Finding stories in noise: Mitochondrial portraits from RAD data

Authors: C.S. STOBIE, M.J. CUNNINGHAM, C.J. OOSTHUIZEN AND P. BLOOMER

Molecular Ecology and Evolution Programme, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Private Bag X20, Hatfield, Pretoria, 0028, South Africa

Keywords: genotyping-by-sequencing, hybridisation, mitochondrial genome, polyploidy, population genomics

Corresponding author: C.S. Stobie

Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Private Bag X20, Hatfield, Pretoria, 0028, South Africa

Fax: +27123625327

E-mail addresses: connor.stobie@gmail.com

Running Title: Mining mitogenomes from RAD sequencing

Abstract

Mitochondrial DNA (mtDNA) has formed the backbone of phylogeographic research for many years, however, recent trends focus on genome-wide analyses. One method proposed for calibrating inferences from noisy Next-Generation data, such as RAD sequencing, is to compare these results with analyses of mitochondrial sequences. Most researchers using this approach appear to be unaware that many Single Nucleotide Polymorphisms (SNPs) identified from genome-wide sequence data are themselves mitochondrial, or assume that these are too few to bias analyses. Here we demonstrate two methods for mining mitochondrial markers using RAD sequence data from three South African species of yellowfish, *Labeobarbus*. First, we use a rigorous SNP discovery pipeline using the program STACKS, to identify variant sites in mtDNA, which we then combine into haplotypes. Secondly, we directly map sequence reads against a mitochondrial genome reference. This method allowed us to reconstruct up to 98% of the *Labeobarbus* mitogenome. We validated these mitogenome reconstructions through BLAST database searches and by comparisons with cytochrome *b* gene sequences obtained through Sanger sequencing. Finally, we investigate the organismal consequences of these data including ancient genetic exchange and a recent translocation among populations of *L. natalensis*, as well as interspecific hybridisation between *L. aeneus* and *L. kimberleyensis*.

1. Introduction

Mitochondrial DNA (mtDNA) sequencing has long been the mainstay of phylogeographic analysis in vertebrates due to the combination of high mutation rates, the absence of recombination and effectively haploid, uniparental inheritance (Avise, 1986; Avise *et al*., 1987; Bermingham & Avise, 1986; Brown, George, & Wilson, 1979; Wilson *et al.*, 1985). This fortuitous confluence of factors leads to rapid lineage sorting within populations and divergence among locations, thus making mtDNA sequences uniquely informative on geographic structure and population history (Avise, 1986; Avise *et al.*, 1987; Avise, Bermingham, Kessler, & Saunders, 1984; Bermingham & Avise, 1986). However, mtDNA polymorphisms also contribute to variance in fundamental physiological and developmental components of fitness such as metabolic performance and age related disease (Avise *et al*., 1987; Ballard & Whitlock, 2004; Wallace, 2008). As a single recombinational locus it is difficult to separate effects of population history and geography from that of selection on standing variation in mtDNA (Ballard & Whitlock, 2004).

More recently mtDNA has experienced resurging popularity as a well understood landmark in phylogeographic analyses, for comparison with genomic data from non-model species (e.g. Jeffries *et al*., 2016; Macher *et al*., 2015; Moura *et al*., 2014; Puckett, Etter, Johnson, & Eggert, 2015; Streicher *et al*., 2014). In contrast to mtDNA, mutations are less frequent in the nuclear genome and sorting of polymorphisms is slower (Brown, George, & Wilson, 1979; Funk & Omland, 2003; Wilson *et al.*, 1985). Consequently, most short genomic sequences show low variation and are relatively uninformative on population history and structure (Edwards & Bensch, 2009; Toews & Brelsford, 2012). Although the use of large numbers of Single Nucleotide Polymorphisms (SNPs) effectively counters this, phylogeography generally uses both mitochondrial and nuclear markers (Avise, 2009; Edwards & Bensch, 2009). There has been a recent trend advocating the multilocus approach of using as many independent markers as possible, including both nuclear and extranuclear loci

(Bermingham & Moritz, 1998; Brito & Edwards, 2009; Edwards & Bensch, 2009). Although the use of genome-wide SNPs may appear to make mtDNA analysis irrelevant, the additional (maternal) perspective to the same complex evolutionary history can add levels of insight to an analysis, particularly in the event of cyto-nuclear discordance (Ballard & Whitlock, 2004; Brito & Edwards, 2009; Toews & Brelsford, 2012; Zhang & Hewitt, 2003). Analysis of mtDNA between hybrids also allows inference of directionality of hybridisation (Avise, 1986). Additionally, bioinformatic filtering of Next Generation Sequencing (NGS) genomic datasets is complex and analyses of mitochondrial DNA contribute to validation of genome-wide genetic signal.

Labeobarbus is a genus of hexaploid (± 150 chromosomes) (Oellermann & Skelton, 1990) freshwater fish distributed throughout Africa. Its position in the family Cyprinidae has been a matter of debate, although a recent phylogeny has placed the genus within the tribe Torini (Yang *et al*., 2015). The hexaploid Torini are thought to have arisen in the late Miocene through hybridisation between a male tetraploid Torin and a female from the diploid genus *Cyprinion*, tribe Barbini (Tsigenopoulos, Kasapidis, & Berrebi, 2010; Yang *et al*., 2015). Spontaneous chromosome duplication occurred within the gametes of this inter-tribal hybrid (Oellermann & Skelton, 1990). The *Labeobarbus* genome retains high levels of paralogy from this event, which must be taken into account during genetic analyses. Some authors (*cf*. Levin, 1983; Van de Peer, Mizrachi, & Marchal, 2017; Zhan, Glick, Tsigenopoulos, Otto, & Mayrose, 2014) have hypothesised that increased ploidy levels may be a driver of adaptation to new environments due to the "extra degree of genetic freedom" experienced by polyploids.

