Investigating interspecific hybridisation in ornamental fishes

Rupert A. Collins
Bio-Protection Research Centre
PO Box 84
Lincoln University 7647
Canterbury
New Zealand
rupertcollins@gmail.com

September 27, 2012

Abstract

Biosecurity agencies dealing with the trade in ornamental fishes have a responsibility to accurately identify the exotic organisms crossing their borders. Molecular methods such as DNA barcoding have been proposed as tools to assist in this process. However, hybrid fishes present in the ornamental aquaculture industry represent a potential problem; standard techniques such as DNA barcoding using mitochondrial genes are known to be inappropriate and potentially misleading in situations where this interspecific hybridisation has occurred. One potential solution is with the addition of genetic data from a nuclear gene. Standardised nuclear barcoding genes have, however, received little attention. The aim of this study is to assess a selection of nuclear genes using genomic resources, and investigate whether a chosen locus can: (1) detect the presence of a hybrid using sequence heterozygosity; and (2) estimate the maternal and paternal species when combined with mtDNA. The nuclear rhodopsin gene (RHO) was chosen as a potential candidate, being relatively variable among species. An experiment using laboratory-bred control hybrids (*Danio* spp.) indicates that sequence heterozygosity can show hybridisation, and that parentage can be elucidated from the nuclear and mitochondrial data. However, when tested upon putative hybrid fishes collected from the aquarium trade, results were mixed, suggesting further work is required before a universal technique can be proposed.

Introduction

In order to comply with international and local regulations, an important aspect of managing biosecurity in the ornamental aquarium trade is the identification of imported fishes to species level (MAF Biosecurity New Zealand, 2011). However, identifying captive bred and mass-produced

domesticated organisms presents unique problems for standard, visual identification procedures. Loss of diagnostic characters may occur in ornamental fishes due to the process of artificial selection for retail purposes (McDowall, 2004), while interspecific hybrids may have unpredictable intermediate phenotypes (Mallet, 2005), making them additionally challenging to identify. 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), and this may have implications in terms of biosecurity (disease vectoring or invasion success).

Interspecific hybrids have long been used in the aquaculture industry 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 trade is now producing increasing numbers of novel hybrid organisms 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, indiscriminate breeding of hybrids at farms. Therefore, when dealing with hybrid-risk groups, it is required that biosecurity agencies be confident in their identifications.

Identifying interspecific hybrids

Using mtDNA DNA barcoding (*sensu* Hebert *et al.*, 2003) using the mitochondrial COI gene has been proposed as a potentially useful tool for fish identification, especially in cases were morphological identification is challenging (Ward *et al.*, 2009). However, despite being well suited to identifying specimens with an atypical phenotype, there are situations where mitochondrial DNA may be inappropriate or may lack properties desirable to make suitable biological inferences.

One such example is in the detection of interspecific hybrid organisms (Aliabadian *et al.*, 2009; Dasmahapatra & Mallet, 2006). The 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 a country based upon positive barcode identification of the maternal species. This may have implications for biosecurity risk assessments, with life history data and nomenclature becoming associated with the maternal species only. Valuable information could be lost by using the standard COI approach alone, and misleading conclusions could be reached regarding the identification of query specimens. Clearly, other methods are required.

Using 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 genotyping 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.

Using 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.

Using 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 a phylogeny for both parental species and related taxa. 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 standalone 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, and is a hypothesis worthy of further investigation.

Nuclear marker selection

The *a priori* choice of an appropriate nuclear marker for this task 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.

Objectives

Here, the aims of this study are to: (1) assess a range of potential nuclear markers for variability at the species level; (2) test suitability of a single marker as a stand-alone hybrid detection test using sequence heterozygosity; and (3) attempt to reconstruct the parentage of interspecific hybrids of known and unknown parentage, using both nDNA and mtDNA.

Materials and methods

Nuclear marker selection

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

Step one: 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 between model species were calculated for each gene using MEGA4 (Tamura *et al.*, 2007).

Step two: 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).

Step three: comparison with COI Finally, a single marker was selected for testing across a wider range of species within the Cyprinidae, and was compared to information from the COI barcode region. This work has been published in Collins *et al.* (2012a); a subset of 200 individuals were amplified for both nuclear rhodopsin (RHO) and COI, comprising 82 species (Collins *et al.*, 2012a, supplementary dataset S2). Patterns in agreement between matched RHO and COI subsets were investigated using the nearest neighbour (k-NN) measure of identification success (Collins *et al.*, 2012a).

