assigning identifications is likely to be addressed by either likelihood-based information-theoretic approaches, or machine learning and statistical tools, such as supervised classification and pattern recognition (e.g. Austerlitz *et al.*, 2009; Zhang *et al.*, 2008). A newly developed fuzzy-set-theory technique (Zhang *et al.*, 2012) appears promising, offering a group membership parameter that provides additional information lacking in threshold-based implementations. Bayesian MCMC coalescent methods promise similar advantages, but may be too computationally inefficient in their current incarnations (Zhang *et al.*, 2012).

In some cases, character-based methods using diagnostic nucleotide combinations may be preferable (DeSalle, 2007), and this is particularly the case for small groups of closely related taxa where similarity methods perform poorly (e.g. Lowenstein *et al.*, 2009). However, character based approaches such as those implemented in the CAOS software (Sarkar *et al.*, 2008), have yet to be fully characterised in terms of their sensitivity to taxon sampling and homoplasy, and are therefore at present perhaps limited to restricted cases (Kerr *et al.*, 2009a). The use of discrete characters could be seen in terms of "DNA barcoding 2.0", potentially offering additional benefits after sampling is extended beyond simply collecting baseline data.

7.2.4 Inappropriate use of bootstrap resampling

The use of bootstrap resampling in DNA barcoding studies typifies the confusion between species discovery and specimen identification. When using DNA barcodes for species discovery—a "molecular parataxonomy" process analogous to sorting specimens into morphospecies (Brower, 2006)—it is required that there is a test of distinctiveness. The bootstrap, along with reciprocal monophyly, is one method among many that can be used to test whether groups (i.e. species-like clusters), are well supported. Bootstrapping in this situation also helps address problems with NJ trees such as taxon-order bias and tied trees (Lowenstein *et al.*, 2009; Meier, 2008).

However, the use of bootstrapping for specimen identification is somewhat perplexing. The aim of DNA barcoding is to maximise congruence with *a priori* defined species, *viz*. the taxonomic names from a morphological identification process. A species with low bootstrap support does not falsify a species hypothesis when this assessment is based on independent data (i.e. morphology from the original description). In many cases, recently diverged sister species on short branches will have low support and therefore fail to be identified, even if they are morphologically distinct *and* diagnosable by unique mutations (Lowenstein *et al.*, 2009). Thus, using a bootstrap value as a cut-off for correct identification severely compromises the efficacy of a reference library (Chapter 3; Zhang *et al.*, 2012), and exacerbates the previously outlined weaknesses of using tree-based methods in general. On top of this, bootstrap resampling does not make an assessment of the uncertainty in identification; an unknown can group with a taxon at 100% bootstrap support, and yet be an entirely different species. Perhaps a better way to measure uncertainty in identification is to calculate group membership probabilities (e.g. Zhang *et al.*, 2012), and to make explicit "caveats in relation to the breath of sampling" (Moritz & Cicero, 2004).

7.2.5 Inappropriate use of fixed distance thresholds

The use of distance thresholds has been extensively debated (Chapter 1; Puillandre *et al.*, 2012; Virgilio *et al.*, 2012; Zhang *et al.*, 2012), but in the context of providing an overview of the challenges for DNA barcoding, the aim here is to re-emphasise these points already made. A threshold is essential when identifying specimens using genetic distance data; in the absence of complete sampling, distance thresholds aim to minimise misidentifications of unknowns that do not have conspecifics represented in the reference library (Virgilio *et al.*, 2012). However, there is no *a priori* reason to assume a universal threshold is applicable, as coalescent depths among species will vary considerably due to differences in population size, rate of mutation, and time since speciation (Monaghan *et al.*, 2009).

A generic threshold such as 1% is perhaps not an unreasonable heuristic in some cases (e.g. Chapter 3), but it can be considered arbitrary, and is likely to suffer from varying rates of false positive and false negative error, depending on the data. Rather than relying on prescribed cut-offs, optimised thresholds can be generated directly from the data itself (Meyer & Paulay, 2005; Virgilio *et al.*, 2012). Computer programs or protocols are now available to calculate optimised thresholds, and for species

discovery, these can even be generated in the absence of taxonomic names (Brown *et al.*, 2012; Puillandre *et al.*, 2012; Virgilio *et al.*, 2012).

7.2.6 Use of the K2P model

As outlined in Chapter 4, DNA barcoding studies use Kimura's two-parameter substitution model (K2P) as the *de facto* standard for constructing genetic distance matrices. Distances generated under this model then provide the basis for most downstream analyses, but uncertainty in model choice is rarely explored and could potentially affect how reliably DNA barcodes discriminate species. This is an important question, as the K2P model is so widely used, and assumed to be correct.

Chapter 4 shows that the K2P is a poorly fitting model at the species level; it was never selected as the best model, and very rarely selected as a credible alternative model. Despite the lack of support for the K2P model, differences in distance between best model and K2P model estimates were usually minimal, and importantly, identification success rates were largely unaffected by model choice even when interspecific threshold values were reassessed. Although these conclusions may justify using the K2P model for specimen identification purposes, simpler metrics such as *p* distance performed equally well, perhaps obviating the requirement for model correction in DNA barcoding. Conversely, when incorporating genetic distance data into taxonomic studies, a more thorough examination of model uncertainty is advocated.

7.2.7 Incorrectly interpreting the barcoding gap

The barcoding gap as proposed by Meyer & Paulay (2005) can represent two distinct scenarios: one for specimen identification (an individual being closer to a member of its own species than a different species), and one for species discovery (a distance that equates to a threshold applicable to all species; see Figure 7.1). The two scenarios are frequently confused, and this again demonstrates conflation of the two objectives of DNA barcoding.

Many DNA barcoding studies present histograms showing frequency distributions of both intra- and interspecific divergences for all pooled species analysed in a study. Overlap between the two distributions can be interpreted as a failure of DNA barcoding, but the only failure demonstrated in this case is that of defining a universal cut-off value. In this regard, and as stated previously, it is widely acknowledged that

coalescent depths vary among species, and substantial overlap between intra- and interspecific distances may be the rule, rather than the exception (Virgilio *et al.*, 2010). Therefore, for specimen identification purposes this type of presentation is wholly uninformative, as intraspecific distances for one species can exceed interspecific distances for other species in the analysis, but without compromising identification success.

A better display of distance data for specimen identification is a dotplot in which, for each individual in the dataset, the distance to the furthest conspecific is plotted against the distance to the nearest non-conspecific, with a 1:1 slope representing the point at which the difference between the two is zero (i.e. no barcoding gap). An example of this method is illustrated in Figure 2.2.

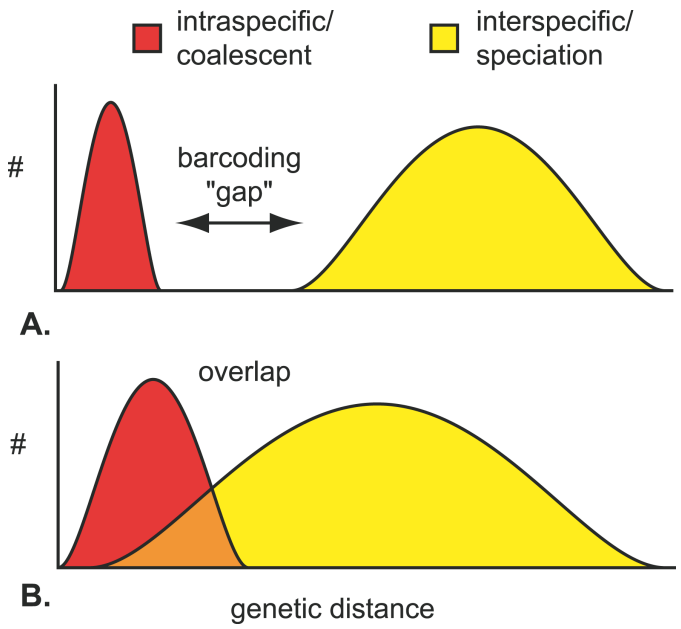

Figure 7.1. An illustrative barcoding gap, showing no overlap (A), and substantial overlap (B) between intraspecific and interspecific variation. This shows how distances are considered overall, but is not informative for specimen identification purposes. Figure copyright (Meyer & Paulay, 2005).

7.2.8 Improving analytical procedures

In conclusion, more care should be taken in setting clear hypotheses for barcoding studies, and choosing appropriate methods for answering each distinct question. Future barcoding studies should make more use of alternative methods, and push forward improvements in data analysis. One possible problem identified in the

limited uptake of many of these methods, has been due to a lack of platform to carry out these analyses (Sarkar & Trizna, 2011). Comparison between different methods is important, and fortunately now increasingly possible in universal open-source environments such as R language, which should supersede the current inflexible and piecemeal software applications (Freckleton, 2009). This will ultimately encourage better use, sharing and benchmarking of new techniques between labs. The publication of the R package SPIDER (Brown *et al.*, 2012), as part of this thesis helps to address this.

7.3 Challenges for biosecurity

7.3.1 Import Health Standard

One potential source of confusion when implementing a DNA barcode reference library such as the one generated in this study, is the discrepancy in names between the identified voucher specimens in the DNA barcode reference libraries, the Import Health Standard (IHS) list of permitted species (MAF Biosecurity New Zealand, 2011), and the trade literature. Some species commonly traded under a well known scientific names may not actually belong to that taxon. Therefore, enforcement of the current names on the IHS may prevent assumed-to-be benign species that are already present in the country from entering the country in future, and could perhaps more worryingly, allow new imports of species that have potentially never been in the country. As follows are several examples of where problems may occur, but it is important to note that the fishes discussed were purchased from the trade in several locations (UK, NZ and Singapore), and comments are based on anecdotal observations of traded species and trade names in these countries, and not just for New Zealand. The IHS status of the fishes collected in the study and any common trade misidentifications, are listed in Appendix C.

A very commonly sold fish in the aquarium trade, the Siamese algae eater "*Crossocheilus siamensis*" (Smith), is a junior subjective synonym of *Crossocheilus oblongus* Kuhl & van Hasselt. Both of these names are listed on the IHS, but *C. oblongus* was not present in this survey of the trade (Chapter 2). All fishes purchased in the trade during this study as *C. siamensis*, were according to morphological features more likely to be *C. langei*, *C.* cf. *atrilimes* or *Garra cambodgiensis* (Appendix C). None of these species are listed on the IHS, and it is possible that *C. oblongus* is rare in the trade and has scarcely been exported.

This may not be an isolated incidence, however, as a similar general pattern was observed across several genera and for several commonly traded species. For example: tinfoil barbs often sold under the name *Barbonymus schwanenfeldii* were frequently *B. altus* (a species not on the approved IHS list); the "arulius" barb named on the IHS was more likely to be *Puntius tambraparniei* rather than *P. arulius*, and so the fishes sold in the trade under this latter name are not therefore listed on the approved IHS list; the fish sold as the clown barb *P. everetti* was more likely to be *P. dunckeri* (not on the approved IHS list); imports of *P. lineatus* were *P. johorensis* (not on the approved IHS list); and the ticto barb "*P. ticto*" was most frequently either *P. stoliczkanus* or *P. padamya* (neither are on the approved IHS list).

Many species not listed on the IHS may also be sold as, or mixed with, species otherwise approved on the IHS list. For example, fishes sold as *Puntius gelius* were often a mixture of bona fide *P. gelius*, and a likely undescribed and not listed as approved *Puntius* (*P.* aff. *gelius*); shipments of *Danio kyathit* may be the more common but undescribed species *D.* aff. *kyathit*, rather than genuine *D. kyathit*; the filament barb *P. filamentosus* can comprise exports of both this species and the not listed as approved *P. assimilis*; and *Devario aequipinnatus* exports were usually *D. malabaricus* (both species are listed as approved on the IHS, however).

There are also scenarios where names have changed due to recent taxonomic work. An example of the latter is *Danio* sp. "pantheri", a species named on the IHS, but now described as *D. aesculapii* (not listed as approved on the IHS). It shows that maintaining a link between these names and keeping up-to-date with taxonomic progress is important, if moving away from qualitative visual identifications to a repeatable system based on often third-party-generated data from DNA barcode reference libraries and vouchered museum specimens. This requires a more adaptable and flexible solution to respond to changing nomenclature, trade patterns and scientific progress.

The current list could perhaps be re-evaluated in light of the problems highlighted above. There are no reasons to assume these discrepancies are limited to the Cyprinidae. Groups such as the loricariid and callichthyiid catfishes are very poorly known taxonomically, and the staggering number of *nomina nuda* listed on the IHS for this latter group suggests a high likelihood of mistaken identities. Due to the plasticity in trade patterns, there is every reason to assume that the species listed above as potentially permitted misidentifications will appear, and therefore be erroneously allowed. This was the case with the arulius barb, known for decades in

the trade as *P. arulius*, until a new species was imported, and the true identities of *P. arulius* and *P. tambraparniei* became known (Ford, 2011).

7.3.2 Risk assessment

Assessment of risk from the ornamental fish trade can be seen in terms of both disease vectoring and of the potential pest status of the fishes themselves (Section 1.1). Although the majority of concern is based upon the risk of the former (Hine & Diggles, 2005), an accurate assessment pertaining to the latter may remain important. Previous management decisions were based upon the best information available at the time, but the potential climate match information for species' invasibility was based upon highly questionable, subjective, and unreferenced data derived from aquarium literature (McDowall & James, 2005). Risk assessment techniques for potentially harmful species using climate modelling and occurrence data have improved since (Hulme, 2012). Based on the Australian Weed Risk Assessment (see McGregor *et al.*, 2012), the Fish Invasiveness Scoring Kit, FISK (Copp *et al.*, 2005, 2009), applies common criteria to prediction of potential problem species. Applying this method to aquarium imports would therefore refine the current IHS list, identify harmful species with a better degree of accuracy, and potentially result in more species being available to the aquarium hobby.

7.3.3 Identification procedures using DNA barcodes

As outlined in Chapter 3, the probability of getting the correct identification for a given query sample can vary according the technique employed, and several other studies have reached the same conclusion using various algorithms under different scenarios (Austerlitz *et al.*, 2009; Little & Stevenson, 2007; van Velzen *et al.*, 2012; Virgilio *et al.*, 2010). The methods outlined and critiqued in the previous section relate to making an academic comparison and assessing empirical support for conclusions as to the effectiveness of a barcode library, but operational considerations should also be taken into account. Ease of use is important, especially when biosecurity officials rather than bioinformaticians are conducting the analyses.

Available online, BOLD-IDS natively uses the most up-to-date reference library, therefore a fresh database version does not need to be downloaded each time a query is made locally. All that is required is that the query sequence is pasted into the browser, and then a species level result is returned on screen. It must be noted,

however, that BOLD will return a higher proportion of ambiguous identifications than other methods tested here (see Chapter 3). A case in point being the differentiation of *Danio albolineatus* from *D. roseus* (Chapter 2). Both are very similar in terms of morphology (Figure 7.2), both are common in the aquarium trade, but unlike *D. albolineatus*, *D. roseus* is not listed on the IHS for import into New Zealand. Telling them apart is therefore important, and this is the kind of problem DNA barcoding was promoted as being able to resolve (Hebert *et al.*, 2003b). Data presented here show that they are indeed closely related, and polyphyletic (Section B.3). The method used by BOLD is unable to separate the two species and gives an ambiguous result, despite discriminating sites existing between the two species. The single closest match methods (*k*-NN or BCM) identify the two species correctly. If operational strategy prioritises ease-of-use over identification accuracy, it must be accepted that the latter will be compromised.

Where conflicts in identifications arise, and BOLD is unable to provide an unambiguous result, it is also important to assess the competency and thoroughness of the work invested in identifying the vouchers that the DNA barcodes are derived from. As outlined above, there now exists the ability to annotate BOLD records and see the confidence in the identifications (Steinke & Hanner, 2011). These features should be used to their fullest potential, to ensure consistency between community curated data.

7.3.4 Possible future goals

Due to the discrepancies outlined above, and the more general difficulty in identifying many imported fishes, an ongoing monitoring program of aquarium fish imports could be implemented, thereby enabling an informed assessment of risk posed to New Zealand (i.e. exactly which species are being traded). In practice, a monitoring program would involve tissue sampling, and identifying using DNA barcodes, individuals from all cyprinid fish imports into New Zealand. For cyprinid fishes having been DNA barcoded in this study, the data generated here can be used as the basis for the reference library. If the monitoring program were required to be extended beyond cyprinids to all imported fishes, it would be required that before being used as reference material, fishes be first accurately identified using demonstrable morphological characters and appropriate taxonomic literature, rather than aquarium guide books which are frequently incorrect (but see Section 7.1.1 regarding taxonomic literature).

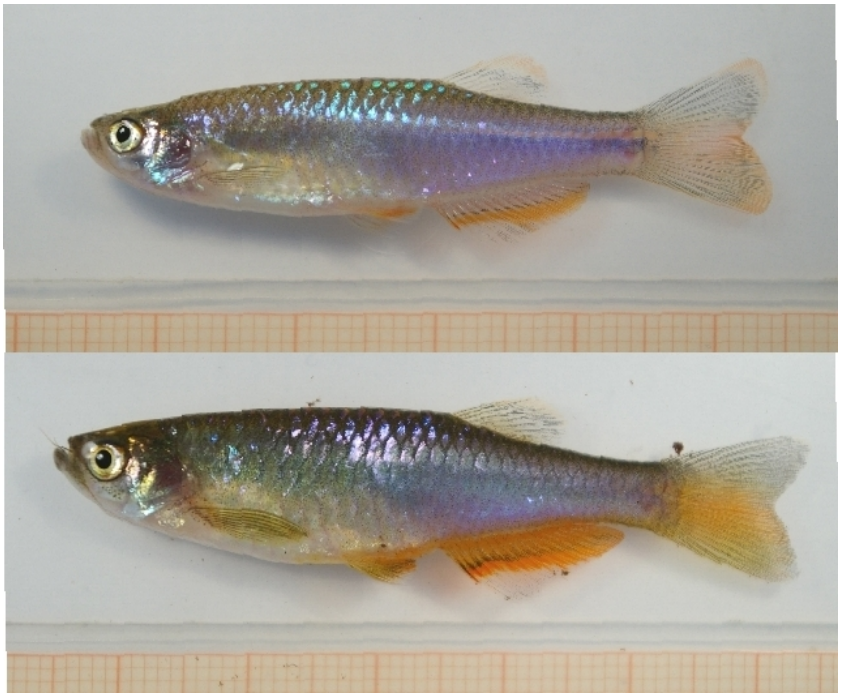

Figure 7.2. Illustration of morphological similarity between the pearl danio, *Danio albolineatus* RC0089 (above), and the rosy danio, *D. roseus* RC0126 (below).

This work could also be carried out in conjunction with an assessment of how effective the reference library compiled for this study actually is in real operational terms, i.e. is it fit for purpose? This would involve sampling from each shipment of cyprinid fish, generating genuine barcode queries, and testing the congruence of names derived from this process against a formal *a posteriori* identification using morphological characters. This would assess the thoroughness of the taxon sampling, the identification power of the DNA barcodes, and the likelihood of encountering unsampled species (Chapter 3). Few studies have conducted this kind of analysis as to the actual end-user benefits of DNA barcoding (Cameron *et al.*, 2006), and this would be a worthwhile study and contribution to the scientific record.

7.4 Concluding remarks

This study provides a comprehensive sampling of the cyprinid fishes in the aquarium trade, together with the publication of reproducible lab protocols to effectively recover DNA barcodes from these fishes. Furthermore, a template is provided for the extension of the library to other groups of problematic ornamentals, especially with

regard to conducting the sampling, storage, and morphological identification. Problems were identified in setting up and using reference libraries, and in particular with regard to a lack of access to taxonomic literature, and the conflict among existing and new barcode data. Nuclear data were found to be useful for detecting interspecific hybrids, and clarifying problems with unrecognised diversity. However, appropriate nuclear sequence data can be difficult to access for species-level identification work, but a comparison among candidates indicated some potentially suitable markers. A critical investigation of some of the widespread assumptions of barcode identification methods was also carried out, and recommendations made as to how best analyse data when conducting future barcoding studies. New diagnostic techniques using traces of environmental DNA in water were also investigated, with this method having the potential to become a powerful tool in the routine detection of high risk species.

References

Abdo, Z. & Golding, G. B. (2007). A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. *Systematic Biology*, 56:44–56.

Ahl, E. (1929). Übersicht über die lebend eingeführten asiatischen arten der gattung *Barbus*. *Das Aquarium*, 1929:165–169.

Alacs, E. A., Georges, A., FitzSimmons, N. N., & Robertson, J. (2010). DNA detective: a review of molecular approaches to wildlife forensics. *Forensic Science, Medicine, and Pathology*, 6:180–194.

Alarcón, J. A. & Alvarez, M. C. (1999). Genetic identification of sparid species by isozyme markers: application to interspecific hybrids. *Aquaculture*, 173:95–103.

Alfaro, M. E. & Huelsenbeck, J. P. (2006). Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Systematic Biology*, 55:89–96.

Alfred, E. R. (1963). Some colourful fishes of the genus *Puntius* Hamilton. *Bulletin of the Singapore National Museum*, 32:135–142.

Ali, A., Raghavan, R., & Dahanukar, N. (2010). *Puntius denisonii*. IUCN Red List of Threatened Species version 2011.1. Accessed 24 August 2011. URL: http://www.iucnredlist.org/.

Ali, B. A., Huang, T. H., Qin, D. N., & Wang, X. M. (2004). A review of random amplified polymorphic DNA (RAPD) markers in fish research. *Reviews in Fish Biology and Fisheries*, 14:443–453.

Aliabadian, M., Kaboli, M., Nijman, V., & Vences, M. (2009). Molecular identification of birds: performance of distance-based DNA barcoding in three genes to delimit parapatric species. *PLoS ONE*, 4:e4119.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410.

Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjaer, K. H., Orlando, L., Gilbert, M. T. P., & Willerslev, E. (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, 21:1966–1979.

Anderson, D. R. (2008). *Model Based Inference In The Life Sciences: A Primer On Evidence*. Springer, New York.

Anderson, I. & Brass, A. (1998). Searching DNA databases for similarities to DNA sequences: when is a match significant? *Bioinformatics*, 14:349–356.

Annandale, N. (1918). Fish and fisheries of the Inlé Lake. *Records of the Indian Museum*, 14:33–64.

Arai, R. & Akai, Y. (1988). *Acheilognathus melanogaster*, a senior synonym of *A. moriokae*, with a revision of the genera of the subfamily Acheilognathinae (Cypriniformes, Cyprinidae). *Bulletin of the National Science Museum, Tokyo, Series A*, 14:199–213.

Armstrong, K. F. & Ball, S. L. (2005). DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1813–1823.

Arunkumar, L. & Tombi Singh, H. (2003). Two new species of puntiid fish from the Yu River system of Manipur. *Journal of the Bombay Natural History Society*, 99:481–487.

Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., Veuille, M., & Laredo, C. (2009). DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics*, 10:S10.

Avise, J. C. (2001). Cytonuclear genetic signatures of hybridization phenomena: rationale, utility, and empirical examples from fishes and other aquatic animals. *Reviews in Fish Biology and Fisheries*, 10:253–263.

Avise, J. C. (2009). Phylogeography: retrospect and prospect. *Journal of Biogeography*, 36:3–15.

Avise, J. C. & Saunders, N. C. (1984). Hybridization and introgression among species of sunfish (*Lepomis*): analysis by mitochondrial dna and allozyme markers. *Genetics*, 108:237–255.

Axelrod, H. R. (1976). *Rasbora brittani*, a new species of cyprinid fish from the Malay Peninsula. *Tropical Fish Hobbyist*, 24:94–98.

Baensch, H. A. & Fischer, G. W. (2007). *Aquarium atlas photo index 1-5*. Mergus Verlag GmbH, Melle, Germany, third edition.

Baldauf, S. L. (2003). Phylogeny for the faint of heart: a tutorial. *Trends in Genetics*, 19:345–351.

Banarescu, P. (1986). A review of the species of *Crossocheilus, Epalzeorhynchos* and *Paracrossochilus* (Pisces, Cyprinidae). *Travaux du Museum d'Histoire Naturelle*, 28:141–161.

Barman, R. P. (1984a). A new freshwater fish of the genus *Danio* Hamilton (Pisces: Cyprinidae) from Assam, India, with the key to the identification of the Indian species of the subgenus *Danio*. *Bulletin of the Zoological Survey of India*, 6:163–165.

Barman, R. P. (1984b). On a new species of the genus *Danio* Hamilton from Burma (Pisces: Cyprinidae). *Bulletin of the Zoological Survey of India*, 5:31–34.

Barman, R. P. (1991). A taxonomic revision of the Indo-Burmese species of *Danio* Hamilton Buchanan (Pisces: Cyprinidae). *Records of the Zoological Survey of India*, 137:1–91.

Barraclough, T. G. & Nee, S. (2001). Phylogenetics and speciation. *Trends in Ecology and Evolution*, 16:391–399.

Bartley, D. M., Rana, K., & Immink, A. J. (2001). The use of inter-specific hybrids in aquaculture and fisheries. *Reviews in Fish Biology and Fisheries*, 10:325–337.

Bauer, A. M., Parham, J. F., Brown, R. M., Stuart, B. L., Grismer, L., Papenfuss, T. J., Böhme, W., Savage, J. M., Carranza, S., Grismer, J. L., Wagner, P., Schmitz, A., Ananjeva, N. B., & Inger, R. F. (2011). Availability of new Bayesian-delimited gecko names and the importance of character-based species descriptions. *Proceedings of the Royal Society B: Biological Sciences*, 278:490–492.

Bazin, E., Glémin, S., & Galtier, N. (2006). Population size does not influence mitochondrial genetic diversity in animals. *Science*, 312:570–572.

Becker, S., Hanner, R., & Steinke, D. (2011). Five years of FISH-BOL: brief status report. *Mitochondrial DNA*, 22 Suppl 1:3–9.

Berra, T. M. (2007). *Freshwater Fish Distribution*. The University of Chicago Press, Chicago.

Bertheau, C., Schuler, H., Krumböck, S., Arthofer, W., & Stauffer, C. (2011). Hit or miss in phylogeographic analyses: the case of the cryptic NUMTs. *Molecular Ecology Resources*, 11:1056–1059.

Bertolazzi, P., Felici, G., & Weitschek, E. (2009). Learning to classify species with barcodes. *BMC Bioinformatics*, 10 Suppl 1:S7.

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., Ingram, K. K., & Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology and Evolution*, 22:148–155.

Boecklen, W. J. & Howard, D. J. (1997). Genetic analysis of hybrid zones: numbers of markers and power of resolution. *Ecology*, 78:2611–2616.