PCR protocols for nuclear genes

Nuclear data for the five shortlisted genes (Table 1) were generated with the following lab protocol. DNA extractions were as outlined in Collins *et al.* (2012a). Optimised PCR reactions were carried out using a GeneAmp 9700 thermocycler (APPLIED BIOSYSTEMS) in 10 µl reactions of: 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 1. Sequencing protocol was as reported in Collins *et al.* (2012a).

Breeding interspecific hybrids

To compare how effectively sequence data can identify introgression and estimate parentage, experimental control 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 25 litre 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 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. 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).

Detecting hybrids

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 200 putative non-hybrid cyprinid fishes collected previously and sequenced for RHO (Collins *et al.*, 2012a, supplementary dataset S2), and from fishes more generally (RHO sequences on GenBank). When assessing heterozygosity in the data generated in this study, the polymorphic positions were scored by visually assessing each chromatogram in FINCHTV 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 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 2.15 (R Development Core Team, 2010) using *grep* and the *seqStat* command of SPIDER (Brown *et al.*, 2012). The "N" code (all bases) was excluded.

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 (Collins *et al.*, 2012a, supplementary dataset S1) was used to reveal the maternal species, so a putative paternal nuclear sequence can be calculated by resolving the ambiguities in the hybrid

Table 1. 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

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 (Collins *et al.*, 2012a, supplementary dataset S2) using the "best close match" (BCM) method of identification (Meier *et al.*, 2006); the threshold was optimised for the RHO data using the *threshOpt* function of SPIDER. This method was tested with both the lab bred *Danio* hybrid and a putative hybrid *Puntius* purchased in the aquarium trade (Collins *et al.*, 2012a, specimen RC0171, BOLD process ID number RCYY136-11).

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 the Barcode of Life Database (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 and added to the downloaded alignments. The primers Glu-2 and Pro-R1 (Hardman & Page, 2003) were used for amplification. 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); 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 cyt *b* sequences were aligned by translated protein in MEGA4 (Tamura *et al.*, 2007), and trimmed to 1,200 bp. Neighbour-joining (NJ) phenograms were constructed using uncorrected *p* distances in the APE package for R (Paradis *et al.*, 2004), following (Collins *et al.*, 2012b). Negative branch lengths were set to zero. The hybrids were also sequenced for RHO using methods outlined previously, to detect polymorphisms.

Results

Nuclear marker selection

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 2. The diversity of these genes across the five model organisms is presented in Figure 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 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.

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. A total of 30 sequences were generated from the six *Danio* species with these nuclear genes. Primers and citations are presented in Table 1. GenBank accession numbers for the sequences generated here are presented in Table 3.

Step three: one locus 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 2). This RHO fragment corresponds 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). The utility of the RHO marker was also tested by Collins *et al.* (2012a), with 200 RHO and COI sequences generated for 82 species of cyprinid fish. The identification success rates "for the nearest neighbour analyses (k-NN) were 99.0% for COI, and 92.2% for RHO" (Collins *et al.*, 2012a).

Interspecific hybrids

Tank bred *Danio* 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 is shown in Figure 3. This hybrid was then sequenced for four of the shortlisted 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 4), with the RHO gene showing the most polymorphic positions in the hybrid (32), compared to the other nuclear genes. Figure 4 shows a section of a trace file chromatogram for the hybrid *Danio*, with corresponding double peaks in both forward and reverse reads.

Baseline heterozygosity For the 200 RHO sequences of putative non-hybrid cyprinid fishes from Collins *et al.* (2012a), 95% had $\leq$ 4 het-

Table 2. Names of 22 candidate nuclear loci, with length (bp), citation, and Ensembl reference data (for *Danio rerio* sequences). Nomenclature follows literature cited. LSU 28S is not available on Ensembl, so GenBank reference is included. Abbreviations: ref. = reference.

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

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

Estimating parentage Using the *Danio rerio* RHO sequence (RC0394, BOLD process RCYY315-11) as the maternal species for the lab bred hybrid, a composite paternal sequence was gener-

Table 3. GenBank accession numbers for sequences generated from five candidate nuclear loci. Notes: (*) *Danio albolineatus* sequence from RC0445; (†) *D.* aff. *kyathit* sequence from RC0120.

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

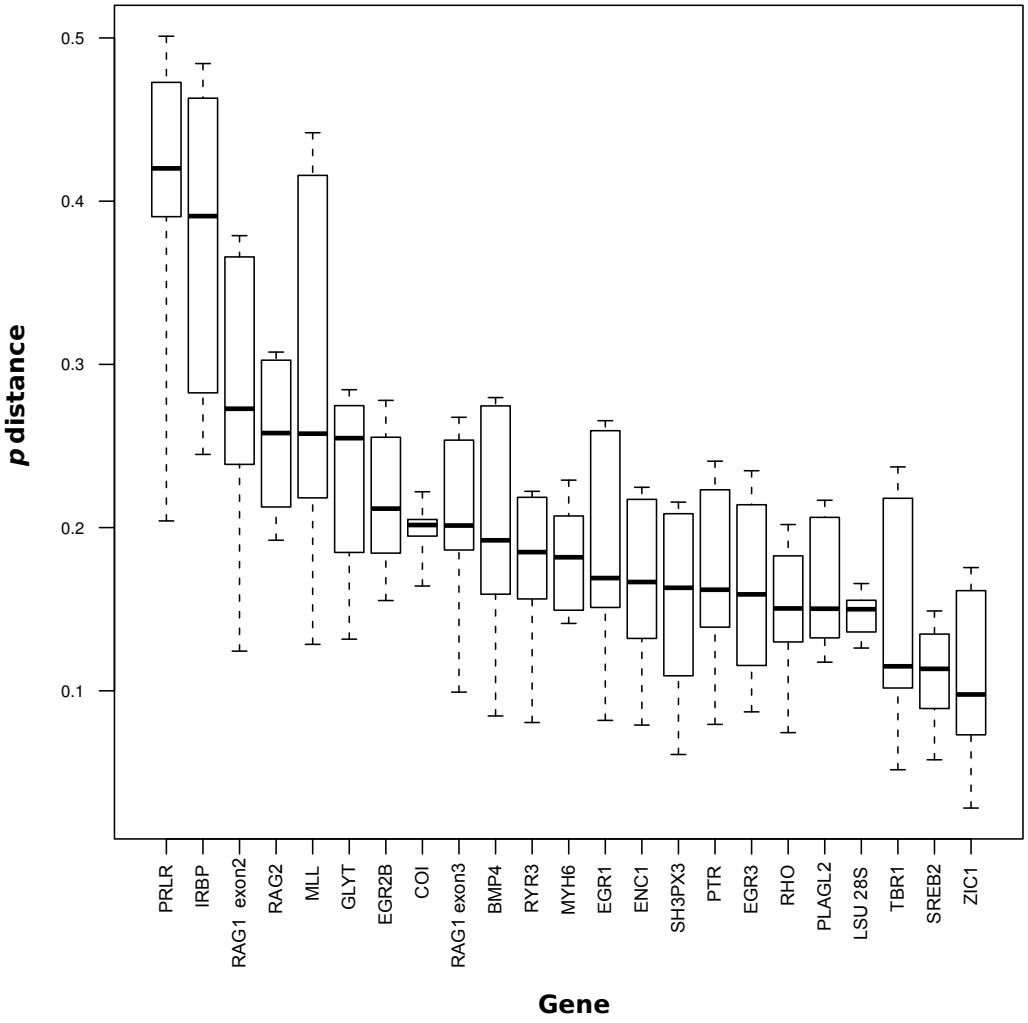


Figure 1. Uncorrected pairwise *p* distance ranges for 22 homologous candidate nuclear loci (and COI) between the following model organisms: *Danio rerio*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Takifugu rubripes* and *Tetraodon nigroviridis*. Whiskers extend to full range of data; boxes represent quartiles; black lines show median values.

ated. 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 (RC0171, BOLD process RCYY136-11), 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 (Collins *et al.*, 2012a, supplementary dataset S2), being over 3% different from the closest match (*P. arulius*), and well above the 1.4% intraspecific threshold for this dataset (Collins *et al.*, 2012a). 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, be-