Bonants, P., Groenewald, E., Rasplus, J. Y., Maes, M., De Vos, P., Frey, J., Boonham, N., Nicolaisen, M., Bertacini, A., Robert, V., Barker, I., Kox, L., Ravnikar,

M., Tomankova, K., Caffier, D., Li, M., Armstrong, K. F., Freitas-Astúa, J., Stefani, E., Cubero, J., & Mostert, L. (2010). QBOL: a new EU project focusing on DNA barcoding of quarantine organisms. *EPPO Bulletin*, 40:30–33.

Bordoloi, S. & Baishya, A. (2006). *Puntius ornatus* from the Brahmaputra drainage in Assam. *Zoos' Print Journal*, 21:2292–2294.

Borisenko, A. V., Sones, J. E., & Hebert, P. D. N. (2009). The front-end logistics of DNA barcoding: challenges and prospects. *Molecular Ecology Resources*, 9:27–34.

Bortolus, A. (2008). Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. *Ambio*, 37:114–118.

Bos, D. H. & Posada, D. (2005). Using models of nucleotide evolution to build phylogenetic trees. *Developmental and Comparative Immunology*, 29:211–227.

Boschung, H. T. & Mayden, R. L. (2004). *Fishes of Alabama*. Smithsonian Institution Press, Washington.

Boulenger, G. A. (1907). Descriptions of three new freshwater fishes discovered by Mr. G. L. Bates in South Cameroon. *Annals and Magazine of Natural History*, 20:485–487.

Boyer, S., Brown, S. D. J., Collins, R. A., Cruickshank, R. H., Lefort, M. C., Malumbres-Olarte, J., & Wratten, S. D. (2012). Sliding window analyses for optimal selection of mini-barcodes, and application to 454-pyrosequencing for specimen identification from degraded DNA. *PLoS ONE*, 7:e38215.

Boykin, L. M., Armstrong, K. F., Kubatko, L., & De Barro, P. (2012). Species Delimitation and Global Biosecurity. *Evolutionary Bioinformatics*, 8:1–37.

Brittan, M. R. (1972). *A revision of the Indo-Malayan fresh-water fish genus* Rasbora. T.F.H. Publications, Neptune.

Brittan, M. R. (1976). *Rasbora axelrodi*, a new cyprinid from Indonesia. *Tropical Fish Hobbyist*, 25:92–98.

Britz, R. (2009). *Danionella priapus*, a new species of miniature cyprinid fish from West Bengal, India (Teleostei: Cypriniformes: Cyprinidae). *Zootaxa*, 2277:53–60.

Britz, R., Conway, K. W., & Rüber, L. (2009). Spectacular morphological novelty in a miniature cyprinid fish, *Danionella dracula* n. sp. *Proceedings of the Royal Society B: Biological Sciences*, 276:2179–2186.

Britz, R. & Kottelat, M. (2008). *Paedocypris carbunculus*, a new species of miniature fish from Borneo (Teleostei: Cypriniformes: Cyprinidae). *The Raffles Bulletin of Zoology*, 56:415–422.

Broughton, R. E., Milam, J. E., & Roe, B. A. (2001). The complete sequence of the zebrafish (*Danio rerio*) mitochondrial genome and evolutionary patterns in vertebrate mitochondrial DNA. *Genome Research*, 11:1958–1967.

Brower, A. V. Z. (2006). Problems with DNA barcodes for species delimitation: 'ten species' of *Astraptes fulgerator* reassessed (Lepidoptera: Hesperiidae). *Systematics and Biodiversity*, 4:127–132.

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

Brown, W. M., George, M., & Wilson, A. C. (1979). Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences*, 76:1967.

Brungs, W. A. (1973). Effects of residual chlorine on aquatic life. *Journal (Water Pollution Control Federation)*, 45:2180–2193.

Buckley, T. R. & Cunningham, C. W. (2002). The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Molecular Biology and Evolution*, 19:394–405.

Buhay, J. E. (2009). "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. *Journal of Crustacean Biology*, 29:96–110.

Cameron, S., Rubinoff, D., & Will, K. (2006). Who will actually use DNA barcoding and what will it cost? *Systematic Biology*, 55:844–847.

Casiraghi, M., Labra, M., Ferri, E., Galimberti, A., & De Mattia, F. (2010). DNA barcoding: a six-question tour to improve users' awareness about the method. *Briefings in Bioinformatics*, 11:440–453.

Cawthorn, D. M., Steinman, H. A., & Witthuhn, R. C. (2011). Establishment of a mitochondrial DNA sequence database for the identification of fish species commercially available in South Africa. *Molecular Ecology Resources*, 11:979–991.

Champlot, S., Berthelot, C., Pruvost, M., Bennett, E. A., Grange, T., & Geigl, E.-M. (2010). An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS ONE*, 5:e13042.

Chang, C., Lin, W. W., Shao, Y. T., Arai, R., Ishinabe, T., Ueda, T., Matsuda, M., Kubota, H., Wang, F. Y., & Jang-Liaw, N. H. (2009). Molecular phylogeny and genetic differentiation of the *Tanakia himantegus* complex (Teleostei: Cyprinidae) in Taiwan and China. *Zoological Studies*, 48:823–834.

Chang, C., Shao, Y. T., & Kao, H. W. (2006). Molecular identification of two sibling species of *Puntius* in Taiwan. *Zoological Studies*, 45:149–156.

Chapin III, F. S., Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., Hooper, D. U., Lavorel, S., Sala, O. E., Hobbie, S. E., Mack, M. C., & Diaz, S. (2000). Consequences of changing biodiversity. *Nature*, 405:234–242.

Chen, W. J., Bonillo, C., & Lecointre, G. (2003). Repeatability of clades as a criterion of reliability: a case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. *Molecular Phylogenetics and Evolution*, 26:262–288.

Chen, W. J., Miya, M., Saitoh, K., & Mayden, R. L. (2008). Phylogenetic utility of two existing and four novel nuclear gene loci in reconstructing tree of life of ray-finned fishes: The order Cypriniformes (Ostariophysi) as a case study. *Gene*, 423:125–134.

Chen, X.-Y., Yang, J.-X., & Chen, Y.-R. (1999). A review of the cyprinoid fish genus *Barbodes* Bleeker, 1859, from Yunnan, China, with descriptions of two new species. *Zoological Studies*, 38:82–88.

Clare, E. L. (2011). Cryptic species? patterns of maternal and paternal gene flow in eight neotropical bats. *PLoS ONE*, 6:e21460.

Clarke, M. (2008). Breeder produces Clown loach hybrids. World Wide Web electronic publication. URL: http://www.practicalfishkeeping.co.uk/content.php?sid=1637.

Coad, B. (2010). Freshwater Fishes of Iran. World Wide Web electronic publication. URL: http://www.briancoad.com/SpeciesAccounts/CyprinidaeGarratoVimba.htm.

Cognato, A. I. (2006). Standard percent DNA sequence difference for insects does not predict species boundaries. *Journal of Economic Entomology*, 99:1037–1045.

Cohen, N. J., Deeds, J. R., Wong, E. S., Hanner, R. H., Yancy, H. F., White, K. D., Thompson, T. M., Wahl, M., Pham, T. D., Guichard, F. M., Huh, I., Austin, C., Dizikes, G., & Gerber, S. I. (2009). Public health response to puffer fish (tetrodotoxin) poisoning from mislabeled product. *Journal of Food Protection*, 72:810–817.

Cole, B., Tamaru, C. S., Bailey, R., Brown, C., & Ako, H. (1999). Shipping practices in the ornamental fish industry. *Center for Tropical and Subtropical Aquaculture Publication*, 131:1–22.

Collins, R. A., Armstrong, K. F., Meier, R., Yi, Y., Brown, S. D. J., Cruickshank, R. H., Keeling, S., & Johnston, C. (2012a). Barcoding and border biosecurity: identifying cyprinid fishes in the aquarium trade. *PLoS ONE*, 7:e28381.

Collins, R. A., Boykin, L. M., Cruickshank, R. H., & Armstrong, K. F. (2012b). Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. *Methods in Ecology and Evolution*, 3:457–465.

Conway, K. W. (2005). Monophyly of the genus *Boraras* (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 16:249–264.

Conway, K. W., Chen, W. J., & Mayden, R. L. (2008). The "celestial pearl danio" is a miniature *Danio* (s.s.) (Ostariophysi: Cyprinidae): evidence from morphology and molecules. *Zootaxa*, 1686:1–28.

Conway, K. W. & Kottelat, M. (2011). *Boraras naevus*, a new species of miniature and sexually dichromatic freshwater fish from peninsular Thailand (Ostariophysi: Cyprinidae). *Zootaxa*, 3002:45–51.

Conway, K. W., Mayden, R. L., & Tang, K. L. (2009). *Devario anomalus*, a new species of freshwater fish from Bangladesh (Ostariophysi: Cyprinidae). *Zootaxa*, 58:49–58.

Conway, K. W. & Moritz, T. (2006). *Barboides britzi*, a new species of minature cyprinid from Benin (Ostariophysi: Cyprinidae), with a neotype designation for *B. gracilis*. *Ichthyological Exploration of Freshwaters*, 17:73–84.

Cooper, W. J., Smith, L. L., & Westneat, M. W. (2009). Exploring the radiation of a diverse reef fish family: phylogenetics of the damselfishes (Pomacentridae), with new classifications based on molecular analyses of all genera. *Molecular Phylogenetics and Evolution*, 52:1–16.

Copp, G. H., Garthwaite, R., & Gozlan, R. E. (2005). Risk identification and assessment of non-native freshwater fishes: a summary of concepts and perspectives on protocols for the UK. *Journal of Applied Ichthyology*, 21:371–373.

Copp, G. H., Vilizzi, L., & Gozlan, R. E. (2010). The demography of introduction pathways, propagule pressure and occurrences of non-native freshwater fish in England. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 20:595–601.

Copp, G. H., Vilizzi, L., Mumford, J., Fenwick, G. V., Godard, M. J., & Gozlan, R. E. (2009). Calibration of FISK, an invasiveness screening tool for nonnative freshwater fishes. *Risk Analysis*, 29:457–467.

Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, 13:3021–3030.

Cottle, P. W. (2010). *Danios and Devarios*. Published by Peter W. Cottle, Rochester, UK.

Cunningham, C. W., Zhu, H., & Hillis, D. M. (1998). Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution*, 52:978–987.

Darling, J. A. & Blum, M. J. (2007). DNA-based methods for monitoring invasive species: a review and prospectus. *Biological Invasions*, 9:751–765.

Darling, J. A. & Mahon, A. R. (2011). From molecules to management: adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Environmental Research*, 111:978–988.

DasGupta, B., Konwar, K. M., Mandoiu, I. I., & Shvartsman, A. A. (2005). DNA-BAR: distinguisher selection for DNA barcoding. *Bioinformatics*, 21:3424–3426.

Dasmahapatra, K. K., Elias, M., Hill, R. I., Hoffman, J. I., & Mallet, J. (2010). Mitochondrial DNA barcoding detects some species that are real, and some that are not. *Molecular Ecology Resources*, 10:264–273.

Dasmahapatra, K. K. & Mallet, J. (2006). DNA barcodes: recent successes and future prospects. *Heredity*, 97:254–255.

Dawnay, N., Ogden, R., McEwing, R., Carvalho, G. R., & Thorpe, R. S. (2007). Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Science International*, 173:1–6.

Day, F. (1865). *The fishes of Malabar*. Bernard Quaritch, London.

Day, F. (1870). Notes on some fishes from the western coast of India. *Proceedings of the General Meetings for Scientific Business of the Zoological Society of London*, 1870:369–374.

Day, F. (1875). *The fishes of India; being a natural history of the fishes known to inhabit the seas and fresh waters of India, Burma, and Ceylon*. Bernard Quaritch, London.

de Bruyn, M. D., Parenti, L. R., & Carvalho, G. R. (2011). Successful extraction of DNA from archived alcohol-fixed white-eye fish specimens using an ancient DNA protocol. *Journal of Fish Biology*, 78:2074–2079.

de Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology*, 56:879–886.

Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P., & Miaud, C. (2011). Persistence of environmental DNA in freshwater ecosystems. *PLoS ONE*, 6:e23398.

Deraniyagala, P. E. P. (1930). The Eventognathi of Ceylon. *The Ceylon Journal of Science*, 16:1–41.

DeSalle, R. (2006). Species discovery versus species identification in DNA barcoding efforts: response to Rubinoff. *Conservation Biology*, 20:1545–1547.

DeSalle, R. (2007). Phenetic and DNA taxonomy; a comment on Waugh. *BioEssays*, 29:1289–1290.

DeSalle, R., Egan, M. G., & Siddall, M. (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1905–1916.

Dettai, A. & Lecointre, G. (2005). Further support for the clades obtained by multiple molecular phylogenies in the acanthomorph bush. *Comptes Rendus Biologies*, 328:674–689.

Devi, K. R., Indra, T. J., & Knight, J. D. M. (2010). *Puntius rohani* (Teleostei: Cyprinidae), a new species of barb in the *Puntius filamentosus* group from the southern Western Ghats of India. *Journal of Threatened Taxa*, 2:1121–1129.

deWaard, J. R., Mitchell, A., Keena, M. A., Gopurenko, D., Boykin, L. M., Armstrong, K. F., Pogue, M. G., Lima, J., Floyd, R., Hanner, R. H., & Humble, L. M. (2010). Towards a global barcode library for *Lymantria* (Lepidoptera: Lymantriinae) tussock moths of biosecurity concern. *PLoS ONE*, 5:e14280.

Dincă, V., Zakharov, E. V., Hebert, P. D. N., & Vila, R. (2011). Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe. *Proceedings of the Royal Society B: Biological Sciences*, 278:347–355.

Doi, A. & Taki, Y. (1994). A new cyprinid fish, *Hampala salweenensis*, from the Mae Pai river system, Salween basin, Thailand. *Japanese Journal of Ichthyology*, 40:405–412.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4:699–710.

Drummond, A. J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7:214.

Drummond, A. J. & Suchard, M. A. (2010). Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, 8:1–12.

Dubey, B., Meganathan, P. R., & Haque, I. (2010). DNA mini-barcoding: an approach for forensic identification of some endangered Indian snake species. *Forensic Science International: Genetics*, 5:181–184.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z. I., Knowler, D. J., Lévêque, C., Naiman, R. J., Prieur-Richard, A. H., Soto, D., Stiassny, M. L. J., & Sullivan, C. A. (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews*, 81:163–182.

Duggan, I. C. (2010). The freshwater aquarium trade as a vector for incidental invertebrate fauna. *Biological Invasions*, 12:3757–3770.

Duncker, G. (1904). Die Fische der malayischen Halbinsel. *Mitteilungen aus dem Naturhistorischen (Zoologischen)*, 21:133–207.

Edwards, S. V. & Bensch, S. (2009). Looking forwards or looking backwards in avian phylogeography? A comment on Zink and Barrowclough 2008. *Molecular Ecology*, 18:2930–2933.

Ekrem, T., Willassen, E., & Stur, E. (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution*, 43:530–542.

Elder, J. F. & Turner, B. J. (1995). Concerted evolution of repetitive DNA sequences in eukaryotes. *Quarterly Review of Biology*, 70:297–320.

Elias, M., Hill, R. I., Willmott, K. R., Dasmahapatra, K. K., Brower, A. V. Z., Mallet, J., & Jiggins, C. D. (2007). Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proceedings of the Royal Society B: Biological Sciences*, 274:2881–2889.

Eschmeyer, W. N. (2010a). Catalog of Fishes electronic version. Accessed 03 March 2011. URL: http://research.calacademy.org/ichthyology/catalog/fishcatmain.asp.

Eschmeyer, W. N. (2010b). Marine fish diversity: history of knowledge and discovery (Pisces). *Zootaxa*, 50:19–50.

Fang, F. (1997a). *Danio maetaengensis*, a new species of cyprinid fish from northern Thailand. *Ichthyological Exploration of Freshwaters*, 8:41–48.

Fang, F. (1997b). Redescription of *Danio kakhienensis*, a poorly known cyprinid fish from the Irrawaddy basin. *Ichthyological Exploration of Freshwaters*, 7:289–298.

Fang, F. (1998). *Danio kyathit*, a new species of cyprinid fish from Myitkyina, northern Myanmar. *Ichthyological Exploration of Freshwaters*, 8:273–280.

Fang, F. (2000). A review of Chinese *Danio* species (Teleostei: Cyprinidae). *Acta Zootaxonomica Sinica*, 25:214–227.

Fang, F. (2003). Phylogenetic analysis of the Asian cyprinid genus *Danio* (Telostei, Cyprinidae). *Copeia*, 2003:714–728.

Fang, F. & Kottelat, M. (1999). *Danio* species from northern Laos, with descriptions of three new species (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 10:281–295.

Fang, F. & Kottelat, M. (2000). *Danio roseus*, a new species from the Mekong basin in northeastern Thailand and northwestern Laos (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 11:149–154.

Fang, F. & Kullander, S. O. (2009). *Devario xyrops*, a new species of danionine fish from south-western Myanmar (Teleostei: Cyprinidae). *Zootaxa*, 2164:33–40.

Fang, F., Norén, M., Liao, T. Y., Källersjö, M., & Kullander, S. O. (2009). Molecular phylogenetic interrelationships of the south Asian cyprinid genera *Danio, Devario* and *Microrasbora* (Teleostei, Cyprinidae, Danioninae). *Zoologica Scripta*, 38:237–256.

Federhen, S. (2011). Comment on 'Birdstrikes and barcoding: can DNA methods help make the airways safer?'. *Molecular Ecology Resources*, 11:937–938.

Ferguson, H. W., Morales, J. A., & Ostland, V. E. (1994). Streptococcosis in aquarium fish. *Diseases of Aquatic Organisms*, 19:1–6.

Ferguson, J. W. H. (2002). On the use of genetic divergence for identifying species. *Biological Journal of the Linnean Society*, 75:509–516.

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., Taberlet, P., & Pompanon, F. (2010). An *in silico* approach for the evaluation of DNA barcodes. *BMC Genomics*, 11:434.

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, 4:423–425.

Finnoff, D., Shogren, J. F., Leung, B., & Lodge, D. M. (2007). Take a risk: preferring prevention over control of biological invaders. *Ecological Economics*, 62:216–222.

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

Foran, D. R. (2006). Relative degradation of nuclear and mitochondrial DNA: an experimental approach. *Journal of Forensic Sciences*, 51:766–770.

Ford, M. (2011). *Puntius tambraparniei*—Arulius Barb. World Wide Web electronic publication. URL: http://www.seriouslyfish.com/profile.php?genus=Puntius&species=tambraparniei&id=1075.

Fowler, H. W. (1934). Zoological results of the third De Schauensee Siamese Expedition, Part V.–Additional fishes. *Proceedings of the Academy of Natural Sciences of Philadelphia*, 86:335–352.

Fowler, H. W. (1935). Zoological results of the third De Schauensee Siamese Expedition, Part VI.–Fishes obtained in 1934. *Proceedings of the Academy of Natural Sciences of Philadelphia*, 87:89–163.

Francis, C. M., Borisenko, A. V., Ivanova, N. V., Eger, J. L., Lim, B. K., Guillén-Servent, A., Kruskop, S. V., Mackie, I., & Hebert, P. D. N. (2010). The role of DNA barcodes in understanding and conservation of mammal diversity in Southeast Asia. *PLoS ONE*, 5:e12575.

Freckleton, R. P. (2009). The seven deadly sins of comparative analysis. *Journal of Evolutionary Biology*, 22:1367–1375.

Fregin, S., Haase, M., Olsson, U., & Alström, P. (2012). Pitfalls in comparisons of genetic distances: A case study of the avian family Acrocephalidae. *Molecular Phylogenetics and Evolution*, 62:319–328.

Freyhof, J. & Herder, F. (2001). *Tanichthys micagemmae*, a new miniature cyprinid fish from Central Vietnam (Cypriniformes: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 12:215–220.

Frézal, L. & Leblois, R. (2008). Four years of DNA barcoding: current advances and prospects. *Infection, Genetics and Evolution*, 8:727–736.

Funk, D. J. & Omland, K. E. (2003). Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, 34:397–423.

Galtier, N. & Gouy, M. (1995). Inferring phylogenies from DNA sequences of unequal base compositions. *Proceedings of the National Academy of Sciences*, 92:11317–11321.

Galtier, N., Nabholz, B., Glémin, S., & Hurst, G. D. D. (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology*, 18:4541–4550.

Gante, H. F., Moreira Da Costa, L., Micael, J., & Alves, M. J. (2008). First record of *Barbonymus schwanenfeldii* (Bleeker) in the Iberian Peninsula. *Journal of Fish Biology*, 72:1089–1094.

Gao, Z., Li, Y., & Wang, W. (2008). Threatened fishes of the world: *Myxocyprinus asiaticus* Bleeker 1864 (Catostomidae). *Environmental Biology of Fishes*, 83:345–346.

Gerson, H., Cudmore, B., Mandrak, N. E., Coote, L. D., Farr, K., & Baillargeon, G. (2008). Monitoring international wildlife trade with coded species data. *Conservation Biology*, 22:4–7.

Gerstner, C. L., Ortega, H., Sanchez, H., & Graham, D. L. (2006). Effects of the freshwater aquarium trade on wild fish populations in differentially-fished areas of the Peruvian Amazon. *Journal of Fish Biology*, 68:862–875.

Glez-Peña, D., Gómez-Blanco, D., Reboiro-Jato, M., Fdez-Riverola, F., & Posada, D. (2010). ALTER: program-oriented conversion of DNA and protein alignments. *Nucleic Acids Research*, 38:W14–W18.

Go, J., Lancaster, M., Deece, K., Dhungyel, O., & Whittington, R. (2006). The molecular epidemiology of iridovirus in Murray cod (*Maccullochella peelii peelii*) and dwarf gourami (*Colisa lalia*) from distant biogeographical regions suggests a link between trade in ornamental fish and emerging iridoviral diseases. *Molecular and Cellular Probes*, 20:212–222.

Go, J. & Whittington, R. (2006). Experimental transmission and virulence of a megalocytivirus (family *Iridoviridae*) of dwarf gourami (*Colisa lalia*) from Asia in Murray cod (*Maccullochella peelii peelii*) in Australia. *Aquaculture*, 258:140–149.

Goldberg, C. S., Pilliod, D. S., Arkle, R. S., & Waits, L. P. (2011). Molecular detection of vertebrates in stream water: a demonstration using rocky mountain tailed frogs and idaho giant salamanders. *PLoS ONE*, 6:e22746.

Goldstein, P. Z. & DeSalle, R. (2011). Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. *BioEssays*, 33:135–147.

Gotelli, N. J. & Colwell, R. K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4:379–391.

Gozlan, R. E., St-Hilaire, S., Feist, S. W., Martin, P., & Kent, M. L. (2005). Disease threat to European fish. *Nature*, 435:1046.

Grant, S. (2002). Zur Identität und Gültigkeit von *Rasbora macrophthalma* Meinken, 1951 (Cyprinidae: Rasborinae). *BSSW-Report, Verband Deutscher für Aquarien- und Terrarienkunde*, 2002:13–17.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52:696–704.

Günther, A. (1868). Catalogue of the fishes in the British Museum. *Catalogue of the fishes in the British Museum*, 7:1–512.

Hajibabaei, M., Janzen, D. H., Burns, J. M., Hallwachs, W., & Hebert, P. D. N. (2006a). DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences*, 103:968–971.

Hajibabaei, M., Singer, G. A. C., Clare, E. L., & Hebert, P. D. N. (2007). Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. *BMC Biology*, 5:1–7.

Hajibabaei, M., Smith, M., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., & Hebert, P. D. N. (2006b). A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, 6:959–964.

Hamilton, F. (1822). *An account of the fishes found in the river Ganges and its branches*. George Ramsay and Co, Edinburgh.

Han, M. V. & Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, 10:356.

Hanner, R. (2009). Data Standards for BARCODE Records in INSDC (BRIs). World Wide Web electronic publication. URL: http://barcoding.si.edu/pdf/dwg_data_standards-final.pdf.

Hardman, M. (2004). The phylogenetic relationships among *Noturus* catfishes (Siluriformes: Ictaluridae) as inferred from mitochondrial gene cytochrome *b* and nuclear recombination activating gene 2. *Molecular Phylogenetics and Evolution*, 30:395–408.

Hardman, M. & Page, L. M. (2003). Phylogenetic relationships among bullhead catfishes of the genus *Ameiurus* (Siluriformes: Ictaluridae). *Copeia*, 2003:20–33.

Hare, M. P. (2001). Prospects for nuclear gene phylogeography. *Trends in Ecology and Evolution*, 16:700–706.

Harris, J. (2003). Can you bank on GenBank? *Trends in Ecology and Evolution*, 18:317–319.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003a). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270:313–321.

Hebert, P. D. N., deWaard, J. R., & Landry, J. F. (2010). DNA barcodes for 1/1000 of the animal kingdom. *Biology Letters*, 6:359–362.

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

Hebert, P. D. N., Ratnasingham, S., & deWaard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, 270:96–99.

Hendrich, L., Pons, J., Ribera, I., & Balke, M. (2010). Mitochondrial *cox*1 sequence data reliably uncover patterns of insect diversity but suffer from high lineage-idiosyncratic error rates. *PLoS ONE*, 5:e14448.

Hensen, R. R., Ploeg, A., & Fosså, S. A. (2010). *Standard names for freshwater fishes in the Ornamental Aquatic Industry*. Ornamental Fish International, Maarssen, Netherlands.

Herre, A. (1940). Additions to the fish fauna of Malaya and notes on rare or little known Malayan and Bornean fishes. *Bulletin of the Raffles Museum*, 16:27–61.

Hickerson, M. J., Meyer, C. P., & Moritz, C. (2006). DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology*, 55:729–739.

Hillis, D. M. & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42:182–192.

Hine, P. M. & Diggles, B. K. (2005). *Import Risk Analysis: Ornamental Fish*. MAF Biosecurity New Zealand, Wellington.

Hoarau, G., Holla, S., Lescasse, R., Stam, W. T., & Olsen, J. L. (2002). Heteroplasmy and evidence for recombination in the mitochondrial control region of the flatfish *Platichthys flesus*. *Molecular Biology and Evolution*, 19:2261–2264.

Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M., & Pääbo, S. (2001). Ancient DNA. *Nature Reviews Genetics*, 2:353–359.

Holmes, B. H., Steinke, D., & Ward, R. D. (2009). Identification of shark and ray fins using DNA barcoding. *Fisheries Research*, 95:280–288.

Hopson, A. J. (1965). *Barbus* (Pisces, Cyprinidae) of the Volta region. *Bulletin of the British Museum (Natural History)*, 13:126–128.

Hora, S. L. (1921). Fish and fisheries of Manipur with some observations on those of the Naga Hills. *Records of the Indian Museum*, 22:165–214.