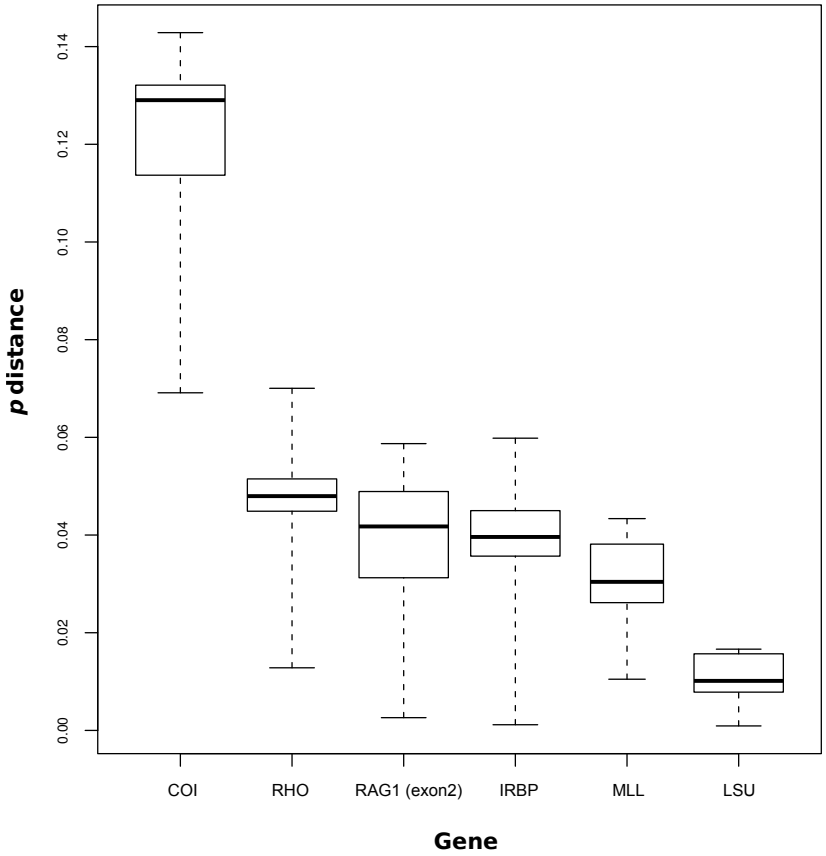


Figure 2. Intrageneric uncorrected pairwise *p* distance ranges between candidate nuclear loci from the following *Danio* species: *Danio* aff. *kyathit*, *D. albolineatus*, *D. dangila*, *D. kyathit*, *D. margaritatus* and *D. rerio*. Whiskers extend to full range of data; boxes represent quartiles; black lines show median values.

ing 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 (Figure 5), four of which were close to species represented in the GenBank data. The full cyt *b* alignment for hybrid and GenBank *Synodontis*

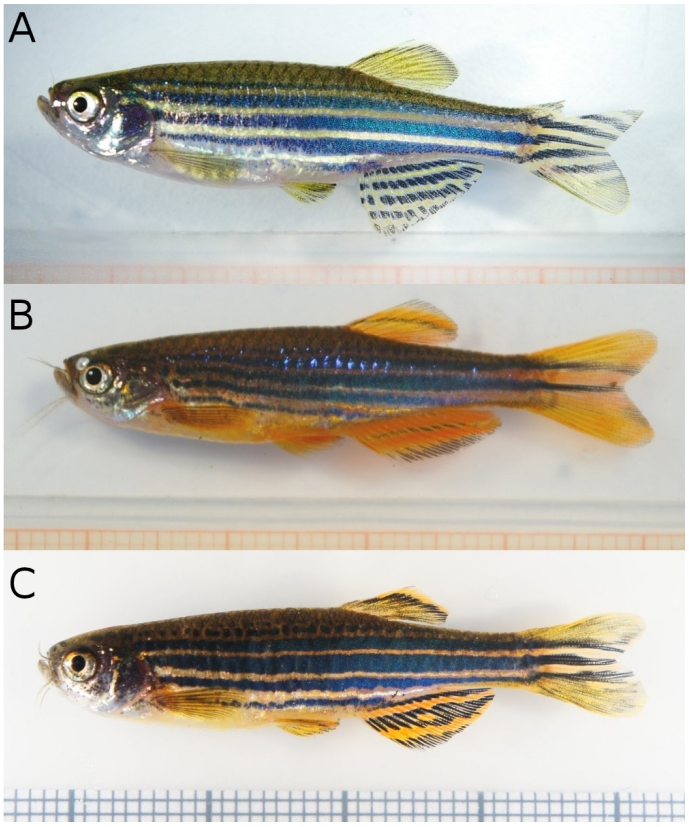


Figure 3. Phenotype of laboratory bred *Danio rerio* × *D*. aff. *kyathit* (C), parental species phenotype of *Danio rerio* RC0067 (A), and *D*. aff. *kyathit* RC0120 (B).

data is presented as supplementary information http://dx.doi.org/10.6084/m9.figshare.96146. Images of the hybrid *Synodontis* are shown in Figure 6. 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 the *Synodontis* RHO data that was generated.

Discussion

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 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 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

Table 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

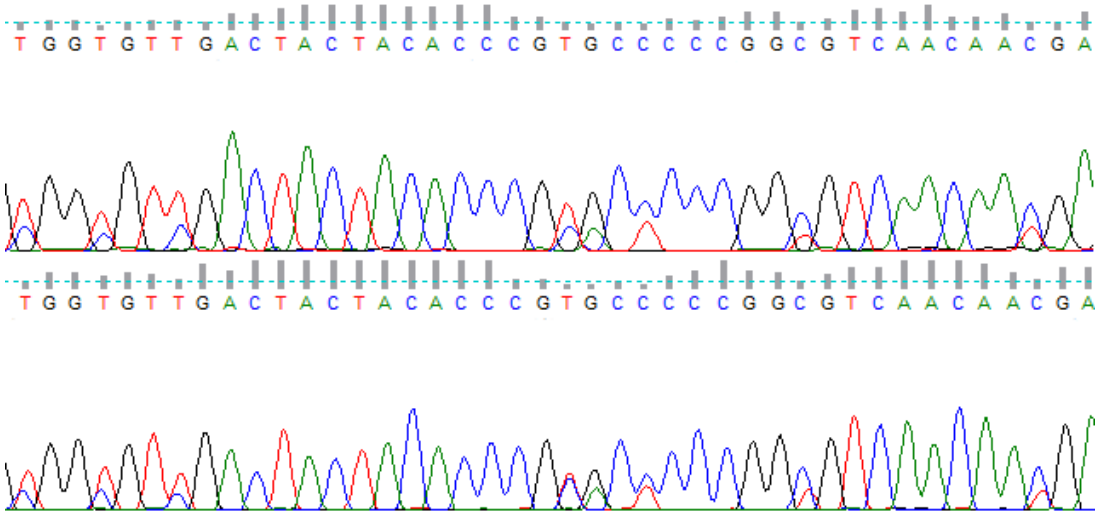


Figure 4. Chromatogram trace files from FINCHTV for interspecific hybrid RC0455 (laboratory bred *Danio rerio* × *D.* aff. *kyathit*), showing multiple heterozygous positions in both forward (top) and reverse (bottom) reads). Note the low quality scores around the polymorphisms.

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 2, Table 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). Collins *et al.* (2012a) reported that 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 may offer a suitable method of verifying mitochondrial results in terms