Hora, S. L. (1928). Notes on fishes in the Indian Museum. XV.– Notes on Burmese fishes. *Records of the Indian Museum*, 30:37–40.

Hora, S. L. (1937). On a small collection of fish from Sandoway, Lower Burma. *Records of the Indian Museum*, 39:323–331.

Hora, S. L. & Mukerji, D. D. (1928). Notes on fishes in the Indian Museum. XVI.–On fishes of the genus *Esomus* Swainson. *Records of the Indian Museum*, 30:41–56.

Hora, S. L. & Mukerji, D. D. (1934). On the collection of fish from the S. Shan states and the Pegu Yomas, Burma. *Records of the Indian Museum*, 36:123–138.

Hubert, N., Hanner, R., Holm, E., Mandrak, N. E., Taylor, E., Burridge, M., Watkinson, D., Dumont, P., Curry, A., Bentzen, P., Zhang, J., April, J., & Bernatchez, L. (2008). Identifying Canadian freshwater fishes through DNA barcodes. *PLoS ONE*, 3:e2490.

Hulme, P. E. (2009). Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, 46:10–18.

Hulme, P. E. (2012). Weed risk assessment: a way forward or a waste of time? *Journal of Applied Ecology*, 49:10–19.

Hurst, G. D. D. & Jiggins, F. M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings of the Royal Society B: Biological Sciences*, 272:1525–1534.

Huxley-Jones, E., Shaw, J. L. A., Fletcher, C., Parnell, J., & Watts, P. C. (2012). Use of DNA barcoding to reveal species composition of convenience seafood. *Conservation Biology*, 26:367–371.

Inger, R. F. & Chin, P. K. (1962). The fresh-water fishes of North Borneo. *Fieldiana Zoology*, 45:1–268.

Ivanova, N. V., Zemlak, T. S., Hanner, R. H., & Hebert, P. D. N. (2007). Universal primer cocktails for fish DNA barcoding. *Molecular Ecology Notes*, 7:544–548.

Jayaram, K. C. (1990). Two new species of the genus *Puntius* Hamilton (Pisces: Cyprinidae) from India. *Journal of the Bombay Natural History Society*, 87:106–109.

Jayaram, K. C. (1991). Systematic status of *Danio malabaricus* (Pisces: Cyprinidae). *Ichthyological Explorations of Freshwaters*, 2:109–112.

Jerde, C. L., Mahon, A. R., Chadderton, W. L., & Lodge, D. M. (2011). "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conservation Letters*, 4:150–157.

Jerdon, T. C. (1849). On the fresh-water fishes of southern India. *Madras Journal of Literature and Science*, 15:302–346.

Jiang, Y. E., Chen, X. Y., & Yang, J. X. (2008). *Microrasbora* Annandale, a new genus record in China, with description of a new species (Teleostei: Cyprinidae). *Environmental Biology of Fishes*, 83:299–304.

Johns, G. C. & Avise, J. C. (1998). A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome *b* gene. *Molecular Biology and Evolution*, 15:1481–1490.

Johnsen, A., Rindal, E., Ericson, P. G. P., Zuccon, D., Kerr, K. C. R., Stoeckle, M. Y., & Lifjeld, J. T. (2010). DNA barcoding of Scandinavian birds reveals divergent lineages in trans-Atlantic species. *Journal of Ornithology*, 151:565–578.

Joly, S., McLenachan, P. A., & Lockhart, P. J. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*, 174:E54–70.

Joseph, L. & Omland, K. E. (2009). Phylogeography: its development and impact in Australo-Papuan ornithology with special reference to paraphyly in Australian birds. *Emu*, 109:1–23.

Kelchner, S. A. & Thomas, M. A. (2007). Model use in phylogenetics: nine key questions. *Trends in Ecology and Evolution*, 22:87–94.

Kemp, B. M. & Smith, D. G. (2005). Use of bleach to eliminate contaminating DNA from the surface of bones and teeth. *Forensic Science International*, 154:53–61.

Kerr, K. C. R., Birks, S. M., Kalyakin, M. V., Red'kin, Y. A., Koblik, E. A., & Hebert, P. D. N. (2009a). Filling the gap - COI barcode resolution in eastern Palearctic birds. *Frontiers in Zoology*, 6:1–13.

Kerr, K. C. R., Lijtmaer, D. A., Barreira, A. S., Hebert, P. D. N., & Tubaro, P. L. (2009b). Probing evolutionary patterns in neotropical birds through DNA barcodes. *PLoS ONE*, 4:e4379.

Kerr, K. C. R., Stoeckle, M. Y., Dove, C. J., Weigt, L. A., Francis, C. M., & Hebert, P. D. N. (2007). Comprehensive DNA barcode coverage of North American birds. *Molecular Ecology Notes*, 7:535–543.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120.

Knight, J. D. M., Devi, K. R., & Atkore, V. (2011). Systematic status of *Systomus rubrotinctus* Jerdon (Teleostei: Cyprinidae) with notes on the *Puntius arulius* group of fishes. *Journal of Threatened Taxa*, 3:1686–1693.

Kochzius, M., Seidel, C., Antoniou, A., Botla, S. K., Campo, D., Cariani, A., Vazquez, E. G., Hauschild, J., Hervet, C., Hjörleifsdottir, S., Hreggvidsson, G., Kappel, K., Landi, M., Magoulas, A., Marteinsson, V., Nölte, M., Planes, S., Tinti, F., Turan, C., Venugopal, M. N., Weber, H., & Blohm, D. (2010). Identifying fishes through DNA barcodes and microarrays. *PLoS ONE*, 5:e12620.

Koski, L. B. & Golding, G. B. (2001). The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution*, 52:540–542.

Kottelat, M. (1982). A small collection of fresh-water fishes from Kalimantan, Borneo, with descriptions of one new genus and three new species of Cyprinidae. *Revue Suisse de Zoologie*, 89:419–437.

Kottelat, M. (1991). Notes on the taxonomy of some Sundaic and Indochinese species of *Rasbora*, with description of four new species (Pisces: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 2:177–191.

Kottelat, M. (1996). The identity of *Puntius eugrammus* and diagnoses of two new species of striped barbs (Teleostei: Cyprinidae) from Southeast Asia. *The Raffles Bulletin of Zoology*, 44:301–316.

Kottelat, M. (1998). Fishes of the Nam Theun and Xe Bangfai basins, Laos, with diagnoses of twenty-two new species (Teleostei: Cyprinidae, Balitoridae, Cobitidae, Coiidae and Odontobutidae). *Ichthyological Exploration of Freshwaters*, 9:1–128.

Kottelat, M. (2000). Diagnoses of a new genus and 64 new species of fishes from Laos (Teleostei: Cyprinidae, Balitoridae, Bagridae, Syngnathidae, Chaudhuriidae and Tetraodontidae). *Journal of South Asian Natural History*, 5:37–82.

Kottelat, M. (2001). *Fishes of Laos*. WHT Publications (Pte) Ltd, Colombo.

Kottelat, M. (2005). *Rasbora notura*, a new species of cyprinid fish from the Malay Peninsula (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 16:265–270.

Kottelat, M. (2008a). *Osteochilus bleekeri*, a new species of fish from Borneo and Sumatra (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 19:249–253.

Kottelat, M. (2008b). *Rasbora dies*, a new species of cyprinid fish from eastern Borneo (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 18:301–305.

Kottelat, M., Britz, R., Tan, H. H., & Witte, K. E. (2006). *Paedocypris*, a new genus of Southeast Asian cyprinid fish with a remarkable sexual dimorphism, comprises the world's smallest vertebrate. *Proceedings of the Royal Society B: Biological Science*, 273:895–899.

Kottelat, M. & Freyhof, J. (2007). *Handbook of European freshwater fishes*. Publications Kottelat, Cornol, Switzerland.

Kottelat, M. & Lim, K. K. P. (1995). Freshwater fishes of Sarawak and Brunei Darussalam: a preliminary annotated check-list. *The Sarawak Museum Journal*, 48:227–256.

Kottelat, M. & Pethiyagoda, R. (1990). *Danio pathirana*, a new species of cyprinid fish endemic to southern Sri Lanka. *Ichthyological Exploration of Freshwaters*, 1:247–252.

Kottelat, M. & Pethiyagoda, R. (1991). Descriptions of three new species of cyprinid fishes from Sri Lanka. In Pethiyagoda, R., editor, *Freshwater fishes of Sri Lanka*, pages 298–313. Wildlife Heritage Trust of Sri Lanka, Colombo.

Kottelat, M. & Tan, H. H. (2009). *Osteochilus flavicauda*, a new species of fish from the Malay Peninsula (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 20:1–5.

Kottelat, M. & Vidthayanon, C. (1993). *Boraras micros*, a new genus and species of minute freshwater fish from Thailand (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 4:161–176.

Kottelat, M., Whitten, A. J., Kartikasari, S. N., & Wirjoatmodjo, S. (1993). *Freshwater fishes of Western Indonesia and Sulawesi*. Periplus Editions, Hong Kong.

Kottelat, M. & Widjanarti, E. (2005). The fishes of Danau Sentarum National Park and the Kapuas Lakes Area, Kalimantan Barat, Indonesia. *The Raffles Bulletin of Zoology*, Suppl 13:139–173.

Kottelat, M. & Witte, K. E. (1999). Two new species of *Microrasbora* from Thailand and Myanmar, with two new generic names for small Southeast Asian cyprinid fishes (Teleostei: Cyprinidae). *Journal of South Asian Natural History*, 4:49–56.

Kubatko, L. S. (2009). Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology*, 58:478–488.

Kullander, S. O. (2008). Five new species of *Puntius* from Myanmar (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 19:59–84.

Kullander, S. O. & Britz, R. (2008). *Puntius padamya*, a new species of cyprinid fish from Myanmar (Teleostei: Cyprinidae). *Electronic Journal of Ichthyology*, 4:56–66.

Kullander, S. O. & Fang, F. (2004). Seven new species of *Garra* (Cyprinidae: Cyprininae) from the Rakhine Yoma, southern Myanmar. *Ichthyological Exploration of Freshwaters*, 15:257–278.

Kullander, S. O. & Fang, F. (2005). Two new species of *Puntius* from northern Myanmar (Teleostei: Cyprinidae). *Copeia*, 2005:290–302.

Kullander, S. O. & Fang, F. (2009a). *Danio aesculapii*, a new species of danio from south-western Myanmar (Teleostei: Cyprinidae). *Zootaxa*, 2164:41–48.

Kullander, S. O. & Fang, F. (2009b). *Danio tinwini*, a new species of spotted danio from northern Myanmar (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 20:223–228.

Kullander, S. O., Liao, T. Y., & Fang, F. (2009). *Danio quagga*, a new species of striped danio from western Myanmar (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 20:193–199.

Lakra, W. S., Verma, M. S., Goswami, M., Lal, K. K., Mohindra, V., Punia, P., Gopalakrishnan, A., Singh, K. V., Ward, R. D., & Hebert, P. D. N. (2011). DNA barcoding Indian marine fishes. *Molecular Ecology Resources*, 11:60–71.

Lane, N. (2009). On the origin of bar codes. *Nature*, 462:272–274.

Larmuseau, M. H. D., Raeymaekers, J. A. M., Ruddick, K. G., van Houdt, J. K. J., & Volckaert, F. A. M. (2009). To see in different seas: spatial variation in the rhodopsin gene of the sand goby (*Pomatoschistus minutus*). *Molecular Ecology*, 18:4227–4239.

Lavoué, S., Miya, M., Arnegard, M. E., McIntyre, P. B., Mamonekene, V., & Nishida, M. (2010). Remarkable morphological stasis in an extant vertebrate despite tens of millions of years of divergence. *Proceedings of the Royal Society B: Biological Sciences*, 287:1003–1008.

Le Roux, J. & Wieczorek, A. M. (2009). Molecular systematics and population genetics of biological invasions: towards a better understanding of invasive species management. *Annals of Applied Biology*, 154:1–17.

Le Vin, A. L., Adam, A., Tedder, A., Arnold, K. E., & Mable, B. K. (2011). Validation of swabs as a non-destructive and relatively non-invasive DNA sampling method in fish. *Molecular Ecology Resources*, 11:107–109.

Lee, J. Y. & Edwards, S. V. (2008). Divergence across Australia's Carpentarian barrier: statistical phylogeography of the red-backed fairy wren (*Malurus melanocephalus*). *Evolution*, 62:3117–3134.

Lefort, M. C., Boyer, S., Worner, S. P., & Armstrong, K. F. (2011). Noninvasive molecular methods to identify live scarab larvae: an example of sympatric pest and nonpest species in New Zealand. *Molecular Ecology Resources*, 12:389–395.

Lemmon, A. R. & Moriarty, E. C. (2004). The importance of proper model assumption in Bayesian phylogenetics. *Systematic Biology*, 53:265–277.

Leschen, R. A. B., Buckley, T. R., & Hoare, R. (2009). The use of tag-names and New Zealand Taxonomy. *New Zealand Entomologist*, 32:85–87.

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

Li, C., Ortí, G., Zhang, G., & Lu, G. (2007). A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evolutionary Biology*, 7:1–11.

Liang, X. F., Chen, G. Z., Chen, X. L., & Yue, P. Q. (2008). Threatened fishes of the world: *Tanichthys albonubes* Lin 1932 (Cyprinidae). *Environmental Biology of Fishes*, 82:177–178.

Liao, T. Y., Kullander, S. O., & Fang, F. (2010). Phylogenetic analysis of the genus *Rasbora* (Teleostei: Cyprinidae). *Zoologica Scripta*, 39:155–176.

Liao, T. Y. & Tan, H. H. (2011). *Brevibora cheeya*, a new species of cyprinid fish from Malay Peninsula and Sumatra. *The Raffles Bulletin of Zoology*, 59:77–82.

Lim, G. S., Balke, M., & Meier, R. (2012). Determining species boundaries in a world full of rarity: singletons, species delimitation methods. *Systematic Biology*, 61:165–169.

Lim, K. K. P. (1995). *Rasbora kottelati*, a new species of cyprinid fish from northwestern Borneo. *The Raffles Bulletin of Zoology*, 43:65–74.

Linacre, A. & Tobe, S. S. (2011). An overview to the investigative approach to species testing in wildlife forensic science. *Investigative Genetics*, 2:1–9.

Lintermans, M. (2004). Human-assisted dispersal of alien freshwater fish in Australia. *New Zealand Journal of Marine and Freshwater Research*, 38:481–501.

Linthoingambi, I. & Vishwanath, W. (2007). Two new fish species of the genus *Puntius* Hamilton (Cyprinidae) from Manipur, India, with notes on *P. ticto* (Hamilton) and *P. stoliczkanus* (Day). *Zootaxa*, 1450:45–56.

Little, D. P. (2011). DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PLoS ONE*, 6:e20552.

Little, D. P. & Stevenson, D. W. (2007). A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. *Cladistics*, 23:1–21.

Lohman, D. J., Ingram, K. K., Prawiradilaga, D. M., Winker, K., Sheldon, F. H., Moyle, R. G., Ng, P. K. L., Ong, P. S., Keng, L., Braile, T. M., Astuti, D., & Meier, R. (2010). Cryptic genetic diversity in "widespread" Southeast Asian bird species suggests that Philippine avian endemism is gravely underestimated. *Biological Conservation*, 143:1885–1890.

López, J. A., Chen, W. J., & Ortí, G. (2004). Esociform phylogeny. *Copeia*, 2004:449–464.

Lowenstein, J. H., Amato, G., & Kolokotronis, S. O. (2009). The real *maccoyii*: identifying tuna sushi with DNA barcodes - contrasting characteristic attributes and genetic distances. *PLoS ONE*, 4:e7866.

Lowenstein, J. H., Burger, J., Jeitner, C. W., Amato, G., Kolokotronis, S. O., & Gochfeld, M. (2010). DNA barcodes reveal species-specific mercury levels in tuna sushi that pose a health risk to consumers. *Biology Letters*, 6:692–695.

Lukhtanov, V. A., Sourakov, A., Zakharov, E. V., & Hebert, P. D. N. (2009). DNA barcoding Central Asian butterflies: increasing geographical dimension does not significantly reduce the success of species identification. *Molecular Ecology Resources*, 9:1302–1310.

MAF Biosecurity New Zealand (2011). Import health standard for ornamental fish and marine invertebrates from all countries. World Wide Web electronic publication accessed 24 August 2011. URL: http://www.biosecurity.govt.nz/files/ihs/fisornic.all.pdf.

Magnacca, K. N. & Brown, M. J. F. (2009). Tissue segregation of mitochondrial haplotypes in heteroplasmic Hawaiian bees: implications for DNA barcoding. *Molecular Ecology Resources*, 10:60–68.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology and Evolution*, 20:229–127.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24:133–141.

Matthews, W. J. (1987). Geographic variation in *Cyprinella lutrensis* (Pisces: Cyprinidae) in the United States, with notes on *Cyprinella lepida*. *Copeia*, 1987:616–637.

Mayden, R. L., Tang, K. L., Conway, K. W., Freyhof, J., Chamberlain, S., Haskins, M., Schneider, L., Sudkamp, M., Wood, R. M., & Agnew, M. (2007). Phylogenetic relationships of *Danio* within the order Cypriniformes: a framework for comparative and evolutionary studies of a model species. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 308:642–654.

McClelland, J. (1839). Indian Cyprinidae. *Asiatic Researches*, 19:217–471.

McDowall, R. M. (2004). Shoot first, and then ask questions: a look at aquarium fish imports and invasiveness in New Zealand. *New Zealand Journal of Marine and Freshwater Research*, 38:503–510.

McDowall, R. M. & James, G. D. (2005). *Freshwater Aquarium Fish Imports and Invasiveness: a New Zealand Evaluation*. National Institute of Water and Atmospheric Research Ltd, Christchurch, New Zealand.

McGregor, K. F., Watt, M. S., Hulme, P. E., & Duncan, R. P. (2012). How robust is the Australian Weed Risk Assessment protocol? A test using pine invasions in the Northern and Southern hemispheres. *Biological Invasions*, 14:987–998.

McKay, B. D. & Zink, R. M. (2010). The causes of mitochondrial DNA gene tree paraphyly in birds. *Molecular Phylogenetics and Evolution*, 54:647–650.

Meier, R. (2008). DNA Sequences in Taxonomy: Opportunities and Challenges. In Wheeler, Q. D., editor, *The New Taxonomy*, chapter 7, pages 95–127. CRC Press, New York.

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

Meier, R., Zhang, G., & Ali, F. (2008). The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. *Systematic Biology*, 57:809–813.

Meinken, H. (1956). Mitteilungen der fischbestimmungsstelle des VDA. XXIII. *Rasbora hengeli* spec. nov., eine sehr hübsche neuheit für das liebhaberbecken. *Aquarien und Terrarien-Zeitschrift*, 9:281–283.

Menon, A. G. K. (1952). Notes on fishes in the Indian Museum. XLVI. –On a new fish of the genus *Laubuca* from Cochin. *Records of the Indian Museum*, 49:1–4.

Menon, A. G. K. (1964). Monograph of the cyprinid fishes of the genus *Garra* Hamilton. *Memoirs of the Indian Museum*, 14:173–260.

Menon, A. G. K., Rema Devi, K., & Thobias, M. P. (1999). *Puntius chalakkudiensis*, a new colourful species of *Puntius* (family: Cyprinidae) fish from Kerala, South India. *Records of the Zoological Survey of India*, 97:61–63.

Menon, A. G. K., Rema Devi, K., & Vishwanath, W. (2000). A new species of *Puntius* (Cyprinidae: Cyprininae) from Manipur, India. *Journal of the Bombay Natural History Society*, 97:263–268.

Meyer, C. P. & Paulay, G. (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, 3:2229–2238.

Meyerson, L. A. & Reaser, J. K. (2002). Biosecurity: moving toward a comprehensive approach. *BioScience*, 52:593–600.

Millennium Ecosystem Assessment (2005). *Ecosystems and Human Well-Being: Biodiversity Synthesis*. World Resources Institute, Washington, DC.

Minamoto, T., Yamanaka, H., Takahara, T., Honjo, M. N., & Kawabata, Z. (2012). Surveillance of fish species composition using environmental DNA. *Limnology*, 13:193–197.

Ministry of Agriculture and Forestry (2011). Biosecurity Act 1993. World Wide Web electronic publication. URL: http://www.legislation.govt.nz/act/public/1993/0095/latest/096be8ed80746fc6.pdf.

Monaghan, M. T., Wild, R., Elliot, M., Fujisawa, T., Balke, M., Inward, D. J. G., Lees, D. C., Ranaivosolo, R., Eggleton, P., Barraclough, T. G., & Vogler, A. P. (2009). Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Systematic Biology*, 58:298–311.

Monbiot, G. (2011). Academic publishers make Murdoch look like a socialist. World Wide Web electronic publication. URL: http://www.guardian.co.uk/commentisfree/2011/aug/29/academic-publishers-murdoch-socialist.

Moritz, C. & Cicero, C. (2004). DNA barcoding: promise and pitfalls. *PLoS Biology*, 2:e354.

Mueller, R. L. (2006). Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. *Systematic Biology*, 55:289–300.

Munch, K., Boomsma, W., Huelsenbeck, J. P., Willerslev, E., & Nielsen, R. (2008). Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, 57:750–757.

Murray, A. G. & Peeler, E. J. (2005). A framework for understanding the potential for emerging diseases in aquaculture. *Preventive Veterinary Medicine*, 67:223–235.

Myers, G. S. (1924). On a small collection of fishes from Upper Burma. *American Museum Novitates*, 150:1–7.

Nakabo, T. (2002). *Fishes of Japan with pictorial keys to the species, English edition*. Tokai University Press, Tokyo.

Naylor, R. L., Williams, S. L., & Strong, D. R. (2001). Aquaculture—a gateway for exotic species. *Science*, 294:1655–1656.

Nei, M. (1996). Phylogenetic analysis in molecular evolutionary genetics. *Annual Review of Genetics*, 30:371–403.

Nei, M. & Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Nelson, J. S. (2006). *Fishes of the World*. John Wiley & Sons, Inc, Hoboken, New Jersey.

Ng, H. H. (2010). Hybrid synos and how to avoid them. World Wide Web electronic publication. URL: http://www.practicalfishkeeping.co.uk/content.php?sid=2928.

Ng, H. H. & Kottelat, M. (2007). *Balantiocheilos ambusticauda*, a new and possibly extinct species of cyprinid fish from Indochina (Cypriniformes: Cyprinidae). *Zootaxa*, 1463:13–20.

Ng, H. H. & Tan, H. H. (1999). The fishes of the Endau drainage, Peninsular Malaysia with descriptions of two new species of catfishes (Teleostei: Akysidae, Bagridae). *Zoological Studies*, 38:350–366.

Ng, P. K. L., Chou, L. M., & Lam, T. J. (1993). The status and impact of introduced freshwater animals in Singapore. *Biological Conservation*, 64:19–24.

Nielsen, R. & Matz, M. (2006). Statistical approaches for DNA barcoding. *Systematic Biology*, 55:162–169.

Nilsson, R. H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K. H., & Kõljalg, U. (2006). Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS ONE*, 1:e59.

Ogden, R. (2008). Fisheries forensics: the use of DNA tools for improving compliance, traceability and enforcement in the fishing industry. *Fish and Fisheries*, 9:462–472.

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., & Hofreiter, M. (2004). Genetic analyses from ancient DNA. *Annual Review of Genetics*, 38:645–679.

Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, 7:1–14.

Padilla, D. K. & Williams, S. L. (2004). Beyond ballast water: aquarium and ornamental trades as sources of invasive species in aquatic ecosystems. *Frontiers in Ecology and the Environment*, 2:131–138.

Page, R. D. M. (2012). Space, time, form: viewing the tree of life. *Trends in Ecology and Evolution*, 27:113–120.

Page, T. J. & Hughes, J. M. (2010). Comparing the performance of multiple mitochondrial genes in the analysis of Australian freshwater fishes. *Journal of Fish Biology*, 77:2093–2122.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290.

Patterson, D. J., Cooper, J., Kirk, P. M., Pyle, R. L., & Remsen, D. P. (2010). Names are key to the big new biology. *Trends in Ecology and Evolution*, 25:686–691.

Pethiyagoda, R. (1991). *Freshwater fishes of Sri Lanka*. The Wildlife Heritage Trust of Sri Lanka, Colombo.

Pethiyagoda, R. & Kottelat, M. (2005). A review of the barbs of the *Puntius filamentosus* group (Teleostei: Cyprinidae) of Southern India and Sri Lanka. *The Raffles Bulletin of Zoology*, Suppl 12:127–144.

Pethiyagoda, R., Kottelat, M., Silva, A., Maduwage, K., & Meegaskumbura, M. (2008). A review of the genus *Laubuca* in Sri Lanka, with description of three new species (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 19:7–26.

Pfenninger, M. & Schwenk, K. (2007). Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evolutionary Biology*, 6:121.

Pimentel, D., Lach, L., Zuniga, R., & Morrison, D. (2000). Environmental and economic costs of nonindigenous species in the United States. *BioScience*, 50:53–65.

Pimentel, D., Zuniga, R., & Morrison, D. (2005). Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics*, 52:273–288.

Ploeg, A. (2008). Invasive species in our industry? *OFI Journal*, 58:21–25.

Ploeg, A., Bassleer, G., & Hensen, R. (2009). *Biosecurity in the Ornamental Aquatic Industry*. Ornamental Fish International, Maarssen, Netherlands.

Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin, W. D., & Vogler, A. P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, 55:595–609.

Pons, J., Ribera, I., Bertranpetit, J., & Balke, M. (2010). Nucleotide substitution rates for the full set of mitochondrial protein-coding genes in Coleoptera. *Molecular Phylogenetics and Evolution*, 56:796–807.

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, 25:1253–1256.

Posada, D. & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53:793–808.

Prasad, G., Ali, A., & Raghavan, R. (2008). Threatened fishes of the world: *Puntius denisonii* (Day 1865)(Cyprinidae). *Environmental Biology of Fishes*, 83:189–190.

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, 21:1864–1877.

Qu, W., Shen, Z., Zhao, D., Yang, Y., & Zhang, C. (2009). MFEprimer: multiple factor evaluation of the specificity of PCR primers. *Bioinformatics*, 25:276–278.

R Development Core Team (2010). R: A language and environment for statistical computing. Vienna, Austria. URL: http://www.r-project.org/.

Rach, J., DeSalle, R., Sarkar, I. N., Schierwater, B., & Hadrys, H. (2008). Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings of the Royal Society B: Biological Sciences*, 275:237–247.

Rachmatika, I. (2004). A new species of cyprinid fish: *Puntius bunau* from the Seturan basin of Indonesian Borneo. *Treubia*, 33:181–190.