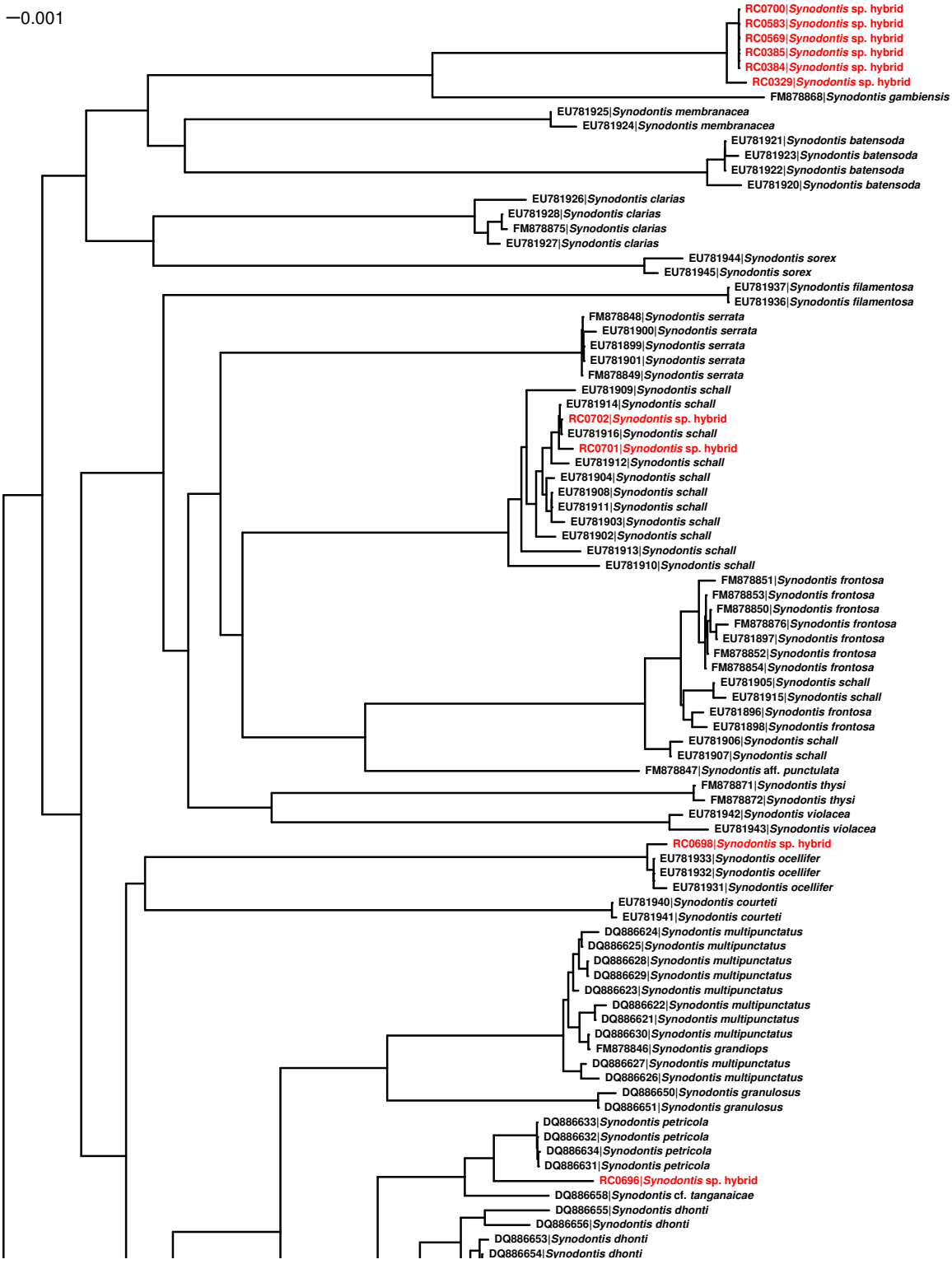


Figure 5. Neighbour-joining phenogram of mitochondrial cyt *b* sequences (uncorrected *p* distances) for 163 *Synodontis* specimens. Putative hybrids collected in this study are labelled in red.

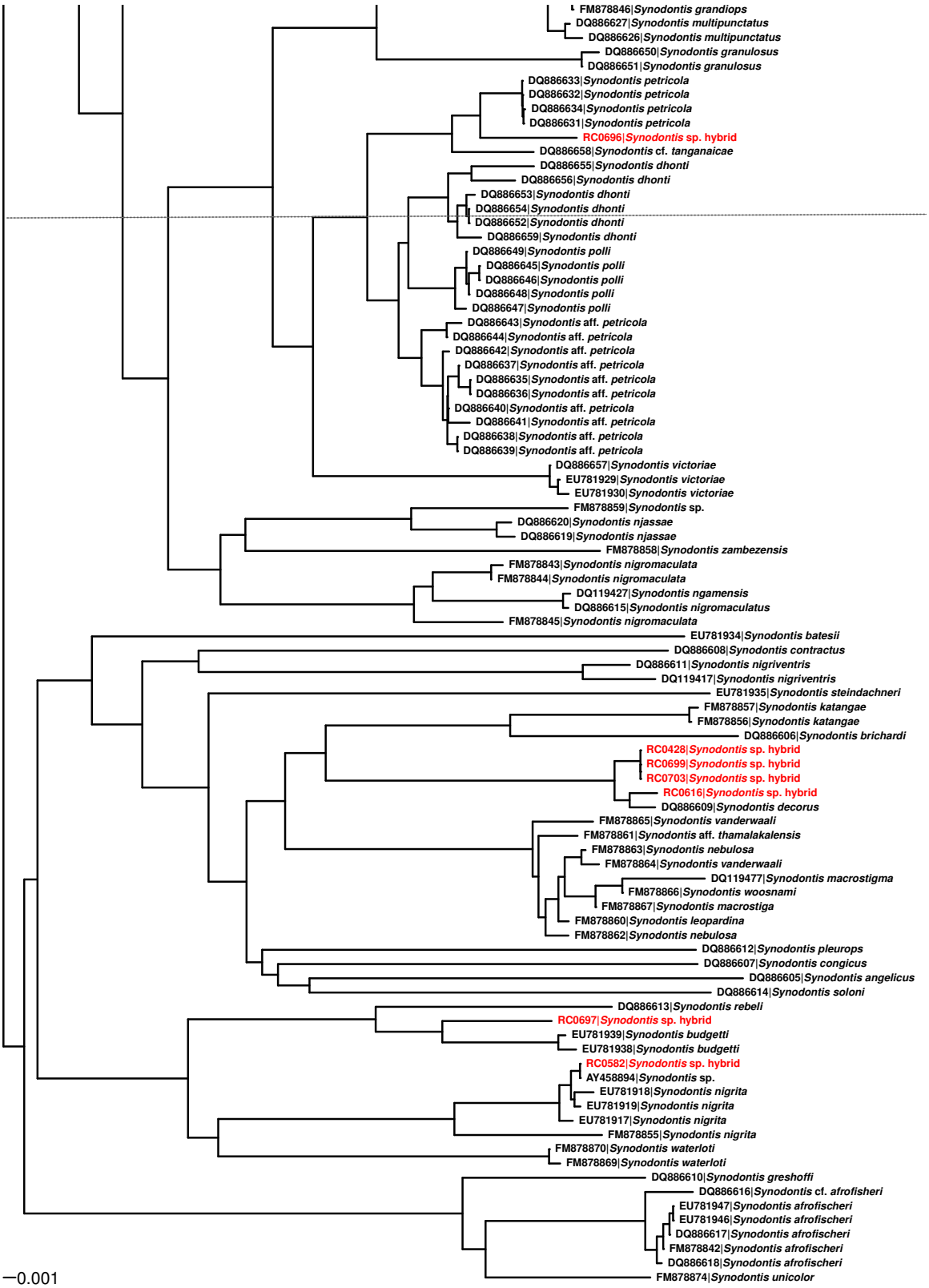
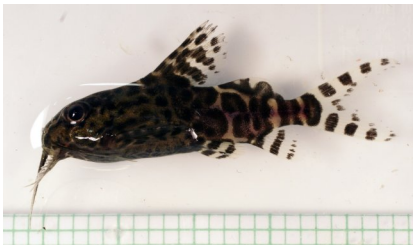
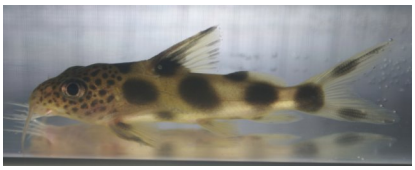


Figure 5. (Cont). Dotted line signifies end of previous part of figure.