Raghavan, R., Prasad, G., Ali, P. H. A., & Sujarittanonta, L. (2007). "Boom and bust fishery" in a biodiversity hotspot - Is the Western Ghats losing its most celebrated native ornamental fish, *Puntius denisonii* Day? *Current Science*, 92:1671–1672.

Rahel, F. J. (2002). Homogenization of freshwater faunas. *Annual Review of Ecology and Systematics*, 33:291–315.

Rahel, F. J. (2007). Biogeographic barriers, connectivity and homogenization of freshwater faunas: it's a small world after all. *Freshwater Biology*, 52:696–710.

Rainboth, W. J. (1996). *Fishes of the Cambodian Mekong*. FAO, Rome.

Rainboth, W. J. & Kottelat, M. (1987). *Rasbora spilocerca*, a new cyprinid from the Mekong river. *Copeia*, 1987:417–423.

Rasmussen, R. S. & Morrissey, M. T. (2008). DNA-based methods for the identification of commercial fish and seafood species. *Comprehensive Reviews in Food Science and Food Safety*, 7:280–295.

Rasmussen, R. S., Morrissey, M. T., & Hebert, P. D. N. (2009). DNA barcoding of commercially important salmon and trout species (*Oncorhynchus* and *Salmo*) from North America. *Journal of Agricultural and Food Chemistry*, 57:8379–8385.

Ratnasingham, S. & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, 7:355–364.

Ratnasingham, S. & Hebert, P. D. N. (2011). BOLD's role in barcode data management and analysis: a response. *Molecular Ecology Resources*, 11:941–942.

Regan, C. T. (1907). Description of a new cyprinid fish of the genus *Danio* from upper Burma. *Records of the Indian Museum*, 1:395.

Reid, B. N., Le, M., McCord, W. P., Iverson, J. B., Georges, A., Bergmann, T., Amato, G., DeSalle, R., & Naro-Maciel, E. (2011). Comparing and combining distance-based and character-based approaches for barcoding turtles. *Molecular Ecology Resources*, 11:956–967.

Remi Devi, K., Indra, T. J., Raghunathan, M. B., & Raagam, P. M. (2005). A note on *Barilius bakeri* (Cyprinidae: Danioninae) from Karnataka with remarks on the status of *Opsarius malabaricus* Jerdon. *Journal of the Bombay Natural History Society*, 102:123–125.

Reyer, H. U. (2008). Mating with the wrong species can be right. *Trends in Ecology and Evolution*, 23:289–292.

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P., & Coissac, E. (2011). ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, 39:e145.

Ricciardi, A. & MacIsaac, H. J. (2000). Recent mass invasion of the North American great lakes by Ponto-Caspian species. *Trends in Ecology and Evolution*, 15:62–65.

Ripplinger, J. & Sullivan, J. (2008). Does choice in model selection affect maximum likelihood analysis? *Systematic Biology*, 57:76–85.

Rixon, C. A. M., Duggan, I. C., Bergeron, N. M. N., Ricciardi, A., & MacIsaac, H. J. (2005). Invasion risks posed by the aquarium trade and live fish markets on the Laurentian Great Lakes. *Biodiversity and Conservation*, 14:1365–1381.

Roberts, T. R. (1986). *Danionella translucida*, a new genus and species of cyprinid fish from Burma, one of the smallest living vertebrates. *Environmental Biology of Fishes*, 16:231–241.

Roberts, T. R. (1989). The freshwater fishes of Western Borneo (Kalimantan Barat, Indonesia). *Memoirs of the California Academy of Sciences*, 14:1–210.

Roberts, T. R. (1994). Systematic revision of the Southeast Asian cyprinid fish genus *Labiobarbus* (Teleostei: Cyprinidae). *The Raffles Bulletin of Zoology*, 41:315–329.

Roberts, T. R. (2007). The celestial pearl danio, a new genus and species of colourful minute cyprinid fish from Myanmar (Pisces: Cypriniformes). *The Raffles Bulletin of Zoology*, 55:131–140.

Roberts, T. R. & Kottelat, M. (1993). Revision of the southeast Asian freshwater family Gyrinocheilidae. *Ichthyological Exploration of Freshwaters*, 4:375–383.

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

Rohland, N. & Hofreiter, M. (2007). Comparison and optimization of ancient DNA extraction. *BioTechniques*, 42:343–352.

Rosenberg, N. A. (2007). Statistical tests for taxonomic distinctiveness from observations of monophyly. *Evolution*, 61:317–323.

Ross, H. A., Lento, G. M., Dalebout, M. L., Goode, M., Ewing, G., McLaren, P., Rodrigo, A. G., Lavery, S., & Baker, C. S. (2003). DNA surveillance: web-based molecular identification of whales, dolphins, and porpoises. *Journal of Heredity*, 94:111–114.

Ross, H. A., Murugan, S., & Li, W. L. S. (2008). Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology*, 57:216–230.

Rozen, S. & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, 132:365–386.

Rubinoff, D. (2006). Utility of mitochondrial DNA barcodes in species conservation. *Conservation Biology*, 20:1026–1033.

Rubinoff, D., Cameron, S., & Will, K. (2006). A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. *Journal of Heredity*, 97:581–594.

Rubinoff, D., Holland, B. S., San Jose, M., & Powell, J. A. (2011). Geographic proximity not a prerequisite for invasion: Hawaii not the source of California invasion by light brown apple moth (*Epiphyas postvittana*). *PLoS ONE*, 6:e16361.

Rutschmann, F. (2006). Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times. *Diversity and Distributions*, 12:35–48.

Ryan, J. R. J. & Esa, Y. B. (2006). Phylogenetic analysis of *Hampala* fishes (subfamily Cyprininae) in Malaysia inferred from partial mitochondrial cytochrome *b* DNA sequences. *Zoological Science*, 23:893–901.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425.

Sanz, N., Araguas, R. M., Fernández, R., Vera, M., & García-Marín, J. L. (2009). Efficiency of markers and methods for detecting hybrids and introgression in stocked populations. *Conservation Genetics*, 10:225–236.

Sarkar, I. N., Planet, P. J., & DeSalle, R. (2008). CAOS software for use in character-based DNA barcoding. *Molecular Ecology Resources*, 8:1256–1259.

Sarkar, I. N. & Trizna, M. (2011). The Barcode of Life Data Portal: bridging the biodiversity informatics divide for DNA barcoding. *PLoS ONE*, 6:e14689.

Schäfer, F. (2009). *Oreichthys crenuchoides*, a new cyprinid from West Bengal, India. *Ichthyological Exploration of Freshwaters*, 20:201–211.

Schindel, D. E. & Miller, S. E. (2005). DNA barcoding a useful tool for taxonomists. *Nature*, 435:17.

Scribner, K. T., Page, K. S., & Bartron, M. L. (2001). Hybridization in freshwater fishes: a review of case studies and cytonuclear methods of biological inference. *Reviews in Fish Biology and Fisheries*, 10:293–323.

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends in Ecology and Evolution*, 19:198–207.

Sen, N. & Dey, S. C. (1985). Two new fish species of the genus *Danio* Hamilton (Pisces: Cyprinidae) from Meghalaya, India. *Journal Assam Scientific Society*, 27:60–68.

Sevilla, R. G., Diez, A., Norén, M., Mouchel, O., Jérôme, M., Verrez-Bagnis, V., Van Pelt, H., Favre-Krey, L., Krey, G., & Bautista, J. M. (2007). Primers and

polymerase chain reaction conditions for DNA barcoding teleost fish based on the mitochondrial cytochrome *b* and nuclear rhodopsin genes. *Molecular Ecology Notes*, 7:730–734.

Shiyang, K., Srivathsan, A., Vaidya, G., & Meier, R. (2012). Is the COI barcoding gene involved in speciation through intergenomic conflict? *Molecular Phylogenetics and Evolution*, 62:1009–1012.

Shokralla, S., Singer, G. A. C., & Hajibabaei, M. (2010). Direct PCR amplification and preservative ethanol. *BioTechniques*, 48:233–234.

Shokralla, S., Zhou, X., Janzen, D. H., Hallwachs, W., Landry, J. F., Jacobus, L. M., & Hajibabaei, M. (2011). Pyrosequencing for mini-barcoding of fresh and old museum specimens. *PLoS ONE*, 6:e21252.

Siebert, D. J. (1997). The identities of *Rasbora paucisqualis* Ahl in Schreitmüller, 1935, and *Rasbora bankanensis* (Bleeker, 1853), with the designation of a lectotype for *R. paucisqualis* (Teleostei: Cyprinidae). *The Raffles Bulletin of Zoology*, 45:29–37.

Siebert, D. J. & Guiry, S. (1996). *Rasbora johannae* (Teleostei: Cyprinidae), a new species of the *R. trifasciata*-complex from Kalimantan, Indonesia. *Cybium*, 20:395–404.

Silas, E. G. (1953). Notes on fishes from Mahableshwar and Wai (Satara district, Bombay state). *Journal of the Bombay Natural History Society*, 51:579–589.

Silva, A., Maduwage, K., & Pethiyagoda, R. (2008). *Puntius kamalika*, a new species of barb from Sri Lanka (Teleostei: Cyprinidae). *Zootaxa*, 64:55–64.

Silva, A., Maduwage, K., & Pethiyagoda, R. (2010). A review of the genus *Rasbora* in Sri Lanka, with description of two new species (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, 21:27–50.

Skelton, P. H. (2001). *A Complete Guide to the Freshwater Fishes of Southern Africa*. Struik Publishers, Cape Town.

Smith, H. M. (1931). Descriptions of new genera and species of Siamese fishes. *Proceedings of the United States National Museum*, 79:1–48.

Smith, H. M. (1934). Contributions to the ichthyology of Siam. *Journal of the Siam Society, Natural History Supplement*, 9:287–325.

Smith, K. M., Anthony, S. J., Switzer, W. M., Epstein, J. H., Seimon, T., Jia, H., Sanchez, M. D., Huynh, T. T., Galland, G. G., Shapiro, S. E., Sleeman, J. M., McAloose, D., Stuchin, M., Amato, G., Kolokotronis, S. O., Lipkin, W. I., Karesh, W. B., Daszak, P., & Marano, N. (2012). Zoonotic viruses associated with illegally imported wildlife products. *PLoS ONE*, 7:e29505.

Smith, M. A., Wood, D. M., Janzen, D. H., Hallwachs, W., & Hebert, P. D. N. (2007). DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae) are not all generalists. *Proceedings of the National Academy of Sciences*, 104:4967–4972.

Smith, M. A., Woodley, N. E., Janzen, D. H., Hallwachs, W., & Hebert, P. D. N. (2006). DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). *Proceedings of the National Academy of Sciences*, 103:3657–3662.

Smits, S. A. & Ouverney, C. C. (2010). jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the Web. *PLoS ONE*, 5:e12267.

Song, H., Buhay, J. E., Whiting, M. F., & 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*, 105:13486–13491.

Sonnenberg, R., Nolte, A. W., & Tautz, D. (2007). An evaluation of LSU rDNA D1-D2 sequences for their use in species identification. *Frontiers in Zoology*, 4:6.

Sota, T. & Vogler, A. P. (2001). Incongruence of mitochondrial and nuclear gene trees in the carabid beetles *Ohomopterus*. *Systematic Biology*, 50:39–59.

Srivathsan, A. & Meier, R. (2012). On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics*, 28:190–194.

Steinke, D. & Hanner, R. (2011). The FISH-BOL collaborators' protocol. *Mitochondrial DNA*, 22 Suppl 1:10–14.

Steinke, D., Zemlak, T. S., Boutillier, J. A., & Hebert, P. D. N. (2009a). DNA barcoding of Pacific Canada's fishes. *Marine Biology*, 156:2641–2647.

Steinke, D., Zemlak, T. S., & Hebert, P. D. N. (2009b). Barcoding Nemo: DNA-based identifications for the ornamental fish trade. *PLoS ONE*, 4:e3600.

Stoeckle, M. Y. (2012). FDA certifies barcoding for seafood ID, opening commercial, educational opportunities. World Wide Web electronic publication. URL: http://phe.rockefeller.edu/barcode/blog/2012/01/05/fda-certifies-barcoding-for-seafood-id-opening-commercial-educational-opportunities/.

Sullivan, J. & Joyce, P. (2005). Model selection in phylogenetics. *Annual Review of Ecology Evolution and Systematics*, 36:445–466.

Summerbell, R. C., Lévesque, C. A., Seifert, K. A., Bovers, M., Fell, J. W., Diaz, M. R., Boekhout, T., de Hoog, G. S., Stalpers, J., & Crous, P. W. (2005). Microcoding: the second step in DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1897–1903.

Swofford, D. (2003). *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.* Sinauer Associates, Sunderland, Massachusetts.

Sykes, W. H. (1839). On the fishes of the Deccan. *Proceedings of the General Meetings for Scientific Business of the Zoological Society of London*, 1838:157–165.

Sykes, W. H. (1841). On the fishes of the Dukhun. *Transactions of the Zoological Society of London*, 2:349–378.

Taberlet, P., Griffin, S., Goossens, B., Questiau, S., Manceau, V., Escaravage, N., Waits, L., & Bouvet, J. (1996). Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, 24:3189–3194.

Taki, Y. & Katsuyama, A. (1979). Differentiation and zoogeography of two species of the cyprinid genus *Puntioplites*. *Japanese Journal of Ichthyology*, 26:253–265.

Talwar, P. K. & Jhingran, A. G. (1991). *Inland fishes of India and adjacent countries*. Oxford & IBH Publishing Co., New Delhi.

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution*, 9:678–687.

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.

Tan, H. H. (1999). *Rasbora vulcanus*, a new species of cyprinid fish from Central Sumatra. *Journal of South Asian Natural History*, 4:111–116.

Tan, H. H. (2009). *Rasbora patrickyapi*, a new species of cyprinid fish from Central Kalimantan, Borneo. *The Raffles Bulletin of Zoology*, 57:505–509.

Tan, H. H. & Kottelat, M. (2008). Revision of the cyprinid fish genus *Eirmotus*, with description of three new species from Sumatra and Borneo. *The Raffles Bulletin of Zoology*, 56:423–433.

Tan, H. H. & Kottelat, M. (2009). The fishes of the Batang Hari drainage, Sumatra, with description of six new species. *Ichthyological Exploration of Freshwaters*, 20:13–69.

Tang, K. L., Agnew, M. K., Hirt, M. V., Sado, T., Schneider, L. M., Freyhof, J., Sulaiman, Z., Swartz, E., Vidthayanon, C., Miya, M., Saitoh, K., Simons, A. M., Wood, R. M., & Mayden, R. L. (2010). Systematics of the subfamily Danioninae (Teleostei: Cypriniformes: Cyprinidae). *Molecular Phylogenetics and Evolution*, 57:198–214.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., & Vogler, A. P. (2003). A plea for DNA taxonomy. *Trends in Ecology and Evolution*, 18:70–74.

Taylor, H. R. & 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.

Taylor, M. (2012). Academic publishers have become the enemies of science. World Wide Web electronic publication. URL: http://www.guardian.co.uk/science/2012/jan/16/academic-publishers-enemies-science.

Teletchea, F. (2009). Molecular identification methods of fish species: reassessment and possible applications. *Reviews in Fish Biology and Fisheries*, 19:265–293.

Teletchea, F., Bernillon, J., Duffraisse, M., Laudet, V., & Hänni, C. (2008). Molecular identification of vertebrate species by oligonucleotide microarray in food and forensic samples. *Journal of Applied Ecology*, 45:967–975.

Thilakaratne, I. D. S. I. P., Rajapaksha, G., Hewakopara, A., Rajapakse, R. P. V. J., & Faizal, A. C. M. (2003). Parasitic infections in freshwater ornamental fish in Sri Lanka. *Diseases of Aquatic Organisms*, 54:157–162.

Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., Orlando, L., & Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, 21:2565–2573.

Tilak, R. & Jain, S. (1990). Description of a new rasborine fish, *Esomus manipurensis* from Manipur, India. *Journal of the Bombay Natural History Society*, 86:408–411.

Timmermans, M. J. T. N., Dodsworth, S., Culverwell, C. L., Bocak, L., Ahrens, D., Littlewood, D. T. J., Pons, J., & Vogler, A. P. (2010). Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research*, 38:e197.

Tobe, S. S., Kitchener, A. C., & Linacre, A. M. T. (2010). Reconstructing mammalian phylogenies: a detailed comparison of the cytochrome *b* and cytochrome *c* oxidase subunit I mitochondrial genes. *PLoS ONE*, 5:e14156.

Townsend, T. M., Alegre, R. E., Kelley, S. T., Wiens, J. J., & Reeder, T. W. (2008). Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles. *Molecular Phylogenetics and Evolution*, 47:129–142.

Tshibwabwa, S. M., Stiassny, M. L. J., & Schelly, R. C. (2006). Description of a new species of *Labeo* (Teleostei: Cyprinidae) from the lower Congo river. *Zootaxa*, 1224:33–44.

Tshibwabwa, S. M. & Teugels, G. G. (1995). Contribution to the systematic revision of the African cyprinid fish genus *Labeo*: species from the Lower Zaire river system. *Journal of Natural History*, 29:1543–1579.

Tweedie, M. W. F. (1961). Notes on Malayan fresh water fishes. *Bulletin of the Raffles Museum*, 26:178–181.

Vähä, J. P. & Primmer, C. R. (2006). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, 15:63–72.

Valdez-Moreno, M., Ivanova, N. V., Elías-Gutiérrez, M., Contreras-Balderas, S., & Hebert, P. D. N. (2009). Probing diversity in freshwater fishes from Mexico and Guatemala with DNA barcodes. *Journal of Fish Biology*, 74:377–402.

Valentini, A., Pompanon, F., & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology and Evolution*, 24:110–117.

Valiere, N. & Taberlet, P. (2000). Urine collected in the field as a source of DNA for species and individual identification. *Molecular Ecology*, 9:2150–2152.

van der Bank, H., van der Bank, M., & van Wyk, B. E. (2001). A review of the use of allozyme electrophoresis in plant systematics. *Biochemical Systematics and Ecology*, 29:469–483.

van Velzen, R., Weitschek, E., Felici, G., & Bakker, F. T. (2012). DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS ONE*, 7:e30490.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H., & Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304:66–74.

Vidthayanon, C. & Kottelat, M. (2003). Three new species of fishes from Tham Phra Wang Daeng and Tham Phra Sai Ngam caves in northern Thailand (Teleostei: Cyprinidae and Balitoridae). *Ichthyological Exploration of Freshwaters*, 14:159–174.

Virgilio, M., Backeljau, T., Nevado, B., & De Meyer, M. (2010). Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics*, 11:206.

Virgilio, M., Jordaens, K., Breman, F. C., Backeljau, T., & De Meyer, M. (2012). Identifying insects with incomplete DNA barcode libraries, African fruit flies (Diptera: Tephritidae) as a test case. *PLoS ONE*, 7:e31581.

Vishwanath, W. & Laisram, J. (2004). Two new species of *Puntius* Hamilton-Buchanan (Cypriniformes: Cyprinidae) from Manipur, India, with an account of *Puntius* species from the state. *Journal of the Bombay Natural History Society*, 101:130–137.

Vishwanath, W., Lakra, W. S., & Sarkar, U. K. (2007). *Fishes of North East India*. National Bureau of Fish Genetic Resources, Lucknow.

Vitousek, P. M., Mooney, H. A., Lubchenco, J., & Melillo, J. M. (1997). Human domination of Earth's ecosystems. *Science*, 277:494–499.

Vogler, A. P. & Monaghan, M. T. (2007). Recent advances in DNA taxonomy. *Journal of Zoological Systematics and Evolutionary Research*, 45:1–10.

Wakeley, J. (1996). The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends in Ecology and Evolution*, 11:158–162.

Ward, R. D. (2009). DNA barcode divergence among species and genera of birds and fishes. *Molecular Ecology Resources*, 9:1077–1085.

Ward, R. D., Hanner, R., & Hebert, P. D. N. (2009). The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology*, 74:329–356.

Ward, R. D. & Holmes, B. H. (2007). An analysis of nucleotide and amino acid variability in the barcode region of cytochrome *c* oxidase I (cox1) in fishes. *Molecular Ecology Notes*, 7:899–907.

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., & Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1847–1857.

Wares, J. P. (2009). Natural distributions of mitochondrial sequence diversity support new null hypotheses. *Evolution*, 64:1136–1142.

Weber, M. & de Beaufort, L. F. (1916). The fishes of the Indo-Australian Archipelago. III. Ostariophysi: II Cyprinoidea, Apodes, Synbranchi. *The Fishes of the Indo-Australian Archipelago*, 3:1–455.

Weitzman, S. H. & Chan, L. L. (1966). Identification and relationships of *Tanichthys albonubes* and *Aphyocypris pooni*, two cyprinid fishes from South China and Hong Kong. *Copeia*, 1966:285–296.

Werren, J. H. & Baldo, L. (2008). *Wolbachia*: master manipulators of invertebrate biology. *Nature Reviews Microbiology*, 6:741–751.

Whittington, R. J. & Chong, R. (2007). Global trade in ornamental fish from an Australian perspective: the case for revised import risk analysis and management strategies. *Preventive Veterinary Medicine*, 81:92–116.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.

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

Willerslev, E. & Cooper, A. (2005). Ancient DNA. *Proceedings of the Royal Society B: Biological Sciences*, 272:3–16.

Winder, L., Phillips, C., Richards, N., Ochoa-Corona, F., Hardwick, S., Vink, C. J., & Goldson, S. (2011). Evaluation of DNA melting analysis as a tool for species identification. *Methods in Ecology and Evolution*, 2:312–320.

Wong, E. H. K., Shivji, M. S., & Hanner, R. H. (2009). Identifying sharks with DNA barcodes: assessing the utility of a nucleotide diagnostic approach. *Molecular Ecology Resources*, 9:243–256.

Wong, Y. T., Meier, R., & Tan, K. S. (2010). High haplotype variability in established Asian populations of the invasive Caribbean bivalve *Mytilopsis sallei* (Dreissenidae). *Biological Invasions*, 13:341–348.

Yancy, H. F., Zemlak, T. S., Mason, J. A., Washington, J. D., Tenge, B. J., Nguyen, N.-L., Barnett, J. D., Savary, W. E., Hill, W. E., Moore, M. M., Fry, F. S., Randolph, S. C., Rogers, P. L., & Hebert, P. D. N. (2008). Potential use of DNA barcodes in regulatory science: applications of the Regulatory Fish Encyclopedia. *Journal of Food Protection*, 71:210–217.

Yassin, A., Markow, T. A., Narechania, A., O'Grady, P. M., & DeSalle, R. (2010). The genus *Drosophila* as a model for testing tree-and character-based methods of species identification using DNA barcoding. *Molecular Phylogenetics and Evolution*, 57:509–517.

Yazdani, G. M. & Talukdar, S. (1975). A new species of *Puntius* (Cypriniformes: Cyprinidae) from Khasi and Jaintia Hills (Meghalaya), India. *Journal of the Bombay Natural History Society*, 72:218–221.

Yokoyama, R., Knox, B. E., & Yokoyama, S. (1995). Rhodopsin from the fish, *Astyanax*: role of tyrosine 261 in the red shift. *Investigative Ophthalmology and Visual Science*, 36:939–945.

Zaldívar-Riverón, A., Martínez, J. J., Ceccarelli, F. S., De Jesús-Bonilla, V. S., Rodríguez-Pérez, A. C., Reséndiz-Flores, A., & Smith, M. A. (2011). DNA barcoding a highly diverse group of parasitoid wasps (Braconidae: Doryctinae) from a Mexican nature reserve. *Mitochondrial DNA*, 21 Suppl 1:18–23.

Zemlak, T. S., Ward, R. D., Connell, A. D., Holmes, B. H., & Hebert, P. D. N. (2009). DNA barcoding reveals overlooked marine fishes. *Molecular Ecology Resources*, 9:237–242.

Zhang, A. B., He, L. J., Crozier, R. H., Muster, C., & Zhu, C. D. (2010). Estimating sample sizes for DNA barcoding. *Molecular Phylogenetics and Evolution*, 54:1035–1039.

Zhang, A. B., Muster, C., Liang, H. B., Zhu, C. D., Crozier, R., Wan, P., Feng, J., & Ward, R. D. (2012). A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Molecular Ecology*, 21:1848–1863.

Zhang, A. B. & Savolainen, P. (2009). BPSI2.0: a C/C++ interface program for species identification via DNA barcoding with a BP-neural network by calling the Matlab engine. *Molecular Ecology Resources*, 6:1–3.

Zhang, A. B., Sikes, D. S., Muster, C., & Li, S. Q. (2008). Inferring species membership using DNA sequences with back-propagation neural networks. *Systematic Biology*, 57:202–215.

Zhang, E. & Kottelat, M. (2006). *Akrokolioplax*, a new genus of Southeast Asian labeonine fishes (Teleostei: Cyprinidae). *Zootaxa*, 1225:21–30.

Zhang, J. (2010). Exploiting formalin-preserved fish specimens for resources of DNA barcoding. *Molecular Ecology Resources*, 10:935–941.

Zhao, X., Li, N., Guo, W., Hu, X., Liu, Z., Gong, G., Wang, A., Feng, J., & Wu, C. (2004). Further evidence for paternal inheritance of mitochondrial DNA in the sheep (*Ovis aries*). *Heredity*, 93:399–403.

Zink, R. M. & Barrowclough, G. F. (2008). Mitochondrial DNA under siege in avian phylogeography. *Molecular Ecology*, 17:2107–2121.

Zou, S., Li, Q., Kong, L., Yu, H., & Zheng, X. (2011). Comparing the usefulness of distance, monophyly and character-based DNA barcoding methods in species identification: a case study of Neogastropoda. *PLoS ONE*, 6:e26619.

Appendix A

Photographing and preserving fishes for molecular studies: a step-by-step guide to voucher preparation

Voucher specimens are important in molecular studies, almost maybe as important as for morphological studies. A good voucher will be useful to both molecular and morphological research for many years to come. A good voucher will also allow any misidentified specimens to be easily corrected, and will permit any interesting molecular results to be effectively corroborated with morphology. But generating good vouchers in molecular studies is hard.

Formalin, the fixative chemical of choice for ichthyologists, degrades DNA and makes extraction/PCR difficult (but see Zhang, 2010). Instead, ethanol can be used as a fixative, but ethanol fixed specimens are often brittle, faded, and of poorer long-term quality. It's often best to take a tissue sample from your specimen, store this in ethanol, and formalin fix the rest of the fish as a voucher. This is fine, but you'll want to know which tissue sample comes from which specimen, and for small fishes it's not possible to permanently attach the label to the specimen without causing damage. Of course, you could put them all in individual jars, but you could soon run out of jars or space. Transporting them is a big problem too, and this is where you really need to save space.

So, after trying out some quite unsatisfactory methods, I have developed a nice method of generating quality molecular vouchers. Of course, these bags have not been tested for long-term (i.e. indefinite) storage, and are only recommended as a temporary (< 5 yr) storage or transport solution. In addition, although I haven't yet tested it, this method could hopefully be adapted for use in the field. As follows are the steps required.

Step 1. (see Figure A.1)

Fill vials for tissue samples with high-grade 100% ethanol. Label the tubes internally with pencil on archive quality “goatskin” paper, and externally with permanent marker pen. The vouchers can be kept separate using small polythene zip-seal bags. They need to be perforated first, however, with a paper hole punch (do several at a time). They should also have their bottom corners cut off to allow the bags to drain. Place another label in the bag.

Figure A.1. Prepare storage vessels.

Step 2. (see Figure A.2)

Get everything ready in advance. Here I have:

- Latex gloves
- 10% formalin (clearly labelled)
- MS-222 (fish anaesthetic)
- Spirit burner to decontaminate tools
- Variety of forceps and scalpel
- Pencil
- Squares of cardboard to use as a clean surface for tissue preparation
- Vials for tissue samples
- Bags for voucher

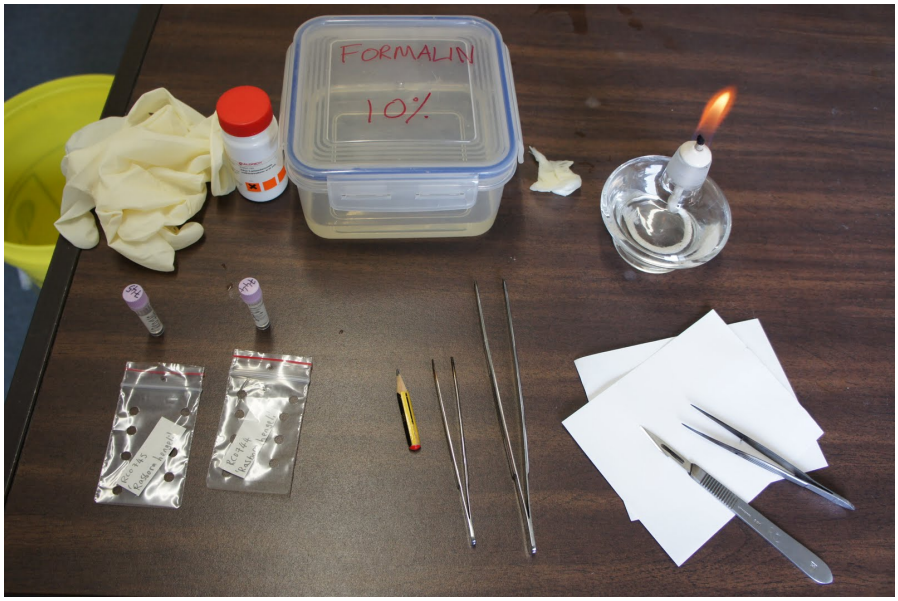

Figure A.2. General preparation.

Step 3. (see Figure A.3)

Assemble your light source and photo rig. Here I use an adjustable microscopy light (halogen desk lamps can be substituted) and a shallow white tray. I used a piece of folded graph paper as a scale for these photos. Now, mix up your MS-222 (overdosed) and water into a shallow clear tray (the lid of a tube rack), and the fish can now be added (wait for 10 mins to ensure death). Make sure the fish is only just covered.

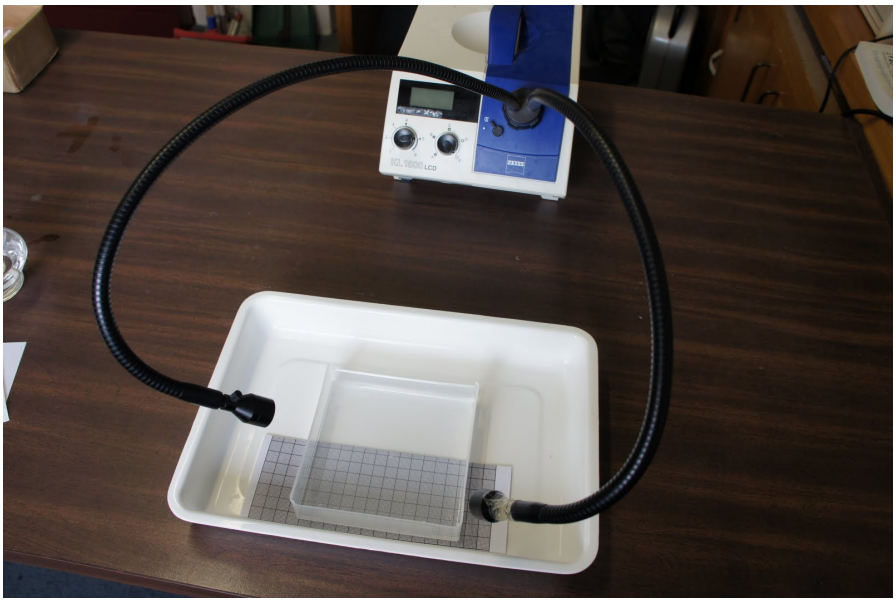

Figure A.3. Photo rig.

Step 4. (see Figure A.4 and Figure A.5)

Adjust the light angle and photograph the left-hand side of the fish, always adding the label. Remember to set your camera's white balance correctly (usually using the custom mode). The picture can then be cropped and the file name changed.

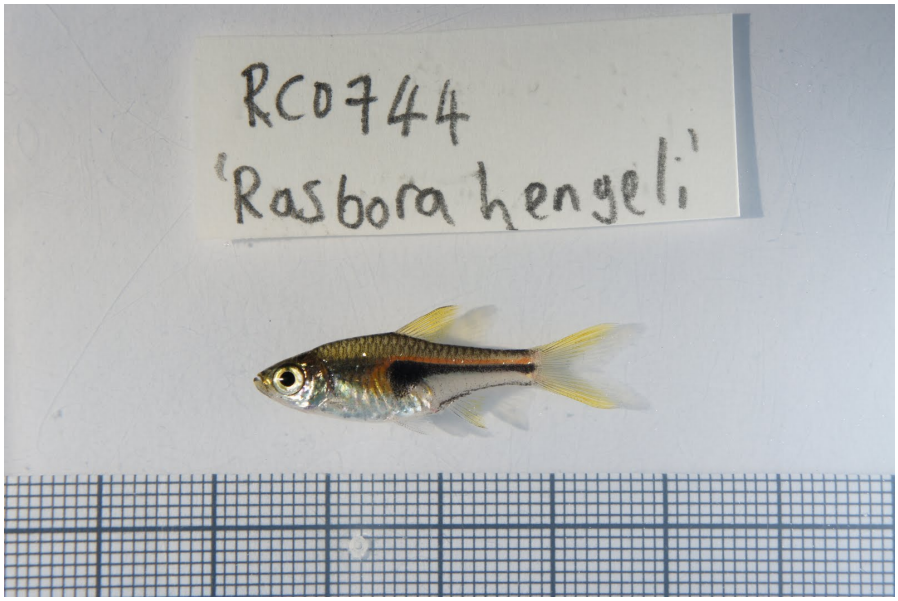

Figure A.4. Set up camera.

Figure A.5. Adjust image.

Step 5. (see Figure A.6)

Take the fish out of the solution and place on the card sheet. Use the scalpel to carefully excise a tissue sample from the right-hand side of the fish. Pectoral fin clips can also be taken to cause less damage, but on small fishes this won't yield much tissue, and using mitochondrion rich muscle may reduce the likelihood of NUMTs (see Section 1.3.1). Note: don't cut from the caudal peduncle area if characters such as caudal peduncle scale counts may be important for identifying your fish.

Figure A.6. Tissue sample.

Step 6. (see Figure A.7 and Figure A.8)

Next, place the fish into the plastic bag with the forceps, and place into the formalin. The position of the fish and fins can be manipulated through the holes in the bag with the forceps. This ensures the fish is not bent and the fins are not folded down.

Figure A.7. Bag and label specimen.

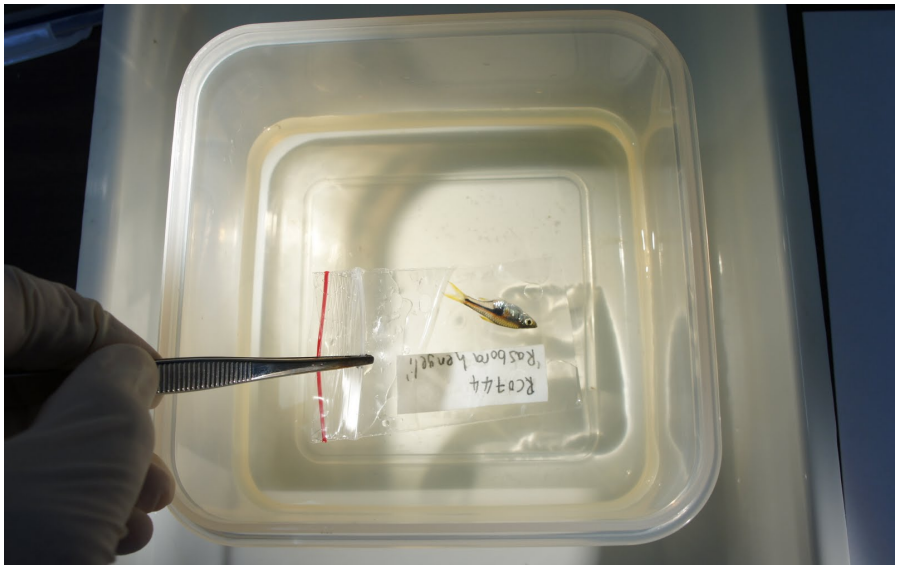

Figure A.8. Formalin fixation.

Step 7.

Throw away the card sheet and replace with new. Clean the implements with a wet tissue and then sterilise with the spirit burner. Repeat process for rest of specimens.

Step 8. (see Figure A.9)

Leave vouchers in formalin for approximately three days (longer for larger fishes). After three days, remove from formalin and wash thoroughly with water. Leave in water for 24 hours to dilute remaining formalin. Place into weak 35% alcohol (ethanol or clear methylated spirit) solution for three days before final storage in 70% alcohol. The voucher will have lost a lot of its colour by now, but can be photographed again to document the preserved colour pattern.

Figure A.9. Preserved colouration (same specimen as previously).

Appendix B

Online supplementary information

B.1 COI sequences

Text file containing all COI sequences used/generated in the study (FASTA format); available online at the following stable and permanent URL: http://goo.gl/N0h22.

B.2 RHO sequences

Text file containing all RHO sequences used/generated in the study (FASTA format); available online at the following stable and permanent URL: http://goo.gl/0GGM8.

B.3 COI NJ tree

Interactive NJ phylogram (COI data) of all specimens (this study plus GenBank data), in phyloXML SVG (scalable vector graphic) format available at the following URL: http://goo.gl/avNuz. Data including identifiers, sequences, trace files, museum voucher codes and specimen images are accessed via the BOLD and GenBank Web sites using URLs embedded in the taxon names. This figure is best viewed with Mozilla Firefox to fully enjoy the benefits of SVG and URL linking. May take up to one minute to load. A scripting "error" may appear in some browsers—this is the browser taking time to render the complex diagram. Phylogram can be saved as a pdf by printing to file using a custom paper size (approximately 3,600 mm height). Links can be opened in a new tab using Ctrl+LeftClick. Stable and permanent archived version is available at: http://goo.gl/Uvokm; may require open-source archiving software such as "7-Zip" to unpack.

B.4 RHO NJ tree

Interactive NJ phylogram (reduced RHO data), in phyloXML SVG (scalable vector graphic) format, available at: http://goo.gl/h9sY5. Data including identifiers, sequences, trace files, museum voucher codes and specimen images are accessed via the BOLD and GenBank Web sites using URLs embedded in the taxon names. This figure is best viewed with Mozilla Firefox to fully enjoy the benefits of SVG and URL linking. May take up to one minute to load. A scripting "error" may appear in some browsers—this is the browser taking time to render the complex diagram. The phylogram can be saved as a pdf by printing to file using a custom paper size (approximately 750 mm height). Links can be opened in a new tab using Ctrl+LeftClick. Stable and permanent archived version is available at: http://goo.gl/oGoyo; may require open-source archiving software such as "7-Zip" to unpack.

B.5 SPIDER tutorial

The R package SPIDER (SPecies IDentity and Evolution in R) was developed in part to address the lack of cross-platform analytical methods for DNA barcode data in this study. A tutorial on the use of this R package can be accessed at http://spider.r-forge.r-project.org/tutorial/tutorial.pdf, and was written with Samuel D. J. Brown.

B.6 Web-log

In addition to publishing work in scientific journals, additional research outputs were published on the Web, and can be found at the following blog address: http://boopsboops.blogspot.com. Appendix A comprises one of these. Examples include:

1. A method of photographing and preserving fishes for molecular studies: URL.
2. Batch extracting GenBank data from journal articles: URL.
3. Summary of the 4th International Barcode of Life Conference, Adelaide 2011: URL.

Appendix C

Table of morphological identifications

Below is presented a table of nomenclature and taxonomic authorities for each species sampled, along with project code numbers (same as BOLD specimen IDs). Nomenclature follows Eschmeyer (2010a), unless otherwise stated. Morphological characters and bibliography of references used to make each identification are included. The use of "sp.", "cf." and "aff." notation in reference specimen identification follows Kottelat & Freyhof (2007). Individuals designated "cf." are treated as conspecific with taxa of the same specific name, while those designated "aff." are treated as non-conspecific.

Taxa highlighted in red are approved to be imported into New Zealand under the current Import Health Standard (MAF Biosecurity New Zealand, 2011). Where common misidentifications occur in the trade, the scientific name of the taxon they are frequently confused with is listed; note that these are personal observations made by the author over a number of years, and do not constitute data collected during this study or any other.

Identification	Characters	Citations	Comments	Specimens
Balantiocheilos melanopterus (Bleeker)	Barbels absent; snout pointed; last unbranched dorsal ray serrated; lower lip extends posteriorly to form pocket; pelvic, anal, caudal and dorsal with wide black margins (>50% in pelvic and anal); body silver (life).	Kottelat (2001); Ng & Kottelat (2007).		RC0215 RC0216 YGN012
Barboides gracilis Brüning	Barbels absent; lateral line absent; visible humeral organ; one pair figure-8 shaped nostrils; dorsal origin anterior to pelvics; prominent axial streak; large eye (approx. 45% HL); 6½ dorsal branched rays; 5½ branched anal rays; scattered melanophores on flanks; black spot on caudal base; orange/red body colour (life).	Conway & Moritz (2006).		RC0628 RC0629
Barbonymus altus (Günther)	Two pairs barbels; short snout; last unbranched dorsal ray strongly serrated; lateral line complete (31–32 pored scales); 7½ scales between dorsal origin and lateral line; dark pigments at base of scales; caudal lobes lacking distinct black submarginal stripe; red colour to pelvics and caudal (life).	Gante *et al.* (2008); Kottelat (2001).	Frequently sold as *Barbonymus schwanenfeldii*.	RC0178 RC0179
Barbonymus schwanenfeldii (Bleeker)	As *B. altus*, but: lateral line with 33–34 pored scales; distinct black submarginal stripe to caudal lobes.	Gante *et al.* (2008); Kottelat (2001).		RC0543 RC0544
Barbus callipterus Boulenger	Two pairs barbels; mouth subterminal; last unbranched dorsal ray not serrated; lateral line complete (23+2 pored scales); dorsal concave with 8½ branched dorsal rays; 5½ branched anal rays; scales with dark bases; dorsal orange anteriorly (life) with black median spot; caudal orange at base; no markings in other fins.	Boulenger (1907).	Description brief, but best match available. Boulenger (1907) reports a terminal mouth. Rows of cephalic papillae noted.	RC0613
Barbus fasciolatus (Günther)	Two pairs barbels (maxillary length = eye diameter); body slender; lateral line complete (25–30 pored scales); 8½ branched dorsal rays; 5½ branched anal rays; approx. 10–15 black vertical bars, last forming spot on caudal peduncle; spot at anal origin.	Günther (1868); Skelton (2001).	Frequently sold as *Barbus barilioides*.	RC0035 RC0036
Barbus trispilos (Bleeker)	Two pairs barbels (rostral as long as eye diameter, maxillary approx. 1.5× eye diameter); mouth subterminal; last unbranched dorsal ray not serrated; lateral line complete, curving ventrally (24–25+2 pored scales); dorsal slightly concave with 8½ branched dorsal rays; 5½ branched anal rays; scales with dark bases; 3 distinct midlateral blotches (second and third slightly elongate).	Günther (1868); Hopson (1965).	Slightly lower lateral line scale count and shorter barbel length than reported by Hopson (1965). Rows of cephalic papillae noted.	RC0606 RC0607
Chela dadyburjori (Menon)	Barbels absent; lateral line incomplete (up to 4 pored scales); supraorbital groove present; dorsal origin posterior to that of anal; 7½ branched dorsal rays; 11½–12½ branched anal rays; elongated pectoral fins; dark midlateral stripe ending at caudal base, with 3–4 indistinct superimposed spots; no markings on fins.	Fang (2003); Menon (1952); Pethiyagoda *et al.* (2008).	Spelling of specific name follows Pethiyagoda *et al.* (2008). Generic assignment follows Tang *et al.* (2010). Frequently sold as *Chela dadiburjori*.	RC0333 RC0334 RC0335 RC0336 RC0337
Crossocheilus cf. *atrilimes* Kottelat	Two pairs barbels (maxillary rudimentary or absent in larger specimens); rostral cap fimbriated; free rostral lobe absent; lower lip papillose; 8½ branched dorsal rays; approx. 1–1½ scales between anus and anal fin; black midlateral stripe extending to end of median caudal rays; fins with no distinct markings; no distinct black marking between anus and anal fin; two rows of dark dots below midlateral stripe (absent in small specimens); proximal yellow colour to fins in large specimens.	Kottelat (2000); Kottelat & Widjanarti (2005); Tan & Kottelat (2009).	Identification tentative, as inconsistency among specimens in some characters (e.g. barbels and markings). Frequently sold as *Crossocheilus siamensis*.	RC0327 RC0521 RC0713 YGN232
Crossocheilus langei Bleeker	Two pairs barbels (maxillary rudimentary in larger specimens); rostral cap fimbriated; free rostral lobe absent; lower lip papillose; 8½ branched dorsal rays; approx. 2–2½ scales between anus and anal fin; black midlateral stripe extending to end of median caudal rays; fins with no distinct markings; distinct black marking between anus and anal fin.	Kottelat (2000); Kottelat & Widjanarti (2005); Tan & Kottelat (2009).	Maxillary barbels reduced/absent in RC0737: treated as *C.* cf. *langei*. Frequently sold as *Crossocheilus siamensis*.	RC0287 RC0288 RC0714 RC0715 RC0737 EUN115
Crossocheilus nigriloba Popta	Two pairs barbels; rostral cap fimbriated; free rostral lobe absent; lower lip papillose; 8½ branched dorsal rays; midlateral black stripe continuing onto lower caudal lobe; red marginal stripes and tips to caudal (life).	Kottelat *et al.* (1993); Rainboth (1996); Roberts (1989).		RC0735 RC0736
Crossocheilus reticulatus (Fowler)	Two pairs barbels (maxillary rudimentary or absent in larger individuals); rostral cap fimbriated; free rostral lobe absent; lower lip papillose; 8½ branched dorsal rays; large dark blotch on caudal base; dark scale margins: reticulate pattern; no distinct markings in fins.	Banarescu (1986); Fowler (1934, 1935); Kottelat (2001); Rainboth (1996); Roberts (1989).		RC0388 RC0517

Cyclocheilichthys janthochir (Bleeker)	One pair barbels (minute); lateral line complete; pores on head forming dense parallel rows; black midlateral stripe; dorsal red with black anterior margin (life); caudal red with black marginal stripe (life).	Kottelat *et al.* (1993); Roberts (1989).		RC0614 RC0615 YGN291
Cyprinella lutrensis (Baird & Girard)	Barbels absent; lateral line complete (33 pored scales); 8½ branched anal rays; well developed tubercles on head; metallic blue body (life); dark bar behind operculum; pectoral, pelvic and caudal red (life); dorsal surface of head red (life); body with reticulate scale pattern.	Boschung & Mayden (2004); Matthews (1987).	Large number of synonyms in this species.	RC0207 RC0208
Cyprinus carpio Linnaeus	Two pairs barbels; lateral line complete (31 +1 pored scales); long concave dorsal; caudal deeply emarginate; last unbranched anal ray spinous and serrated posteriorly.	Kottelat & Freyhof (2007).	The ornamental "koi" variety is hypothesised to belong to *Cyprinus rubrofuscus* Lacepède by Kottelat & Freyhof (2007). Wild *C. rubrofuscus* should have 29–33 pored lateral line scales and this specimen agrees with the diagnosis, but due to support from a single character, and the selective breeding in ornamental varieties, the "koi" is retained here for now as *C. carpio*.	EUN226
Danio aesculapii Kullander & Fang	Two pairs barbels (rostral not extending past pectoral base); 6½ branched dorsal rays; lateral line incomplete; approx. 6 short lateral bars anteriorly, continuing into parallel rows of spots/dots; distinct A-stripe.	Kullander & Fang (2009a).	Frequently sold as *Danio* sp. "pantheri", or *D.* sp. "TW03".	RC0111 RC0112 RC0706 RC0707 RC0708
Danio albolineatus (Blyth)	Two pairs long barbels (rostral extending to eye); lateral line incomplete (up to 9 pored scales); 7½ branched dorsal rays; body devoid of stripes except a dark P-stripe posterior on body, bordered above by light I-stripe, ending on caudal base; blue/pink colouration in life.	Fang & Kottelat (1999, 2000).	The *D. albolineatus* complex is poorly characterised and requires systematic attention. Numerous synonyms exist, but these specimens are regarded by the oldest available name.	RC0076 RC0077 RC0089 RC0443 RC0445
Danio choprae Hora	Two pairs barbels (rostral not extending past eye, maxillary not extending past pectoral base); 7½ branched dorsal rays; lateral line absent; 6–8 short lateral bars anteriorly, continuing into rows of spots and P-stripe on caudal peduncle; P+1 and P–1 stripes continue onto caudal; distinct A and D stripes.	Hora (1928); Kullander & Fang (2009a).	Spelling of specific name follows Kullander & Fang (2009a). Frequently sold as *Danio choprai*.	RC0059 RC0060 RC0079 RC0163 RC0164 RC0446
Danio aff. *choprae* Hora	As *D. choprae*, but barbels longer (rostral extending past eye, maxillary extending past pectoral base); lateral line incomplete (1–3 pored scales); anterior lateral bars broken up with intermediate spots; larger size; overall grey rather than orange colouration (life).	Hora (1928); Kullander & Fang (2009a).	Likely an undescribed species, differing in several characters from *D. choprae*. Spelling of specific name follows Kullander & Fang (2009a).	RC0523 RC0524 RC0525 RC0669 RC0670
Danio dangila (Hamilton)	Two pairs long barbels (maxillary reach past operculum); supraorbital groove absent; lateral line complete (32–36 pored scales); 9½–11½ branched dorsal rays; 15½ branched anal rays; well defined vertically elongated cleithral spot; network of P-stripes (blue in life) and interspaces forming spots and rings; P-stripes continue onto caudal; anal with 2–3 A-stripes.	Day (1875); Hamilton (1822); Sen & Dey (1985); Talwar & Jhingran (1991).	RC0343 appears different, with darker pattern, larger size; wider P-stripes, smaller interspace spots, a distinct axial streak, and a cleithral spot not elongated vertically. This specimen is regarded here as *Danio* cf. *dangila*.	RC0122 RC0123 RC0343 RC0344 RC0345 RC0346 RC0347 RC0348

Danio aff. *dangila* (Hamilton)	As *D. dangila*, but with stripes on dorsal and caudal forming distinct and discreet spots.	Day (1875); Hamilton (1822); Sen & Dey (1985); Talwar & Jhingran (1991).	Likely an undescribed *Danio* closely related to *D. dangila*. Purportedly sourced from Myanmar.	RC0560 RC0561 RC0562 RC0563 RC0564
Danio erythromicron (Annandale)	Barbels absent; lateral line absent; 7½ branched dorsal rays; supraorbital groove absent; snout short and blunt; well defined dark spot at caudal base; fins without stripes; up to 12 narrow lateral bars, from operculum to caudal peduncle.	Annandale (1918); Conway *et al.* (2008).		RC0552 RC0553 RC0599 RC0704 RC0705 YGN172 YGN340
Danio feegradei Hora	Two pairs long barbels (maxillary extends past operculum); lateral line complete (approx. 36 pored scales); 8½ branched dorsal rays; 12½ branched anal rays; cleithral spot present; dark P-stripe narrowing posteriorly and terminating in spot on caudal base, with light I-stripe above posteriorly (on caudal peduncle and base); light spots in two rows anteriorly.	Hora (1937).		RC0245 RC0246 RC0247 RC0248 RC0249
Danio cf. *kerri* Smith	Two pairs barbels (rostral extends past eye, maxillary beyond pectoral base); 7½ branched dorsal rays; lateral line incomplete (up to 9 pored scales); two complete lateral stripes (P and P+1) with two light interspaces, widening posteriorly and joining in a loop behind the operculum; fins dusky with weak pigmentation.	Smith (1931).	Smith (1931) reports no pored lateral line scales in *D. kerri*, so this material is regarded as *D.* cf. *kerri*.	EUN035 RC0267 RC0268 RC0269 RC0270 RC0271
Danio kyathit Fang	Two pairs long barbels (maxillary extends past operculum); supraorbital groove absent; lateral line incomplete (5–9 pored scales); 13½–14½ branched anal rays; D-stripe and 3 A-stripes present; 5–7 P-stripes broken almost entirely into spots; P, P+1 and P–1 extending onto caudal; caudal without stripes on lobes.	Fang (1998); Kullander *et al.* (2009).	Conforms to holotype of *D. kyathit* Fang (1998).	RC0064 RC0090 RC0129 RC0130 RC0131 YGN014 YGN338
Danio aff. *kyathit* Fang	As *D. kyathit*, but: P-stripes as stripes rather than spots; P–1 and P–2 stripes slightly ventrally slanting.	Fang (1998); Kullander *et al.* (2009).	A likely undescribed species with distinct colour pattern from *D. kyathit* s.s. holotype (Fang, 1998). A paratype of *D. kyathit* from Kamaing (Ayeyarwaddy drainage) shows a similar pattern. Similar also to *D. quagga* Kullander, Liao & Fang, but barbels appear longer here, and *D. quagga* is a poorly known species. Frequently sold as *D. kyathit*.	EUN041 EUN179 RC0065 RC0066 RC0120 RC0121 RC0405
Danio margaritatus (Roberts)	Barbels absent; lateral line absent; 7½ branched dorsal rays; supraorbital groove absent; snout short and blunt; D-stripe, A-stripe and A-1 stripe present; P+1 and P–1 stripes extend onto caudal; 5–6 irregular rows of spots; distinctive blue, red, gold colouration (life).	Conway *et al.* (2008); Roberts (2007).		RC0032 RC0033 RC0107 RC0138 RC0139