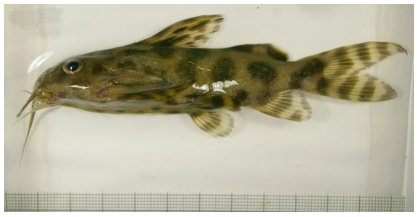
(a) RC0329

(b) RC0384

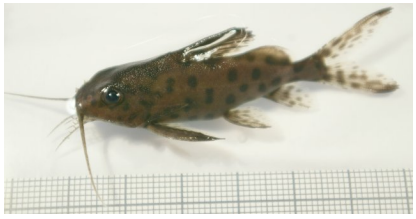


(c) RC0428

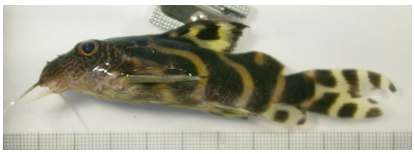
(d) RC0569



(e) RC0582

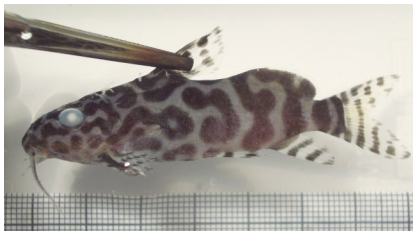


(f) RC0583

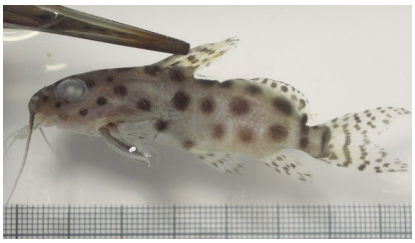


(g) RC0616

(h) RC0696

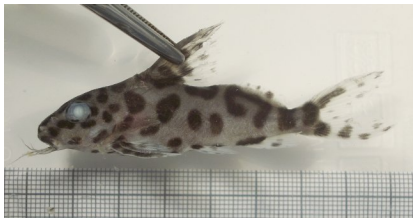


(i) RC0697



(j) RC0698

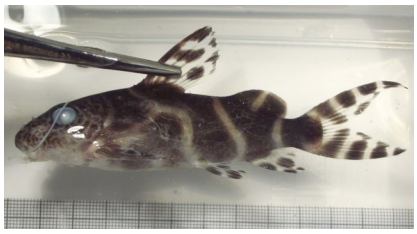
(k) RC0699



(l) RC0700

(m) RC0701

(n) RC0702



(o) RC0703

Figure 6. Images of 15 putative *Synodontis* hybrid specimens collected in the trade.

of hybridisation.

However, when using some nuclear protein-coding genes, some potential pitfalls may be apparent. 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?

Interspecific hybrids

The breeding of aquarium hybrids in a controlled environment provided an important opportunity to test how 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 (Collins *et al.*, 2012a, supplementary table S1).

The *Synodontis* catfishes are well known subjects of hybridisation in the aquarium trade (Ng, 2010), and appear to derive from a diverse selection of maternal species (Figure 5). However, the RHO protocol used here failed to yield consistently clean PCR products or sequence data. From those that were sequenced, the amount of poly-

morphism 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 or SNP markers. 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.

Summary

In this study, a range of nuclear loci were assessed in their ability to detect ornamental fishes of hybrid origin using a simple nDNA heterozygosity test, a hypothesis which worked as predicted for controlled, lab bred hybrids, as well as some examples from the trade. Identification of both parental species was even possible when unambiguous reference data were available. Unfortunately, other hybrid fishes purchased from aquarium stores were unable to be identified as such, indicating a universal and simple method to detect fish hybrids through nDNA sequencing requires further work.

Acknowledgements

This work represents part of an unpublished chapter of my PhD thesis, and has not undergone formal peer review. Therefore, I take sole responsibility for its content. I thank: my supervisors Karen Armstrong (Bio-Protection Research Centre), Rob Cruickshank (Lincoln University), and Suzanne Keeling (MPI); Patrick Cambell (NHM, London) for providing tissue samples of a hybrid *Clarias* catfish; Richard Broadbent (Warwick, UK) for help with sourcing some of the hybrids from the aquarium trade; Jeremy Wright (University of Michigan) for discussion about *Synodontis*; and Robert Ward (CSIRO) and Ian Hogg (University of Waikato) for comments on the thesis version of this manuscript. The work was funded by a Ministry of Primary Industries Biosecurity New Zealand scholarship.

References

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

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.

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

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.

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

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.

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.

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

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.

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.

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.

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

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.

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.

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

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.

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.

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.

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

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.

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.

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.

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

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.

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

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.

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.

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.

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

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

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

Rozen, S. & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, 132:365–386.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Zink, R. M. & Barrowclough, G. F. (2008). Mitochondrial DNA under siege in avian phylogeography. *Molecular Ecology*, 17:2107–2121.