Danio meghalayensis Sen & Dey	Two pairs barbels (maxillary not reaching past operculum, rostral just extending past eye); supraorbital groove absent; lateral line complete (33–34 pored scales); 8½ branched dorsal rays; 10½–11½ branched anal rays; no distinct cleithral spot; 5 P-stripes, with interspaces forming broken golden (life) spots and stripes anteriorly; P-stripes continue onto caudal; anal with A-stripes.	Day (1875); Hamilton (1822); Sen & Dey (1985); Talwar & Jhingran (1991).		RC0565 RC0566 RC0567 RC0568
Danio nigrofasciatus (Day)	One pair barbels (maxillary, reaching past eye); P and P+1 stripes uniform unbroken, extending into caudal; no stripe above P+1; stripes below P broken into spots; anal and pelvics spotted; D-stripe present.	Fang (1998); Kullander & Fang (2009b).		EUN034 RC0081 RC0082 RC0242 RC0243 RC0244
Danio rerio (Hamilton)	Two pairs long barbels (maxillary extends past operculum, rostral not extending past eye); lateral line absent, except in RC0679 (4 pored scales); D-stripe and 3 A-stripes present; 5 well defined parallel P-stripes, with P, P+1 and P–1 extending onto caudal; caudal with stripes on lobes.	Fang (1998); Hamilton (1822); Kullander *et al.* (2009).	Hamilton (1822) reports lateral line "scarcely observable", so it's hard to discern if an abbreviated or absent lateral line conforms to description. Here, the Indian wild-caught specimen (RC0679) is referred to *D.* cf. *rerio*. Several specimens were the "leopard" variety *D. frankei* (Meinken), understood to be a selective breeding form and junior subjective synonym of *D. rerio* (Mayden *et al.*, 2007). Sometimes sold as *D. frankei*.	EUN228 RC0067 RC0068 RC0069 RC0070 RC0071 RC0072 RC0088 RC0105 RC0394 RC0679 YGN413
Danio roseus Fang & Kottelat	As *D. albolineatus*, but: smaller; slimmer; slightly shorter barbels; posterior light and dark P/I stripes absent or v. indistinct.	Fang & Kottelat (1999, 2000).	The *D. albolineatus* complex is poorly characterised and requires systematic attention.	RC0126 RC0127 RC0128 RC0547 RC0548
Danio sp. "hikari"	Two pairs barbels (rostral extends past operculum, maxillary beyond pectoral); 7½ branched dorsal rays; lateral line incomplete; two complete lateral stripes (P and P+1) with two light interspaces, not joining in a loop behind the operculum; distinct D-stripe, A-stripe and A-1 stripe.	Smith (1931).	Similar to *D. kerri*, but likely an undescribed species.	EUN039 RC0262 RC0263 RC0264 RC0265 RC0266
Danio tinwini Kullander & Fang	One pair barbels (maxillary); lateral line absent; 6½ branched dorsal rays; 3 P-stripes, broken into rows of discrete spots; anal, dorsal and pelvics spotted.	Fang (1998); Kullander & Fang (2009b).	Frequently sold as *Danio* sp. "Burma" or *D.* sp. "TW02".	RC0062 RC0063 RC0158 RC0159 RC0160 YGN426 YGN511
Danionella dracula Britz, Conway & Rüber	Scales absent; miniature size (up to 17 mm SL); remnant larval caudal fin-folds; 13 total anal rays; 16 principal caudal rays; genital papilla not developed as a conical projection; body transparent with yellow/green lateral stripe (life).	Britz (2009); Britz *et al.* (2009); Roberts (1986)		YGN118
Devario cf. *acuticephala* (Hora)	Barbels absent; lateral line absent; supraorbital groove present; 10½ branched anal rays; caudal not truncate; pectorals not pointed and not reaching pelvic base; broad longitudinal stripe; no markings on fins.	Barman (1991); Hora (1921); Hora & Mukerji (1934); Talwar & Jhingran (1991).	Specimen in poor condition, and identification therefore tentative. Does not disagree with *D. acuticephala*.	RC0115

Devario cf. *aequipinnatus* (McClelland)	Two pairs barbels (rostral longer than maxillary); lateral line complete (31–36 pored scales); infraorbital process IO1 present; 10½–11½ branched dorsal rays; 12½–13½ branched anal rays; cleithral spot round and well defined; P-stripes interrupted anteriorly; P-stripe extending onto median caudal rays.	Barman (1984a); Day (1875); Fang (1997b, 2000); Jayaram (1991); McClelland (1839); Talwar & Jhingran (1991).	Identification tentative, as the concept of *D. aequipinnatus* varies considerably among authors, and is poorly characterised: following Day (1875) here.	RC0349 RC0350 RC0351 RC0352 RC0464
Devario auropurpureus (Annandale)	Barbels absent; snout sharply pointed; narrow elongate body; origin of dorsal slightly anterior to anal; lateral line complete (approx. 37 pored scales); branched dorsal rays 7½; branched anal rays 14½–16½; approx. 14 bluish (life) lateral bars; fine dark granulation on fins.	Annandale (1918); Barman (1984b).		RC0610 RC0689 RC0691 YGN246 YGN398 YGN485 YGN509
Devario cf. *browni* (Regan)	Two pairs barbels (v. small); infraorbital process IO1 present; lateral line complete (approx. 32 pored scales); branched dorsal rays 9½–10½; branched anal rays 12½–13½; predorsal scales 14–15; cleithral spot present; 3 wavy P-stripes (P-stripe continues onto caudal).	Fang (2000); Fang & Kullander (2009); Regan (1907).	Tentative identification: not entirely consistent with characters of *D. browni* presented by Fang (2000). The P+1 and P–1 stripes should meet to form a loop anteriorly: this character is not present in all material here, and the loop is positioned too far anteriorly for *D. browni* (above end of pectorals). Fin ray counts are reported to be quite varied in different populations of *D. browni* (Fang, 2000).	RC0196 RC0197 RC0198 RC0199 RC0200 YGN154
Devario cf. *chrysotaeniatus* (Chu)	Two pairs barbels (rostral approx. ½ eye diameter, maxillary tiny); infraorbital process IO1 present; branched dorsal rays 7½–8½; branched anal rays 12½; cleithral spot present; dorsal and anal with faint median stripe; P-stripe strong: starting above pelvics and continuing onto caudal; weak P+1 and P+2 stripes; interspace stripes break up anteriorly into dots.	Fang (2000); Fang & Kottelat (1999); Kottelat (2001).	Tentative identification: *D. chrysotaeniatus* should not have a process on infraorbital IO1. Alternative identification could be *D. laoensis* (Pellegrin & Fang).	RC0258 RC0259 RC0261
Devario cf. *devario* (Hamilton)	One pair barbels (small); lateral line complete (44–46 pored scales); infraorbital process IO1 absent; 15½–16½ branched dorsal rays; 16½–17½ branched anal rays; deep rhomboidal body shape; cleithral spot absent; three stripes on posterior of body (blue in life); network of spots and stripes in anterior of body (blue and yellow in life).	Conway *et al.* (2009); Hamilton (1822); Talwar & Jhingran (1991).	*Devario devario* is reported as having no barbels. This material has small but obvious barbels, so may not be conspecific with *D. devario*.	RC0510 RC0585 RC0586 RC0587
Devario malabaricus (Jerdon)	Two pairs barbels; lateral line complete (36–40 pored scales); infraorbital process IO1 absent; 11½ branched dorsal rays; 14½–15½ branched anal rays; snout pointed; cleithral spot present as vertical mark; 4–5 lateral stripes breaking up into spots anteriorly (blue in life).	Jayaram (1991); Jerdon (1849); Kottelat & Pethiyagoda (1990); Talwar & Jhingran (1991).	Frequently sold as *Devario aequipinnatus*.	RC0406 RC0407 RC0408 RC0409 RC0410 RC0462 RC0733
Devario pathirana (Kottelat & Pethiyagoda)	Two pairs barbels; lateral line complete; infraorbital process IO1 present; 7–11 irregular parallel bars (dark blue in life); longitudinal stripe on caudal peduncle continuing onto median caudal rays; dark median stripe in dorsal.	Kottelat & Pethiyagoda (1990).		RC0529 RC0530 RC0692 RC0693
Devario sondhii (Hora & Mukerji)	Barbels absent; lateral line incomplete (8–10 pored scales); supraorbital groove present; dorsal 7½ branched rays; cleithral spot present; iridescent lateral stripe on posterior of body; sides covered with small pigmented dots; no markings on fins.	Hora & Mukerji (1934).		RC0113 RC0114 RC0165 RC0166 RC0167

Devario sp. "giraffe"	Two pairs barbels (v. small); infraorbital process IO1 present; deep, bulky body shape; lateral line complete (approx. 31–34 pored scales); branched dorsal rays 9½–11½; branched anal rays 12½–14½; predorsal scales 14–15; cleithral spot not distinct; P-stripes and interspaces broken up anteriorly into spots, rings and vertical bars.	Cottle (2010); Fang (2000); Fang & Kottelat (1999); Fang & Kullander (2009); Kottelat (2001); Regan (1907).	Presented here as an undescribed species: does not match literature, although many nominal *Devario* spp. are very poorly known. Appears very similar to *D.* sp. "giraffe" and *D.* cf. *malabaricus* as presented by Cottle (2010).	EUN042 RC0257 RC0260 RC0511 RC0634 RC0635 RC0687 RC0694 RC0695
Devario sp. "purple cypris"	Barbels absent; snout blunt, round; supraorbital groove present; infraorbital process IO1 absent; lateral line complete; approx. 9–10 lateral bars; fine dark granulation on fins (no stripes).	Annandale (1918); Barman (1984b); Fang (1997a); Fang & Kottelat (1999).	Presented here as an undescribed species: does not match literature, although many nominal *Devario* spp. are poorly known.	RC0250 RC0251 RC0252 RC0253
Devario sp. "TW04"	Barbels absent; infraorbital process IO1 absent; lateral line complete (approx. 33 pored scales); branched dorsal rays 9½; branched anal rays 10½; predorsal scales 14; cleithral spot absent; three P-stripes, with P+1 and P–1 stripes joining irregularly; two rows of metallic pink coloured scales along dorsal midline.	Cottle (2010); Fang (2000); Fang & Kottelat (1999).	Unable to confidently place to known species. Strong visual match to *D.* sp. "TW04" as presented in Cottle (2010).	YGN072
Devario sp. "undet. (1)"	Two pairs barbels (rostral longer than maxillary, and less than half eye width); lateral line complete (29–30 pored scales); infraorbital process IO1 present; 11½–12½ branched dorsal rays; 12½–13½ branched anal rays; cleithral spot present; 4–5 P-stripes, breaking up anteriorly; P-stripe wider, and extending onto median caudal rays; dusky median stripe in dorsal.	Fang (1997b, 2000); Fang & Kottelat (1999); Kottelat (2001); Myers (1924).	Literature unable to discriminate. *Devario acrostomus* (Fang and Kottelat) and *D. kakhienensis* (Anderson) are similar. Conservatively, it is presented as an undetermined (i.e. an unidentified or undescribed) species. Many nominal *Devario* spp. are poorly known. Sold as *D. strigillifer* (Myers).	RC0187 RC0188 RC0189 RC0190
Devario sp. "undet. (2)"	Two pairs barbels (rostral longer than maxillary); lateral line complete (30–32 pored scales); infraorbital process IO1 present; 9½–11½ branched dorsal rays; 10½–11½ branched anal rays; cleithral spot present; 3–4 P-stripes; P-stripe wider, and extends onto median caudal rays; bright green/yellow colouration (life).	Fang (1997b, 2000); Kottelat (2001); Myers (1924).	Possibly conspecific with *D. kakhienensis* (Anderson), but not positive enough to apply the name. Conservatively, it is presented as an undetermined (i.e. unidentified or undescribed) species. Many nominal *Devario* spp. are poorly known. Purportedly sourced from Myanmar, and sold as *D.* sp. "fluoro" or "Himalayan lemon".	RC0480 RC0481 RC0531 RC0532 RC0533
Eirmotus furvus Tan & Kottelat	Barbels absent; mouth terminal; cephalic papillae present on head (arranged in rows); lateral line incomplete; last unbranched dorsal ray serrated; 8 dark conspicuous bars, with width of bar 5 greater than 1½ scales; mark on posterior of dorsal adjacent to bar 6; last unbranched dorsal ray entirely pigmented; distinct black mark anterior to anus; back upper margin of pectoral; body and fins dusky with scattered chromatophores on fin rays.	Tan & Kottelat (2008).	Frequently sold as *Eirmotus octozona*.	YGN345
Eirmotus cf. *insignis* Tan & Kottelat	Barbels absent; mouth terminal; cephalic papillae present on head (arranged in rows); lateral line incomplete (2–6 pored scales); last unbranched dorsal ray serrated (approx. 21 serrae); 8 dark bars, with width of bar 5 approx. 1–1½ scales; row median dark spots on dorsal; mark on posterior of dorsal adjacent to bar 6; unbranched dorsal rays entirely pigmented; last unbranched anal ray pigmented in some specimens.	Tan & Kottelat (2008).	Identification tentative, as pigmentation on last unbranched dorsal and anal rays extending entire length of ray rather than proximal half/base. Diagnoses in Tan & Kottelat (2008) difficult to reconcile with these specimens. Frequently sold as *Eirmotus octozona*.	EUN052 RC0667 RC0668 YGN050
Eirmotus cf. *octozona* Schultz	Barbels absent; mouth terminal; cephalic papillae present on head (arranged in rows); lateral line incomplete; last unbranched dorsal ray serrated (approx. less than 20 serrae); 8 dark bars, with width of bar 5 approx. 1 scale; row median dark spots on dorsal absent; unbranched dorsal rays entirely pigmented; unbranched anal rays unpigmented.	Tan & Kottelat (2008).	Identification tentative, as count of unbranched dorsal ray serrae fall short of the 25–31 expected in *E. octozona*. Diagnoses in Tan & Kottelat (2008) difficult to reconcile with these specimens.	YGN077 YGN233

Epalzeorhynchos bicolor (Smith)	Two pairs barbels (black); fimbriate rostral cap with free lateral lobe not terminating in sharp tubercle; upper lip poorly developed; lower lip not papillose; body and fins uniform dark colour; caudal orange/red (life); dorsal with white edge; dark spots behind operculum and above pectorals.	Kottelat *et al.* (1993); Roberts (1989); Smith (1931); Zhang & Kottelat (2006).		EUN080 RC0321 RC0322 YGN019
Epalzeorhynchos frenatum (Fowler)	Two pairs barbels; fimbriate rostral cap with free lateral lobe not terminating in sharp tubercle; upper lip poorly developed; lower lip not papillose; dark blotch at caudal base; no black or white margin to dorsal, pelvic and pectoral; all fins dusky orange/red (life).	Kottelat (1998, 2001); Rainboth (1996); Roberts (1989); Zhang & Kottelat (2006).		EUN081 RC0213 RC0214 YGN032
Epalzeorhynchos kalopterus (Bleeker)	Two pairs barbels (rostral black, maxillary pale); fimbriate rostral cap with free lateral lobe terminating in sharp tubercle; upper lip poorly developed; lower lip not papillose; well defined, broad lateral stripe (snout tip to median caudal rays).	Kottelat *et al.* (1993); Roberts (1989); Zhang & Kottelat (2006).		EUN079 RC0519 RC0520 YGN061 YGN127 YGN373 YGN400 YGN489
Esomus metallicus Ahl	Two pairs barbels (rostral extending past eye, maxillary extending past pelvic base); supraorbital groove absent; lateral line single and incomplete (extends to approx. between pelvic and anal); lateral stripe and more intense posteriorly, terminating at caudal base; no markings on fins.	Fang (2003); Hora & Mukerji (1928); Kottelat (2001); Talwar & Jhingran (1991); Tilak & Jain (1990).		RC0653 RC0654 RC0655 RC0656 RC0657 YGN090
Garra cambodgiensis (Tirant)	Mouth inferior; upper and lower lips continuous, with lower lip modified into sucking disc; snout tuberculated; one pair barbels (rostral); wide midlateral stripe (approx. 2 scales width); two dark bands (proximal and distal) in dorsal; caudal plain with red margins (life).	Kottelat (2001); Rainboth (1996).	Frequently sold as *Crossocheilus siamensis*.	RC0716 RC0717
Garra cf. ceylonensis Bleeker	Mouth inferior; ventral surface of head and body flattened; upper and lower lips continuous, with lower lip modified into sucking disc; proboscis absent; two pairs barbels; lateral line complete (32 pored scales); dark spot on gill opening; distance of anus from anal fin origin less than 4× in distance between pelvic fin origin and anal fin origin; interorbital width greater than 0.5× HL; dark spots at dorsal base absent; dark midlateral stripe with several narrow light and dark longitudinal stripes posteriorly.	Menon (1964); Talwar & Jhingran (1991)	Tentative identification as many *Garra* spp. are poorly known. Keys out as *G. ceylonensis* in Talwar & Jhingran (1991), but *G. mullya* Sykes is a plausible alternative identification, a species with a wider distribution.	YGN399
Garra flavatra Kullander & Fang	Mouth inferior; ventral surface of head and body flattened; upper and lower lips continuous, with lower lip modified into sucking disc; proboscis absent; lateral line complete (28 pored scales); 7½ branched dorsal rays; shallow rostral furrow; rostral lobe present; tubercles on rostral lobes and snout; abdomen scaled; black spot at gill opening; 3 yellow contrasting bars (life); wide, dark distal band and white tip to dorsal; subdistal band to caudal; spots on caudal.	Kullander & Fang (2004).		EUN163 RC0317 RC0318 YGN016 YGN155 YGN376
Garra gotyla (Gray)	Mouth inferior; ventral surface of head and body flattened; upper and lower lips continuous, with lower lip modified into sucking disc; two pairs barbels; upper lip not tuberculate; chest and ventral surface scaled; no distinct proboscis or rostral fold; lateral line complete (31–32 pored scales); 8½ branched dorsal rays; dark blotch/bar at caudal base; longitudinal stripes on posterior of body; dark posterior margin to dorsal and caudal; red/pinkish fins (life).	Menon (1964); Talwar & Jhingran (1991); Vishwanath *et al.* (2007).	Individuals appear juvenile, and lacking proboscis.	YGN062 YGN166 YGN219 YGN478 RC0390 RC0391

Garra gravelyi (Annandale)	Mouth inferior; ventral surface of head and body flattened; upper and lower lips continuous, with lower lip modified into sucking disc; unilobed indistinct square proboscis; transverse groove across upper lip; two pairs barbels (maxillary shorter than rostral); 8½ branched dorsal rays; lateral line complete (32 pored scales); 8 predorsal scales; dark spot on gill opening; dark spots at dorsal base; dark midlateral stripe.	Kottelat (2000); Menon (1964).	Unable to count diagnostic circumpeduncular scales due to tissue excision from this area: estimated from photograph to be approx. 12.	RC0272 RC0273 YGN046
Garra rufa (Heckel)	Mouth inferior; ventral surface of head and body flattened; upper and lower lips continuous, with lower lip modified into sucking disc; lateral line complete (35 pored scales); proboscis absent; 8½ branched dorsal rays; 17 branched caudal rays; 4–5 dark spots at base of dorsal; black spot at upper opening of operculum; dark blotch at caudal base; lower lobe of caudal dark; darkly mottled flanks.	Coad (2010); Menon (1964).		RC0526 RC0527 YGN105 YGN159 YGN199
Garra sp. "undet. (1)"	Mouth inferior; ventral surface of head and body flattened; upper and lower lips continuous, with lower lip modified into sucking disc; proboscis absent; two pairs barbels; snout rounded; lateral line complete (approx. 33 pored scales; 8½ branched dorsal rays; no spots at dorsal base; dark bar at base of caudal; fins with no distinct markings; no longitudinal stripes posteriorly; no spot behind gill opening; fins with no distinct markings.	Menon (1964); Talwar & Jhingran (1991); Vishwanath *et al.* (2007).	Unable to confidently place to known species. *G. annandalei* Hora and *G. manipurensis* Vishwanath & Sarojnalini appear close.	RC0386 RC0387
Gyrinocheilus aymonieri (Tirant)	Spiracle above operculum; dorsal with 9½ branched rays; caudal spotted; dark spot posterior to spiracle.	Roberts & Kottelat (1993).	*Gyrinocheilus* is a gyrinocheilid.	EUN164 RC0395 RC0396 YGN018 YGN033 YGN230
Hampala macrolepidota Kuhl & van Hasselt	One pair barbels; mouth large, extending past anterior margin of eye; last unbranched dorsal ray finely serrated; lateral line complete (25–27 pored scales); narrow black bar between dorsal and anal origin; black bar on caudal peduncle; caudal red (life) with black submarginal stripes.	Doi & Taki (1994); Inger & Chin (1962); Kottelat (1998, 2001); Ryan & Esa (2006).	Discrepancies in lateral line scale counts and presence of black markings on posterior of body make identification as *H. macrolepidota* tentative. However, inconsistency between authors suggest the name be maintained here as most likely identification. Specimens were immature.	RC0367 RC0368
Hypsibarbus wetmorei (Smith)	Lateral line complete; 4½ scales between lateral line and dorsal origin; 2 rows of scales between anus and anal origin; last unbranched dorsal ray serrated; distance between distal dorsal serrae greater than width of their base; 8 branched pelvic rays; shallow groove in lower lip between jaw; dark scale bases, reticulated pattern; pectorals, pelvics and anal yellow/orange colour (life).	Kottelat (2001); Rainboth (1996).	Unable to count circumpeduncular scales, so cannot entirely rule out *H. malcolmi* (Smith).	RC0180 RC0181 YGN430
Labeo cf. *boga* (Hamilton)	One pair minute maxillary barbels; upper lip covered by rostral cap; lateral line complete (38 pored scales); 4½ scales between lateral line and pelvic base; 9½ branched dorsal rays; 5½ branched anal rays; dark spot above pectoral; dark bar on caudal peduncle.	Hamilton (1822); Talwar & Jhingran (1991).	Identification tentative, as literature cannot rule out alternative such as *L. ariza* (Hamilton), *L. bata* (Hamilton) and *L. kawrus* (Sykes). Most likely *L. boga*, however.	RC0671 RC0672
Labeo chrysophekadion (Bleeker)	Two pairs barbels; lips fimbriated; upper lip covered by rostral cap with broad lateral folds; dorsal large, with straight margin and 18½ branched rays; black body and fin colour.	Kottelat (2001).		RC0369 RC0370
Labeo cyclorhynchus Boulenger	Two pairs barbels (maxillary large and visible); lips plicate; snout large and rounded; upper lip covered by broad rostral cap; dorsal deeply concave with 12½ branched rays; variegated body colour pattern.	Tshibwabwa *et al.* (2006); Tshibwabwa & Teugels (1995).		RC0506 RC0507
Labiobarbus leptocheilus (Valenciennes)	Two pairs barbels (maxillary extending to not beyond centre of eye, rostral short); lips fimbriated; lateral line complete (36 pored scales); long dorsal fin (24½ branched rays); 5½ branched anal rays; approx. 10 rows spots forming longitudinal stripes.	Kottelat (2001); Roberts (1994).		RC0376
Labiobarbus ocellatus (Heckel)	Two pairs barbels; lips plicate; scales small (61 pored lateral line scales); long dorsal fin (28½ branched rays); no lateral stripes; ocellated humeral spot; ocellated spot on caudal peduncle and caudal base; fins without markings.	Kottelat *et al.* (1993); Roberts (1994).		RC0274 RC0275
Leptobarbus rubripinna (Fowler)	Two pairs barbels (maxillary barbel not reaching past centre of eye); lateral line complete, terminating on ventral half of caudal peduncle; 4½ scales between lateral line and dorsal origin; 7½ branched dorsal rays; no back blotch posterior to operculum; black midlateral stripe approx. ½–1 scale width; caudal lobes without black submarginal stripes; pelvic, anal, caudal red/orange (life).	Kottelat (2001); Kottelat *et al.* (1993); Rainboth (1996); Roberts (1989); Tan & Kottelat (2009).		RC0296 RC0460

Leuciscus idus (Linnaeus)	Barbels absent; mouth terminal; lateral line complete (53–56 pored scales); 8½–9½ branched dorsal rays; 11½ branched anal rays; posterior margin of anal concave.	Kottelat & Freyhof (2007).	Ornamental blue variety.	RC0570 RC0571
Luciosoma setigerum (Valenciennes)	Two pairs barbels (well developed); mouth large; snout strongly pointed; 7½ branched dorsal rays; 6½ branched anal rays; pelvic filaments extend to anal origin; semicircle of tubercles between nostrils absent; scattered tubercles on lower jaw and snout; dorsal positioned in posterior half of body; dark spots on caudal absent; midlateral stripe of indistinct spots, continuing onto caudal as submarginal stripe of upper lobe; median caudal rays not pigmented.	Kottelat (2001); Kottelat *et al.* (1993); Rainboth (1996); Roberts (1989).		RC0294 RC0295 YGN026 YGN488
Microdevario kubotai (Kottelat & Witte)	Barbels absent; lateral line absent; predorsal scales 10; narrow infraorbital 4; 7½ branched dorsal rays; 9½–10½ branched anal rays; concave distal margins of anal and dorsal; wide midlateral stripe, diffuse anteriorly; cleithral spot absent; no stripes on fins; black anal papilla absent; thin axial streak from above anus to caudal base.	Fang *et al.* (2009); Jiang *et al.* (2008); Kottelat & Witte (1999).		RC0234 RC0235 RC0492 RC0601 RC0602 YGN510
Microdevario nana (Kottelat & Witte)	As *M. kubotai*, but: distinct dark spot on tip of dorsal; diffuse spot on tip of anal; 10½–11½ branched anal rays; thin midlateral stripe, diffuse anteriorly; unpaired fins yellowish (life).	Fang *et al.* (2009); Jiang *et al.* (2008); Kottelat & Witte (1999).		EUN161 RC0618 RC0619 RC0620 RC0621 RC0622
Microrasbora rubescens Annandale	Barbels absent; supraorbital groove present; wide infraorbital 4; lateral line absent; predorsal scales 13; 7½–8½ branched dorsal rays; 10½–11½ branched anal rays; cleithral spot absent; no stripes on fins; black anal papilla; bright orange/red colouration with greenish lateral stripe (life).	Annandale (1918); Cottle (2010); Fang (2003); Fang *et al.* (2009); Jiang *et al.* (2008); Kottelat & Witte (1999).	These are a smaller, narrower, more colourful fish (2.8 cm TL), and perhaps better fit the description of *M. rubescens* (Annandale, 1918) than the *M.* cf. *rubescens* specimens. Found as possible bycatch with another lake Inle species, *Danio erythromicron*.	EUN162 RC0662
Microrasbora cf. *rubescens* Annandale	As *Microrasbora rubescens*, but: larger (4.3 cm TL), deeper bodied and bulkier; duller pinkish/orange hue (life).	Annandale (1918); Cottle (2010); Fang (2003); Fang *et al.* (2009); Jiang *et al.* (2008); Kottelat & Witte (1999).	These are larger fish than described by Annandale (1918). They are also less colourful. It is not exactly clear which of the *M. rubescens* specimens here are conspecific with the types, but these a poorer fit than the other specimens (RC0662, EUN162), and so are regarded for now as *M.* cf. *rubescens*. Additionally, Fang (2003) reports the supraorbital groove absent in her *M. rubescens* material. Very similar in appearance to *Devario* sp. "TW04" as presented by Cottle (2010).	RC0681 RC0682 RC0683 RC0684 RC0685
Mystacoleucus argenteus (Day)	Two pairs barbels; lateral line complete; procumbent predorsal spine; body deep and laterally compressed; eyes large; 8½ branched dorsal rays; last unbranched dorsal ray serrated; 6½ branched anal rays; dorsal origin anterior to pelvic origin; anal with concave distal margin; dorsal with black distal margin, becoming fainter posteriorly; strong black margin to caudal absent; dark scale base crescents absent.	Kottelat (2001); Talwar & Jhingran (1991).		EUN049
Myxocyprinus asiaticus (Bleeker)	Barbels absent; mouth small and inferior; lips papillated; ventral surface flat; high body, strongly laterally compressed; dorsal origin just posterior to pectoral base; dorsal, sail-like, terminating at caudal peduncle; variegated colouration with 4 dark bars.	Gao *et al.* (2008).	*Myxocyprinus* is a catostomid.	RC0203 RC0204
Neolissochilus cf. *stracheyi* (Day)	Two pairs barbels; lateral line complete (24+2 pored scales); last unbranched dorsal spine not serrated; 9½ branched dorsal rays; post labial groove interrupted (no median fleshy lobe on lower lip); tubercles on sides of snout and below eye; 3½ rows scales between dorsal origin and lateral line; dark midlateral stripe; back bronze and belly silver (life).	Chen *et al.* (1999); Day (1875); Kottelat (2001); Vidthayanon & Kottelat (2003).	Systematics of *Neolissochilus* is confused. Both *N. baoshanensis* (Chen & Yang) and *N. wynaadensis* (Day) are possible identifications, but tentatively, *N.* cf. *stracheyi* appears the most likely fit.	RC0365

Opsarius bakeri (Day)	One pair barbels (minute); lateral line complete; 10½–11½ branched dorsal rays; 13½ branched anal rays; single row 10–12 midlateral short bars/spots, becoming more elongated anteriorly; anal, dorsal and pelvics with black distal and white proximal stripes; caudal with white margins to lobes, and upper lobe with submarginal black blotch anteriorly.	Day (1865); Remi Devi *et al.* (2005); Talwar & Jhingran (1991).	Generic nomenclature follows Tang *et al.* (2010).	RC0377 RC0378
Oreichthys cosuatis (Hamilton)	Barbels absent; snout pointed; scales between pelvic origin and dorsal midline: ½, 6, ½; cephalic papillae present on head (arranged in rows); lateral line incomplete (4–5 pored scales); last unbranched dorsal ray not serrated; 8½ branched dorsal rays; 5½ branched anal rays; scales with dark bases: reticulate pattern; no spot on caudal peduncle; anal with indistinct median stripe/blotch; black subdistal margin on dorsal.	Schäfer (2009).	Schäfer (2009) reports 2–3 pored lateral line scales.	RC0470 RC0471
Oreichthys crenuchoides Schäfer	Barbels absent; snout blunt; scales between pelvic origin and dorsal midline: ½, 7, ½; cephalic papillae present on head (arranged in rows); lateral line incomplete; last unbranched dorsal ray not serrated; 8½ branched dorsal rays; 5½ branched anal rays; scales with dark bases: reticulate pattern; no spot on anal; spot on caudal base greater than ⅓ of peduncle depth; distal-anterior blotch on dorsal in females.	Schäfer (2009).	Frequently sold as *Oreichthys cosuatis*.	RC0050 RC0051
Oreichthys parvus Smith	Barbels absent; snout pointed; scales between pelvic origin and dorsal midline: ½, 6, ½; cephalic papillae present on head (arranged in rows); lateral line incomplete (6 pored scales); last unbranched dorsal ray not serrated; 8½ branched dorsal rays; 5½ branched anal rays; scales with dark bases: reticulate pattern; spot on caudal base less than ⅓ of peduncle depth; anal with spot; dark marking on tip of dorsal.	Schäfer (2009).		EUN207
Oreichthys sp. "red fin"	Barbels absent; snout blunt; scales between pelvic origin and dorsal midline: ½, 6, ½; cephalic papillae present on head (arranged in rows); lateral line incomplete (5–6 pored scales); last unbranched dorsal ray not serrated; 8½ branched dorsal rays; 5½ branched anal rays; scales with dark bases: reticulate pattern; blotch covering almost entire caudal peduncle; anal with spot; anterior subdistal blotch on dorsal continuing as median stripe (females), with no spot on dorsal in male; red colouration on body, caudal, dorsal and pelvics, anal in males (life).	Schäfer (2009).	Differs from *O. parvus* in snout shape and size of blotch on caudal base. Likely an undescribed species.	RC0638 RC0639
Osteochilus bleekeri Kottelat	Two pairs barbels; lips plicate; dorsal strongly concave anteriorly (11½ branched rays); last unbranched dorsal ray not serrated; 5½ branched anal rays; black blotch on proximal-anterior of dorsal; 6–7 rows lateral spots.	Kottelat (2008a); Kottelat *et al.* (1993); Roberts (1994).		RC0276 RC0659
Osteochilus microcephalus (Valenciennes)	Two pairs barbels; lips fimbriated and folded; mouth subinferior; tubercle at end of snout; 22 gill rakers; dorsal with 11½ branched rays; last unbranched dorsal ray not serrated; 5½ branched anal rays; wide midlateral stripe from operculum to caudal base; two rows of spots on dorsal.	Kottelat (2001, 2008a); Kottelat & Tan (2009); Kottelat *et al.* (1993); Roberts (1989).	More gill rakers (27–35) are reported by Kottelat (2008a), but fishes here are juveniles.	RC0217 RC0218
Osteochilus vittatus (Valenciennes)	Two pairs barbels; lips fimbriated and folded; mouth subinferior; snout tubercles absent; 5½ scale rows between dorsal origin and lateral line; last unbranched dorsal ray not serrated; scale rows with dark spots forming faint stripes; midlateral stripe absent; medium-sized blotch on caudal peduncle; fins red colour (life).	Kottelat (2001); Kottelat *et al.* (1993); Tan & Kottelat (2009)	Identification tentative as unable to count circumferential scales rows, so cannot effectively distinguish between *O. vittatus* and *O. kappenii* Bleeker. Specimens were wild-caught in Singapore, so based on distribution, *O. vittatus* is a more likely occurrence.	EUN038 YGN045
Paedocypris cf. *carbunculus* Britz & Kottelat	Scales absent; miniature size (up to 10 mm SL); modified pelvic fin in males forming keratinised "flange and hook" on anterior ray; pre-anal larval fin fold in females; single irregular row of mid-dorsal chromatophores; head blotch v-shaped; head-kidney pigment present; chest spots present; well developed chest blotch; opercular and branchiostegal rows of pigment; lips not heavily pigmented; red colouration (life).	Britz & Kottelat (2008); Kottelat *et al.* (2006).	*Paedocypris carbunculus* should have three rows of mid-dorsal chromatophores, and does not have a v-shaped head blotch (Britz & Kottelat, 2008). Likely an undescribed species, but conservatively regarded here as *P.* cf. *carbunculus*.	RC0222 RC0223
Paedocypris cf. *micromegethes* Kottelat, Britz, Tan, & Witte	Scales absent; miniature size (up to 10 mm SL); modified pelvic fin in males forming keratinised "flange and hook" on anterior ray; single row of mid-dorsal chromatophores; head-kidney pigment absent; overall, lightly pigmented; chest blotch present (distinct); red colour (life).	Britz & Kottelat (2008); Kottelat *et al.* (2006).	*Paedocypris micromegethes* should have a poorly developed or absent chest blotch, so these specimens are best referred as *P.* cf. *micromegethes*. Both specimens have different head blotch patterns, however, and are not regarded as conspecific with one another.	YGN554 EUN045

Pectenocypris korthausae Kottelat	Barbels absent; symphyseal knob present; pointed snout; elongate body shape; v. large number comb-like gill rakers (not counted); 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; dorsal origin above pelvic; lateral line incomplete (8 pored scales); round black spot on caudal base occupying 50% of peduncle; axial streak from operculum to caudal peduncle.	Kottelat (1982); Tan & Kottelat (2009).		RC0590
Poropuntius normani Smith	Two pairs barbels; mouth inferior; lateral line complete (28 +2–3 pored scales); lateral line with accessory ventral pore; last unbranched dorsal ray serrated; well defined dark stripe along margins of caudal lobes; yellow caudal (life).	Kottelat (2000, 2001).		RC0545 RC0546
Puntioplites proctozystron (Bleeker)	Barbels absent; lateral line complete; last unbranched anal ray thick and serrated posteriorly; last unbranched dorsal ray short, not reaching caudal; body plain with no markings; fins without orange colour.	Kottelat (2001); Kottelat *et al.* (1993); Taki & Katsuyama (1979).		RC0176 RC0177
Puntius arulius (Jerdon)	One pair maxillary barbels; mouth subterminal; lateral line complete; last unbranched dorsal ray smooth; dark band across caudal lobes absent; three large blotches on body (> 2 scales): large blotch mid body above pelvic origin, dark blotch above anal, dark blotch on caudal base; dorsal filaments absent in males.	Devi *et al.* (2010); Knight *et al.* (2011); Pethiyagoda & Kottelat (2005).	Frequently sold as *Puntius tambraparniei*.	RC0555 RC0556 RC0557 RC0558 RC0559
Puntius assimilis (Jerdon)	Lateral line complete; smooth last unbranched dorsal ray; one pair maxillary barbels (long); mouth inferior; dark band across caudal lobes; dark posterior lateral blotch; no markings on body anterior to anal origin.	Devi *et al.* (2010); Pethiyagoda & Kottelat (2005).	Some specimens small, but salient features discernible. There is diversity in the species, with three populations tentatively treated as conspecific, plus one synonym (*P. lepidus* Day). Frequently sold as *P. filamentosus*.	RC0132 RC0133 RC0134 RC0490 RC0491
Puntius aff. *banksi*	Two pairs long barbels; lateral line complete; last unbranched dorsal ray serrated; wedge-shaped marking beneath dorsal covering 3–4 scales; spot above anterior of anal; blotch on caudal peduncle.	Herre (1940); Kottelat & Lim (1995); Ng & Tan (1999); Rachmatika (2004).	Type material of *P. banksi* comprises two batches, viz. Singapore and Sarawak; Sarawak material (lectotype) comprises a species with elongate black bar at base of dorsal 1–2 scales in width, so likely not conspecific with Singapore material which matches these fish. Frequently sold as *P. banksi*.	RC0303 RC0393
Puntius chalakkudiensis Menon, Rema Devi & Thobias	One pair maxillary barbels; mouth inferior; lateral line complete (28 pored scales); smooth last unbranched dorsal ray; pronounced snout; black midlateral stripe with scarlet stripe above anteriorly; caudal with oblique dark distal band; dark median spot anteriorly on dorsal.	Day (1865); Menon *et al.* (1999); Prasad *et al.* (2008); Talwar & Jhingran (1991).		RC0537 RC0538 RC0539 RC0540 RC0541
Puntius chola (Hamilton)	One pair barbels (maxillary); mouth subterminal; 8½ branched dorsal rays; spot on caudal peduncle; proximal-anterior spot on dorsal branched rays 1–4; median-proximal row of dots above spot on dorsal.	Hamilton (1822); Silva *et al.* (2008); Talwar & Jhingran (1991).	Individual lacks iridescent pigments.	RC0730
Puntius conchonius (Hamilton)	Barbels absent; lateral line incomplete (8–13 pored scales); 8½ branched dorsal rays; deep body; dark blotch on caudal peduncle (no anterior blotches); dorsal with thick distal band.	Hamilton (1822); Talwar & Jhingran (1991); Vishwanath *et al.* (2007).		RC0001 RC0002 RC0084 RC0156 RC0371 RC0372 RC0373

Puntius denisonii (Day)	One pair barbels (maxillary); lateral line complete (28 pored scales); smooth last unbranched dorsal ray; mouth inferior; no pronounced snout; black midlateral stripe with scarlet stripe above anteriorly; caudal with oblique dark distal band.	Day (1865); Menon *et al.* (1999); Prasad *et al.* (2008); Talwar & Jhingran (1991).		RC0020 RC0106 RC0119 RC0150 RC0151 RC0712 YGN015 YGN114
Puntius dunckeri (Ahl)	Two pairs long barbels; lateral line complete; 8½ branched dorsal rays; last unbranched dorsal ray not serrated; colour pattern: see comments.	Ahl (1929); Kottelat *et al.* (1993).	Kottelat *et al.* (1993) and Ahl (1929) report *P. everetti* (Boulenger) with five round black spots, two above lateral line and two below, with a fifth spot on the caudal peduncle, and a bar posterior to the operculum. Examination of the type series [BMNH 1893.3.6.213–218(6)] confirms this. Specimens examined here do not appear to be conspecific with *P. everetti*, and although the description of *P. dunckeri* Ahl (1929) reveals little information and no types are known, the fish illustrated superficially matches these presented there, with strikingly larger blotches, and the midlateral bar above pelvics elongated to form a distinct bar. Frequently sold as *P. everetti*.	RC0017 RC0018 RC0145 RC0146 RC0147
Puntius erythromycter Kullander	Barbels absent; lateral line incomplete; lateral scale row curved; last unbranched dorsal ray serrated; 8½ branched dorsal rays; humeral marking absent; dark band around caudal peduncle; snout red (life).	Kullander (2008).		RC0603 RC0675 RC0676 RC0677 RC0678
Puntius fasciatus (Jerdon)	Two pairs barbels (maxillary longer than eye diam.); last unbranched dorsal ray not serrated; three scale rows between mid-dorsal row and lateral line; lateral line complete; four wide, irregular dark bars viz. oblique band between eyes, bar above pelvic, bar above anal, bar on caudal base.	Jayaram (1990); Jerdon (1849); Pethiyagoda & Kottelat (2005); Talwar & Jhingran (1991).	Possible diversity within the species, as four other names available in synonymy of *P. fasciatus*. Have chosen oldest available name in absence of modern treatment. Frequently sold as *P. melanampyx*.	RC0021 RC0022 RC0101 RC0102 RC0168 RC0169 RC0170 RC0353 RC0354 YGN267 YGN395

Puntius filamentosus (Valenciennes)	One pair maxillary barbels (short); lateral line complete; last unbranched dorsal ray not serrated; mouth subterminal; dark band across caudal lobes; dark posterior lateral blotch; no markings on body anterior to anal origin.	Pethiyagoda & Kottelat (2005). Devi *et al.* (2010).	Frequently sold as *Puntius assimilis*.	RC0007 RC0008 RC0116 RC0117 RC0118 RC0293 RC0299 RC0688
Puntius foerschi (Kottelat)	Two pairs barbels; lateral line complete (24 pored scales); 5½ branched anal rays; six dark bars; up to four spots between second, third and fourth bars.	Kottelat (1982); Kottelat *et al.* (1993).		RC0098 RC0099 RC0100 RC0665 RC0666
Puntius gelius (Hamilton)	Barbels absent; lateral line incomplete (up to 5 pored scales); last unbranched dorsal ray strongly serrated; 8½ branched dorsal rays; black band around caudal peduncle; black anterior spot on anal (not extending onto body); distinct black spots on pelvics; black spot on anterior base of dorsal; last unbranched dorsal ray not pigmented posterior to spot.	Bordoloi & Baishya (2006); Hamilton (1822); McClelland (1839); Vishwanath & Laisram (2004).	RC0135–RC0137 appear a larger fish with different form, but do not deviate significantly from the description. Frequently sold as *Puntius canius*.	RC0038 RC0039 RC0135 RC0136 RC0137 RC0604 RC0605
Puntius aff. *gelius*	Barbels absent; lateral line incomplete (up to 4 scales); last unbranched dorsal ray strongly serrated; 8½ branched dorsal rays; black band around caudal peduncle; black anterior spot on anal (extending onto body); distinct black spots on pelvics absent; black spot on anterior base of dorsal; last unbranched dorsal ray pigmented posterior to spot.	Bordoloi & Baishya (2006); Hamilton (1822); McClelland (1839); Vishwanath & Laisram (2004).	Differs from description of *P. gelius* in lacking spots on pelvics (RC0741 has v. faint marking). Also differs from my *P. gelius* in the anal fin spot extending well on to body and the pigmentation of last unbranched dorsal extending to tip (vs. not extending, and no dark pigmentation to tip). Appears as a smaller, more translucent fish. The description of *P. canius* (Hamilton) does not mention the pelvic spots, but Hamilton's illustrations published by McClelland (1839) show spots. *Puntius canius* is described as a smaller fish with a reddish hue; my material does not show a red colour, but this may be a seasonal, breeding effect. Bordoloi & Baishya (2006) report this colouration from specimens of "*P. ornatus*" Vishwanath & Laisram from Assam, and the specimens they picture appear similar, but are not *P. ornatus* as described (only markings being a band around caudal peduncle). I am reluctant to call my specimens *P. canius* or *P. ornatus*, and await further study. Frequently sold as *P. canius* or *P. gelius*.	RC0468 RC0469 RC0600 RC0740 RC0741

Puntius hexazona (Weber & de Beaufort)	Two pairs barbels; lateral line complete (but see comments); 5½ scales between dorsal and lateral line; six dark bars; dark spot below posterior base of dorsal absent.	Alfred (1963); Kottelat *et al.* (1993).	Specimens RC0361 and RC0362 appear to have incomplete lateral lines. They are referred to as *Puntius* cf. *hexazona*. Frequently sold as *P. pentazona*.	RC0046 RC0047 RC0048 RC0361 RC0362
Puntius jerdoni (Day)	Two pairs barbels (maxillary = eye diameter, rostral shorter); last unbranched dorsal ray not serrated; lateral line complete; 9½ branched dorsal rays; 6½ branched anal rays; 12 predorsal scales; colour silvery (life); fins orange (life) and tipped with black.	Day (1870, 1875); Talwar & Jhingran (1991).	Perhaps better referred to *Hypselobarbus*, but will follow Talwar & Jhingran (1991) in the absence of a modern treatment.	RC0611 RC0612
Puntius johorensis (Duncker)	Two pairs barbels; 4-5 dark stripes (wide, approx. 1 scale); stripes +1 and -1 on scale rows +2 and -2; no distinct axial streak below dorsal fin base.	Kottelat (1996).	Assigned as *P. johorensis*, but indistinct axial streak present on RC0641; number of stripes mostly lower than that reported by Kottelat (1996), but fits *P. johorensis* better than alternative species.	RC0379 RC0380 RC0381 RC0382 RC0383 RC0641
Puntius lateristriga (Valenciennes)	Two pairs barbels; deep body; lateral line complete; last unbranched dorsal ray serrated; two wide (2–4 scales) dark bars: first above pectoral, second wider, between dorsal and pelvics; dark midlateral stripe (1–2 scales) commencing anterior to anal, continuing onto caudal; spot above anterior of anal; RC0515 and RC0516 with more indistinct patterning comprising series of dark scale bases rather than solid lines, and midlateral stripe not extending into caudal.	Talwar & Jhingran (1991).	Six forms from the Malay Peninsula were recognised by Tweedie (1961): RC0302, RC0019 and RC0298 conform to the Johore form, while RC0515 and RC0516 conform to Perlis and Kedah form; these forms are not regarded as as conspecific in analysis, but the name *Barbus zelleri* Ahl may apply to Malay fishes.	RC0019 RC0298 RC0302 RC0515 RC0516
Puntius lineatus (Duncker)	Barbels absent; 5½ scale rows between dorsal origin and lateral line; mouth subinferior; fleshy lower lip forming continuous postlabial groove; longitudinal dark stripes.	Kottelat (1996).		EUN047
Puntius manipurensis Arunkumar & Tombi Singh	Barbels absent; lateral line incomplete (4 pored scales); 8½ branched dorsal rays; last unbranched dorsal ray serrated; small (one scale) humeral spot (not bar); small (one scale) caudal peduncle spot; 2–3 faint rows of spots in dorsal; spots absent from pelvic and anal; pigmented scale base; red colouration (life).	Arunkumar & Tombi Singh (2003); Kullander & Britz (2008); Linthoingambi & Vishwanath (2007); Menon *et al.* (2000).		RC0646 RC0647 RC0648 RC0649
Puntius nigrofasciatus (Günther)	Barbels absent; mouth subterminal; lateral line complete; last unbranched dorsal ray serrated; three complete dark bars above pectoral, pelvic and anal fins; oblique bar between eyes; scales with dark pigment at base.	Günther (1868); Kottelat & Pethiyagoda (1991); Pethiyagoda (1991); Talwar & Jhingran (1991).		RC0094 RC0095 RC0096 RC0149 RC0710
Puntius oligolepis (Bleeker)	One pair barbels; lateral line incomplete (6–7 pored scales); last unbranched dorsal ray not serrated; parallel rows of papillae on head; no bars or stripes; black distal margin to dorsal and anal; dark crescents along scale rows.	Kottelat *et al.* (1993); Tan & Kottelat (2008).		RC0014 RC0015 RC0016 RC0104 RC0311
Puntius orphoides (Valenciennes)	Two pairs barbels; last unbranched dorsal ray serrated; lateral line complete (29–31 pored scales); blotch on caudal peduncle; spot below dorsal origin; dark bar immediately anterior to operculum; caudal red with dark marginal stripes; dots along scale rows.	Kottelat (2001); Rainboth (1996).		RC0182 RC0183 RC0184 RC0185 RC0186 YGN004

Puntius padamya Kullander & Britz	One pair barbels (maxillary, small); lateral line incomplete (5–8 scales); last unbranched dorsal ray serrated; 2–3 rows dark spots on dorsal, pelvic and anal (males); vertical humeral blotch covering 3 scales; dark blotch on caudal peduncle; red colouration; base of scales heavily pigmented.	Kullander & Britz (2008).	Frequently sold as *Puntius ticto*.	RC0043 RC0044 RC0045 RC0152 RC0153 RC0711 YGN041 YGN056 YGN196 YGN404
Puntius pentazona (Boulenger)	Two pairs barbels; lateral line complete; 5½ scales between dorsal and lateral line; six dark bars; dark spot below posterior base of dorsal.	Alfred (1963); Kottelat *et al.* (1993).		RC0013 RC0304 RC0305 RC0306
Puntius rhomboocellatus Koumans	Two pairs barbels; lateral line complete; 5½ branched anal rays; 4½ scales between dorsal origin and lateral line; six irregular black bars with"ocellate rhombi" widening midlaterally; no spots between bars.	Alfred (1963); Kottelat (1982); Kottelat *et al.* (1993); Roberts (1989).		EUN232 RC0023 RC0024 RC0025 RC0154 RC0155 YGN076 YGN129
Puntius sahyadriensis Silas	Barbels absent; mouth subterminal; dorsal profile strongly convex; last unbranched dorsal ray not serrated, and also dark; pelvics black with white distal margins; scales with dark margin; up to seven irregular spots or vertical marks on sides.	Silas (1953).		RC0338 RC0339 RC0340 RC0341 RC0342
Puntius cf. *sarana* (Hamilton)	Two pairs barbels; lateral line complete (31+2 scales); last unbranched dorsal ray serrated; deep body; diffuse dark round blotch on caudal peduncle; rows of spots forming indistinct lateral stripes running along base of scales.	Hamilton (1822); Kottelat & Pethiyagoda (1991); Pethiyagoda (1991).	Much uncertainty this in identification, with 22 available names in the synonymy of *P. sarana*. Hamilton (1822) states two minute barbels, so maybe not this fish; here I follow Pethiyagoda (1991) and use the oldest available name pending a critical review.	RC0074
Puntius semifasciolatus (Günther)	One pair barbels, small; last unbranched dorsal ray serrated and shorter than adjacent branched ray; lateral line complete; series (up to seven) of irregular lateral marks (spots or bars), with last bar forming spot on caudal base.	Chang *et al.* (2006); Günther (1868); Kottelat (2001).	Frequently sold as *Puntius sachsii*.	RC0040 RC0041 RC0042 RC0093 RC0142 RC0673 RC0674
Puntius shalynius Yazdani & Talukdar	Barbels absent; lateral line incomplete (up to 11 pored scales); dark axial streak; last unbranched dorsal ray strongly serrated; 7½ branched dorsal rays; prominent first dark spot on peduncle above posterior of anal; indistinct second spot on caudal base; base of scales dark.	Yazdani & Talukdar (1975).	Yazdani & Talukdar (1975) reports orange/black fins, perhaps this material is immature?	RC0485 RC0486 RC0487 RC0488 RC0489

Puntius cf. *sophore* (Hamilton)	Barbels absent; mouth terminal; lateral line complete; last unbranched dorsal ray smooth; 8½ branched dorsal rays; dark proximal spot on branched dorsal rays 3, 4 and 5; dark spot on caudal peduncle and base; golden blotch on operculum; pelvic and anal yellow (life).	Hamilton (1822); Silva *et al.* (2008); Talwar & Jhingran (1991).	Much uncertainty in identification, with five available names in synonymy of *P. sophore*. Hamilton (1822) states four minute barbels, so probably not this fish. *Puntius stigma* (Valenciennes) may apply here, but I conservatively use the diagnosis of Talwar & Jhingran (1991), citing the oldest available name pending a critical review.	RC0658 RC0729
Puntius sp. "hybrid"	See comments.		Purported to be a hybrid of *P. denisonii* and *P. everetti*. Does not convincingly match any known *Puntius* species. The presence of a weak red stripe above the black midlateral stripe suggests *P. denisonii* may indeed be a parent.	RC0171 RC0172 RC0173 RC0174 RC0175
Puntius stoliczkanus (Day)	Barbels absent; lateral line complete; 8½ branched dorsal rays; last unbranched dorsal ray serrated (11-16 serrae); black vertical blotch on scales 3–4 above pectoral; black blotch on caudal peduncle; 2 black rows of spots on dorsal.	Hamilton (1822); Kottelat (2001); Linthoingambi & Vishwanath (2007).	Frequently sold as *Puntius ticto*.	RC0473 RC0474 RC0512 RC0576 RC0577 RC0718
Puntius tambraparniei Silas	One pair barbels; mouth terminal; lateral line complete; last unbranched dorsal ray not serrated; dark band across caudal lobes absent; four large blotches on body: two dark narrow bars under dorsal; dark blotch above anal, dark bar on caudal base; dorsal filaments present in males.	Devi *et al.* (2010); Knight *et al.* (2011); Pethiyagoda & Kottelat (2005).	Some specimens small, but salient features discernible. Frequently sold as *Puntius arulius*.	RC0010 RC0011 RC0012 RC0097 RC0528 RC0732
Puntius tetrazona (Bleeker)	One pair barbels; last unbranched dorsal ray serrated; lateral line incomplete; four vertical dark bars; dark proximal band on dorsal not extending onto body.	Alfred (1963); Kottelat *et al.* (1993).	Specimens here have an incomplete lateral line, but with 10–13 pored scales. Kottelat *et al.* (1993) reports 8–9 pored scales for *P. tetrazona*, and illustrates a fish with black pelvics (as does BMNH syntype 1867.11.28.178), but there is no mention on this in the literature. Identified as *P. tetrazona* (Bleeker) over *P. anchisporus* (Vaillant). Additional material (RC0742–RC0743) has 6–7 pored scales and 12 circumpeduncular scales, also conforming to *P. tetrazona*. Photos of wild (live) *P. anchisporus* with a clearly complete lateral line are nearly identical looking to the aquarium tiger barb. Photos of wild putative *P. tetrazona* with black pelvics are a quite different looking fish, although there has been a long history of selective breeding this fish. Retained for time being as *P. tetrazona*.	EUN103 EUN233 RC0004 RC0005 RC0006 RC0083 RC0140

Puntius tiantian Kullander & Fang	One pair barbels (maxillary, rudimentary); mouth subterminal; lateral line complete; 8½ branched dorsal rays; last unbranched dorsal ray thin and weakly serrated; large dark humeral bar; large dark blotch on caudal peduncle forming indistinct band.	Kullander & Fang (2005).		RC0501 RC0502 RC0503 RC0504 RC0505
Puntius ticto (Hamilton)	Barbels absent; lateral line incomplete (up to 11 pored scales); 24 scales in lateral series; 8½ branched dorsal rays; last unbranched dorsal ray serrated (13–15 serrae); dark spot on 3rd–4th lateral line scale; dark midlateral blotch above posterior of anal (on 17th–19th lateral scale); 1–2 rows of irregular spots on dorsal.	Hamilton (1822); Linthoingambi & Vishwanath (2007); Menon *et al.* (2000).	Linthoingambi & Vishwanath (2007) reports 15–17 serrae on last unbranched dorsal ray. *Puntius ticto* appears to vary geographically, and may comprise a complex of species.	RC0623 RC0624 RC0625
Puntius titteya Deraniyagala	One pair barbels; incomplete lateral line (3–5 pored scales); last unbranched dorsal ray weakly serrated; dark midlateral stripe from lip extending into caudal; bright red colour (life).	Deraniyagala (1930); Pethiyagoda (1991); Talwar & Jhingran (1991).		EUN230 RC0053 RC0054 RC0103 RC0141 RC0709
Puntius vittatus Day	Barbels absent; mouth terminal; last unbranched dorsal ray not serrated; lateral line incomplete (3–4 pored scales); scales with dark base and dotted margins; vertical blotch on dorsal; dark spot at base of caudal; pigmented anus.	Day (1865). citeTalwar1991.	Day (1865) describes and illustrates a fish with "four black spots" on the body viz. "one just before the dorsal, one under its posterior margin, another at the base of the caudal, and the fourth at the base of the anal. The dorsal has a black streak down it . . ." This fish only has three spots (only two on body), so identification may need to be revisited when modern literature is available.	RC0356 RC0357 RC0358 RC0359 RC0360 RC0650
Rasbora cf. *aurotaenia* Tirant	Barbels absent; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete (27+3 pored scales); 4½ scales between lateral line and dorsal origin; 2½ scale rows between lateral line and pelvic origin; dorsal origin closer to eye than caudal base; weak midlateral stripe (1 scale width) from operculum to caudal peduncle, superimposed onto axial streak.	Kottelat (1998, 2001, 2005); Kottelat *et al.* (1993).	Specimens in poor condition, so identification tentative.	RC0192 RC0193
Rasbora bankanensis (Bleeker)	Barbels absent; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete (20–22 pored scales); diffuse midlateral stripe superimposed over prominent axial streak; supra-anal stripe; fins unpigmented except prominent anterior subdistal spot on anal.	Siebert (1997).	Much variation in the size and position of the anal spot between batches. Perhaps a complex of species?	EUN012 EUN053 EUN203 RC0283 RC0284 YGN124
Rasbora borapetensis Smith	Barbels absent; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line incomplete (13–14 pored scales); midlateral stripe from operculum to caudal base, yellow iridescent stripe above (life); supra-anal stripe and subpeduncular streak present; caudal base red/orange (life); fins otherwise without colour.	Kottelat (2001); Smith (1934).		RC0591 RC0592
Rasbora brigittae Vogt	As *R. merah*, but: midlateral blotch and midlateral stripe confluent; red spots on caudal lobes (life).	Conway (2005); Conway & Kottelat (2011); Kottelat (1991); Kottelat & Vidthayanon (1993).	Characters do not appear consistent between *R. brigittae* and *R. merah*. Some examples of *R. merah* have confluent lateral stripe, but red spots on caudal, and examples of *R. brigittae* have red spots on caudal, but midlateral blotch resembling *R. merah*. Generic assignment follows Tang *et al.* (2010).	EUN223 RC0230 RC0231 YGN169 YGN179

Rasbora brittani (Axelrod)	Barbels absent; symphyseal knob absent; pointed snout; elongate body shape; 15 predorsal scales; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; dorsal origin posterior to pelvic; lateral line incomplete (10 pored scales), descending in steps; black spot on caudal base occupying 50% of peduncle.	Axelrod (1976); Kottelat (1991, 2008b); Liao *et al.* (2010); Tan & Kottelat (2009).	Generic assignment follows Tang *et al.* (2010).	EUN017 RC0636
Rasbora caudimaculata Volz	Barbels absent; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete; midlateral stripe present, but v. weak axial streak present; supra-anal stripe confluent with sub-peduncular streak; scale pigments giving distinct reticulated pattern throughout body; caudal with black tips; other fins without markings.	Brittan (1972); Kottelat *et al.* (1993).		EUN050 RC0595 RC0596
Rasbora cf. *cheeya* (1) (Liao & Tan)	Barbels absent; body bulky; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete, and not arranged in "step-like" pattern; dorsal origin anterior to pelvic origin; 9 predorsal scales; large eye; dark blotch in centre of dorsal, more like a bar; dorsal anterior to blotch, green-yellow colour (life).	Brittan (1972); Duncker (1904); Grant (2002); Liao *et al.* (2010); Liao & Tan (2011).	A larger fish than *Rasbora dorsiocellata*. Appears similar to *R. macrophthalma* Meinken, a species which should have an abbreviated lateral line. The positions of the dorsal fin as described by Grant (2002) is inconsistent with photographs in that article, so these are not regarded as *R. macrophthalma* until the original description or type material become available. Closest to *Brevibora cheeya*, but differs in predorsal scale count (should be 10–11), shape of dorsal blotch (should be round), and lateral line shape (should be "step-like". Generic assignment follows Tang *et al.* (2010).	RC0686
Rasbora cf. *cheeya* (2) (Liao & Tan)	Barbels absent; body bulky; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete, and not arranged in "step-like" pattern; dorsal origin anterior to pelvic origin; 9 predorsal scales; large eye; dark blotch in dorsal; fine, dark granulated chromatophores scattered evenly on head, body and fins.	Brittan (1972); Duncker (1904); Grant (2002); Liao *et al.* (2010); Liao & Tan (2011).	Specimens in poor condition, but closest to *Brevibora cheeya*. Differs, however, in predorsal scale count (should be 10–11) and lateral line shape (should be "step-like"). Not regarded as conspecific to RC0686 due to distinct pigment colour pattern on body and fins. Generic assignment follows Tang *et al.* (2010).	YGN431 EUN204
Rasbora cf. *dandia* (Valenciennes)	Barbels absent; symphyseal knob not pronounced; mouth terminal; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete (28–30 pored scales); ½, 4, 1, 1½ scales in transverse line between dorsal and pelvic origin; 13 predorsal scales; midlateral dark stripe greater than one scale width on caudal peduncle, and extending to median caudal rays; greenish lateral stripe above dark stripe (life).	Kottelat (1998, 2001); Silva *et al.* (2010).	Identification tentative. Does not conform to *R. daniconius* (Hamilton) s.s., but could be conspecific with Indochinese *R. daniconius* s.l. However, does not disagree with diagnosis of *D. dandia*, and so the name is used here conservatively in the absence of information on Indochinese *R. daniconius*.	RC0651 RC0652
Rasbora dorsiocellata Duncker	Barbels absent; body slender; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; 10–11 predorsal scales; dorsal origin approx. above pelvics; last unbranched dorsal ray not serrated; lateral line incomplete (7–8 pored scales), arranged in "step-like" pattern (see comments); round, dark blotch in centre of dorsal, not reaching last 2 branched rays, not bar-like.	Brittan (1972); Duncker (1904); Grant (2002); Liao *et al.* (2010); Liao & Tan (2011).	Liao *et al.* (2010) reports symphyseal knob absent. The "step-like" pattern of the pored lateral line scales was not clear in all specimens (some damaged), with variation apparent. Generic assignment follows Tang *et al.* (2010).	EUN051 RC0291 RC0663
Rasbora dusonensis (Bleeker)	Barbels absent; mouth subterminal; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete (26+3 pored scales); 10–11 predorsal scales; dorsal origin posterior to pelvic origin; 4½ scales between lateral line and dorsal origin; 1½ scale rows between lateral line and pelvic origin; 3 scale rows between lateral line and mid-ventral row; dorsal origin closer to eye than caudal base; diffuse midlateral stripe from operculum to caudal peduncle; axial streak ventral to midlateral stripe; weak black posterior margin to caudal.	Kottelat (1998, 2001, 2005); Kottelat *et al.* (1993).		RC0419

Rasbora einthovenii (Bleeker)	Barbels absent; symphyseal knob present on lower jaw; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete (28+2 pored scales); uneven, ventrally curved lateral stripe from snout to end of median caudal rays; reticulated scale pattern on dorso-anterior of body; purple hue (life).	Brittan (1972); Kottelat *et al.* (1993); Tan (2009).		RC0363 RC0364
Rasbora cf. *ennealepis* Roberts	Barbels absent; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete (26–27 pored scales); 10–11 predorsal scales; 2 rows of scales between lateral line and pelvic origin; caudal peduncle narrow; wide midlateral stripe (2 scales width), more intense posteriorly and superimposed over axial streak; precaudal spot absent; supra-anal stripe present; reticulate pattern weak; anterior anal rays weakly pigmented.	Kottelat (2000); Kottelat *et al.* (1993); Roberts (1989); Siebert (1997); Siebert & Guiry (1996).	Poor match to *R. ennealepis*, a species with 24–25 pored lateral line scales, 9 predorsal scales and a strongly reticulated scale pattern (Roberts, 1989). Roberts (1989) reported a sample from the Kapuas drainage with 10–11 predorsal scales and a lighter pattern. He regarded these as *R.* cf. *ennealepis*.	RC0660 RC0661
Rasbora espei Meinken	As *R. heteromorpha*, but: slimmer, less deep bodied; triangular, posterior black stripe smaller, markedly concave ventrally, forming distinct "lambchop" shape.	Brittan (1972); Duncker (1904); Kottelat *et al.* (1993); Kottelat & Witte (1999); Meinken (1956).	Generic assignment follows Tang *et al.* (2010).	EUN054 EUN235 RC0202 RC0496 RC0508 RC0509 YGN280 YGN282 YGN448
Rasbora gracilis Kottelat	Barbels absent; symphyseal knob absent; slender body shape; pointed snout; triangular-shaped operculum; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; dorsal high and strongly pointed; lateral line incomplete (0–4 pored scales); anal concave with elongated anterior rays; conspicuous, wide midlateral stripe continuing onto caudal; slender caudal peduncle.	Kottelat (1991); Liao *et al.* (2010).	Generic assignment follows Tang *et al.* (2010).	YGN117 YGN432
Rasbora hengeli Meinken	As *R. heteromorpha*, but: slimmer, less deep bodied; triangular, posterior black stripe markedly smaller: distance between pelvic origin and lower anterior edge of stripe equal to greatest width of stripe; colouration generally muted, with grey background colour and bright orange stripe above lateral stripe (life).	Brittan (1972); Duncker (1904); Kottelat *et al.* (1993); Kottelat & Witte (1999); Meinken (1956).	Generic assignment follows Tang *et al.* (2010).	YGN480
Rasbora heteromorpha Duncker	Barbels absent; symphyseal knob present on lower jaw; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; deep body, strongly laterally compressed; convex body (back) shape posterior to occiput; lateral line incomplete (up to 8 pored scales); conspicuous black stripe commencing posterior to dorsal origin, broader anteriorly covering most of body as triangle, or wedge shape, not concave ventrally; dark pigmentation to anterior dorsal and anal rays; pink/orange/red background colour to body (life).	Brittan (1972); Duncker (1904); Kottelat *et al.* (1993); Kottelat & Witte (1999); Meinken (1956).	Generic assignment follows Tang *et al.* (2010).	EUN236 RC0308 RC0597 YGN460 YGN506
Rasbora cf. *heteromorpha* Duncker	As *R. heteromorpha*, but: more slender, lacking convexity posterior to occiput; pigmentation on anterior dorsal/anal rays less distinct; orange/yellow anterior-subdistal blotch in anal.	Brittan (1972); Duncker (1904); Kottelat *et al.* (1993); Kottelat & Witte (1999); Meinken (1956).	Possibly an undescribed species. Generic assignment follows Tang *et al.* (2010).	RC0201 RC0307 YGN496
Rasbora kalochroma (Bleeker)	Barbels absent; symphyseal knob present on lower jaw; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; red colouration (life); two midlateral blotches (above pectoral and anal); no blotch on peduncle; indistinct posterior stripe from second blotch to end of median caudal rays.	Lim (1995); Tan (2009).		RC0450 RC0451 YGN133 YGN170 YGN377
Rasbora maculata Duncker	Barbels absent; scales present; lateral line absent; symphyseal knob weak or absent; miniature size; slender caudal peduncle; 7½ branched dorsal rays; 5½ branched anal rays; 10+9 principal caudal rays; dark lateral blotch anterior to pelvics (larger than pupil); black spot at caudal base; red and black pigmentation on anterior of dorsal and anal (life); conspicuous pigmentation absent between eye and maxilla.	Conway (2005); Conway & Kottelat (2011); Kottelat (1991); Kottelat & Vidthayanon (1993).	Generic assignment follows Tang *et al.* (2010).	RC0228 RC0229 YGN132 YGN178

Rasbora merah Kottelat	Barbels absent; scales present; lateral line absent; symphyseal knob weak or absent; miniature size; slender caudal peduncle; 7½ branched dorsal rays; 5½ branched anal rays; 7 pelvic rays; oval, longitudinally elongate midlateral blotch between pectoral and pelvic origin (surrounded by area free of pigment); irregular midlateral stripe from above anal origin to caudal peduncle; supra-anal spot; black spot on caudal base; black spot at caudal base; red spot on anterior of dorsal (life); conspicuous pigmentation absent between eye and maxilla; last unbranched anal ray pigmented; red colouration to body (life).	Conway (2005); Conway & Kottelat (2011); Kottelat (1991); Kottelat & Vidthayanon (1993).	See comments for *R. brigittae*. Generic assignment follows Tang *et al.* (2010).	RC0226 RC0227 YGN123
Rasbora naevus	As *R. maculata*, but: 9+8 principal caudal rays; sexually dimorphic lateral blotch (smaller in females).	Conway (2005); Conway & Kottelat (2011); Kottelat (1991); Kottelat & Vidthayanon (1993).	Generic assignment follows Tang *et al.* (2010). Conway & Kottelat (2011) report specimens of *Boraras* cf. *micros* in Tang *et al.* (2010) (GenBank EF452885 & HM224235) correspond to *R. naevus*. Frequently sold as *B.* sp. "red micros" or *B.* sp. "Thailand".	RC0224 RC0225
Rasbora pauciperforata Weber & de Beaufort	Barbels absent; symphyseal knob not distinct; slender body shape; pointed snout; triangular-shaped operculum; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line incomplete (6 pored scales); anal concave with elongated anterior rays; midlateral stripe ending at caudal base, with lighter red stripe above (life); series vertical streaks on anterior scales below midlateral stripe; supra-anal stripe and subpeduncular streak confluent.	Brittan (1972); Kottelat (1991); Kottelat *et al.* (1993); Liao *et al.* (2010); Weber & de Beaufort (1916).	Liao *et al.* (2010) reports symphyseal supra-anal stripe and subpeduncular streak absent. Generic assignment follows Tang *et al.* (2010).	RC0240 RC0241 YGN116 YGN290
Rasbora cf. *paucisqualis* Ahl	Barbels absent; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line incomplete (13–14 pored scales); no dorsal tubercles; elongate body; midlateral stripe diffuse anteriorly, ventral to axial streak anteriorly, becoming intense posteriorly and ending on caudal base; width of midlateral stripe 1½ scale rows; no precaudal spot; supra-anal stripe distinct; reticulate pattern weak, fins with no colouration.	Kottelat (2000, 2001, 2008b); Siebert (1997); Siebert & Guiry (1996).	*Rasbora paucisqualis* should have 22–27 pored lateral line scales (Siebert, 1997), so have conservatively named these fish *R.* cf. *paucisqualis*.	EUN032 EUN229 RC0255 RC0256
Rasbora paviana Tirant	Barbels absent; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete; distinct midlateral stripe starting at operculum, narrow anteriorly (½ scale row width), terminating in contiguous diamond-shaped blotch on caudal base; axial streak superimposed on midlateral stripe for much of length; weak supra-anal pigments; fins without markings.	Kottelat (1998, 2001, 2005).		RC0194 RC0195
Rasbora rasbora (Hamilton)	Barbels absent; symphyseal knob present; mouth terminal; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete; weak supra-anal stripe; diffuse lateral stripe from operculum to caudal base; subpeduncular streak present; scale pigments giving weak reticulated pattern; caudal yellow (life) with black lobes and posterior margin.	Brittan (1972); Hamilton (1822); Silva *et al.* (2010).		RC0191 RC0513 RC0514
Rasbora rubrodorsalis Donoso-Büchner & Schmidt	As *R. borapetensis*, but with: (7–8 pored lateral line scales); red/orange blotch on anterior dorsal base (life).	Kottelat (2001).		RC0630 RC0631
Rasbora sarawakensis Brittan	Barbels absent; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; body depth 30% in SL; lateral line complete (25 pored scales); tubercles present on dorsal surface; midlateral stripe distinct and of even intensity throughout; supra-anal stripe distinct; subpeduncular streak absent; dorsal and anal fins with dark pigmentation to anterior rays.	Brittan (1972); Kottelat *et al.* (1993); Roberts (1989).		RC0632 RC0633
Rasbora sp. "undet. (1)"	Barbels absent; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete; midlateral stripe from operculum to caudal peduncle, widest under dorsal, and terminating in triangular spot; axial streak above, but not separate from midlateral stripe until anterior to anal origin; supra-anal stripe present; distinct reticulate scale pattern; caudal yellow (life) with black tips and thin posterior margin.	Kottelat (1998, 2001, 2005); Kottelat *et al.* (1993); Tan & Kottelat (2009).	Likely member of the *R. sumatrana* group. Similar to *R. vulgaris* Duncker, *R. notura* Kottelat and *R. hosii* Boulenger, but cannot confidently match due to differences in midlateral stripe arrangement.	RC0574 RC0575
Rasbora trilineata Steindachner	Barbels absent; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete; midlateral stripe fading anteriorly and widening posteriorly; supra-anal stripe confluent with sub-peduncular streak; scale pigments giving weak reticulated pattern (anteriorly); caudal with oblique subterminal bars and white tips.	Brittan (1972); Kottelat *et al.* (1993); Rainboth & Kottelat (1987); Roberts (1989).		RC0205 RC0206

Rasbora urophthalmoides Kottelat	Barbels absent; scales present; lateral line absent; symphyseal knob weak or absent; minature size (up to 12.4 mm SL); slender caudal peduncle; 7½ branched dorsal rays; 5½ branched anal rays; midlateral stripe from operculum to caudal peduncle; black spot at caudal base; conspicuous pigmentation present between eye and maxilla; last unbranched dorsal ray pigmented; red spots on caudal lobes absent (life).	Conway (2005); Conway & Kottelat (2011); Kottelat (1991); Kottelat & Vidthayanon (1993).	Generic assignment follows Tang *et al.* (2010).	RC0232 RC0233
Rasbora vulcanus Tan	Barbels absent; symphyseal knob present; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete; 10 predorsal scales; midlateral stripe from operculum to caudal base; supra-anal stripe and subpeduncular streak present; dorsal, anal and caudal with weak subdistal dark margins; axial streak not distinct; distinct reticulate scale pattern; golden orange colour of body and fins (life).	Tan (1999).		RC0279 RC0588 YGN034 YGN182 YGN342
Rasbora wilpita Kottelat & Pethiyagoda	Barbels absent; symphyseal knob pronounced; well developed lateral maxillary process; body depth 25–28% in SL; 7½ branched dorsal rays; 5½ branched anal rays; last unbranched dorsal ray not serrated; lateral line complete (29–31 pored scales); ½, 4, 1, 1½ scales in transverse line between dorsal and pelvic origin; 13 predorsal scales; midlateral dark stripe greater than one scale width on caudal peduncle; upper margin of stripe distinct, with lower margin indistinct giving jagged appearance.	Silva *et al.* (2010).		RC0285 RC0584
Rasboroides vaterifloris (Deraniyagala)	Barbels absent; symphyseal knob present; deep laterally compressed body shape; 7½ branched dorsal rays; 6½ branched anal rays; last unbranched dorsal ray not serrated; lateral line incomplete (up to 3 pored scales); anal strongly concave with rays elongated anteriorly; orange colour of body and fins, with caudal hyaline and orange lower lobe (life).	Brittan (1972); Deraniyagala (1930); Pethiyagoda (1991).		EUN048 RC0281 RC0282
Rhodeus ocellatus (Kner)	Barbels absent; anal origin before end of dorsal base; lateral line incomplete (up to 4 pored scales); 12½ branched dorsal and anal rays; posterior midlateral stripe, starting after pelvic base; caudal with red median stripe (life); white anterior margin of pelvics (life); 2 rows of white spots along median dorsal rays (life).	Arai & Akai (1988); Nakabo (2002).	Conforms to *R. ocellatus ocellatus*.	RC0572 RC0573
Rohtee ogilbii Sykes	Barbels absent; lateral line complete; 8½ branched dorsal rays; 13½ branched anal rays; last unbranched dorsal ray serrated; ventral edge of body sharp and keel-like between pelvics and anal; procumbent predorsal spine (concealed by scales); body deep and laterally compressed; silvery colour (life) with 5 black bars; spot on caudal peduncle.	Day (1865); Sykes (1839, 1841); Talwar & Jhingran (1991).	Matches Talwar & Jhingran (1991) and Day (1865) well, but Sykes (1839) does not mention black bars. Specimen may be a juvenile.	RC0609
Sawbwa resplendens Annandale	Barbels absent; scales absent; last unbranched dorsal ray serrated; 7½ branched dorsal rays; 5½ branched anal rays; body with scattered chromatophores.	Annandale (1918).		EUN173 RC0161 RC0162 YGN396
Sundadanio cf. *axelrodi* (Brittan)	Barbels absent; lateral line absent; symphyseal knob present; head blunt; caudal peduncle slender; miniature size (up to 20 mm TL); 6½ branched dorsal rays; 5½ branched anal rays; posterior margin of anal concave; sexually dichromatic, males with more intense colouration.	Brittan (1976); Kottelat & Witte (1999); Roberts (1989).	Sold in aquarium trade as three colour varieties: red, blue, green. Likely a complex of species. Mostly female specimens here, so hard to characterise diagnostic male colour patterns and match specimens to type material, so all regarded here as *S.* cf. *axelrodi*.	EUN099 EUN231 RC0236 RC0237 RC0238 RC0239 YGN073 YGN119 YGN120 YGN121

Tanakia himantegus (Günther)	One pair barbels (greater than eye diameter); anal origin before end of dorsal base; lateral line complete; 8½ branched dorsal rays; 10½ branched anal rays; median row of elongated spots on dorsal membrane; anal with black distal stripe and red median stripe (life); midlateral stripe starting above pelvis base, widening posteriorly and continuing onto caudal; red distal band on dorsal (life); upper of iris red (life); midlateral spot above pectoral.	Arai & Akai (1988); Chang *et al.* (2009); Günther (1868); Nakabo (2002).	Conforms to *T. himantegus himantegus*.	RC0466 RC0467
Tanichthys albonubes Lin	Barbels absent; symphyseal knob absent; lateral line absent; posterior and anterior nostrils confluent; 6½ branched dorsal rays; 8½ branched anal rays; row cornified tubercles on snout of male; dark midlateral stripe terminating as spot on caudal base, with light stripe above; dark stripe narrower than light stripe; distance between dorsal origin and top of light stripe half of distance between anal origin and bottom of dark stripe; body below dark midlateral stripe dark coloured; dusky caudal with red blotch at centre and base (life).	Freyhof & Herder (2001); Liang *et al.* (2008); Weitzman & Chan (1966).		EUN234 RC0442 RC0449
Tanichthys micagemmae Freyhof & Herder	As *T. albonubes*, but: dark midlateral stripe wider than light midlateral stripe; distance between dorsal origin roughly equal or greater than distance between anal origin and dark stripe; body below dark midlateral stripe light coloured.	Freyhof & Herder (2001); Liang *et al.* (2008); Weitzman & Chan (1966).	Tubercles not observed in these specimens, as all female.	EUN011 RC0478 RC0479 YGN259 YGN420