Currently seven species of *Labeobarbus* are recognized from South Africa, with *Varicorhinus* considered a synonym (Tsigenopoulos et al., 2010; Vreven, Musschoot, Snoeks & Schliewen, 2016; Yang *et al*., 2015). Members of this genus are known locally as yellowfish and are prized for game fishing, and serve as 'flagship' species indicative of river health (Skelton & Bills, 2008). Two of the most abundant species in the region are the KwaZulu-Natal yellowfish *L. natalensis* and the Orange-

Vaal smallmouth yellowfish *L. aeneus*. The KwaZulu-Natal yellowfish are endemic to KwaZulu-Natal Province, whereas Orange-Vaal smallscale yellowfish are distributed widely across South Africa, primarily within the Orange-Vaal river system. Mitochondrial phylogeographic studies have shown that KwaZulu-Natal yellowfish consists of a number of lineages associated with separate drainage systems (Bloomer *et al*., 2007). Genome-wide SNP data are congruent with this, and separated the species into five major lineages: the Umfolozi, Tugela, Mbokodweni, Mkomaas and Umgeni populations (Stobie, Oosthuizen, Cunningham, & Bloomer, 2018). Orange-Vaal smallscale yellowfish are thought to hybridise with Orange-Vaal largemouth yellowfish *L. kimberleyensis* (Eccles, 1986; Gaigher, 1976; Mulder, van Vuuren, Ferreira, & van der Bank, 1990; van Vuuren, Mulder, Ferreria, & van der Bank, 1989). However, morphometrics, allozyme and mitochondrial DNA studies of introgression between these forms have been inconclusive (Bloomer *et al*., 2007). Our ongoing research, using genomic data, is aimed at resolving this.

To investigate the phylogeography of these hexaploid species we analysed genome-wide SNPs complementing our previous analyses of mitochondrial sequences. Restriction-site Associated DNA (RAD) sequencing (Baird *et al*., 2008; Miller, Dunham, Amores, Cresko, & Johnson, 2007) has become a popular technique to obtain molecular markers distributed across an organism's genome. Interest in the method has boomed exponentially since its inception, and it has been used in a wide array of applications (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Davey *et al*., 2013; Davey *et al*., 2011; Stobie *et al*., 2018).

Double-digest RAD (ddRAD) is a variant of this technique that removes the need for random fragmentation by sonication. Instead ddRAD fragments the genome by digestion with two restriction enzymes, one with a commonly encountered recognition sequence and the other targeting a less frequent motif (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). Fragments within a given size range with different recognition sequences at either end are then sequenced (Peterson *et al*., 2012). This method offers several advantages over traditional RAD sequencing: that it allows greater specificity

and repeatability in fragments sequenced, reduction in library preparation costs, it maximises coverage of both fragment ends, and it requires less genomic DNA (Peterson *et al*., 2012). However, this method also has several pitfalls, such as allele frequency estimation bias from restriction site mutations (Arnold, Corbett-Detig, Hartl, & Bomblies, 2013; Eaton, Spriggs, Park, & Donoghue, 2017), indels eliminating alleles from size-selection (DaCosta & Sorenson, 2014), and overrepresentation of particular genomic regions due to enzyme choice (DaCosta & Sorenson, 2014).

A popular way to verify inferences from RAD data is to sequence other markers for independent analysis, such as one or more mitochondrial DNA fragments (Herrera, Watanabe, & Shank, 2015; Jezkova *et al*., 2015; Macher *et al*., 2015; Pante *et al*., 2015; Streicher *et al*., 2014). This allows one to show whether observed trends are corroborated by other markers, although in some cases a different pattern will emerge from mitochondrial sequences (e.g. Cruaud *et al*., 2014; Jezkova *et al*., 2015; Macher *et al*., 2015; Puckett *et al*., 2015). Mito-nuclear discordance here may be indicative of a problem during the RAD data analysis or an evolutionary process responsible for the incongruity such as introgression (reviewed in Ballard & Whitlock, 2004). To our knowledge, few studies have attempted to mine the mitochondrial data present in RAD datasets for independent analysis (but see Marrano *et al*., 2017; Perry, Pederson, & Baxter, 2017; Pujolar *et al*., 2014; Terraneo, Arrigoni, Benzoni, Forsman, & Berumen, 2018; Truong *et al*., 2012). Here we compare results from our previous genome-wide SNP study of KwaZulu-Natal yellowfish (Stobie *et al*., 2018) with analyses of mitochondrial SNP markers mined from the ddRAD sequence data. This is supplemented with additional nuclear analyses for all species. We construct partial mitogenomes for each sample to investigate the evolutionary relationships within and between three *Labeobarbus* species. Finally, we validate these mitochondrial mappings by comparing them to Sanger sequences of a mitochondrial gene.

2. Materials and Methods

2.1. Sampling and DNA extraction

We used the 23 samples previously described in Stobie *et al*. (2018) for the KwaZulu-Natal yellowfish (*L. natalensis*), and an additional 39 samples from species in the Orange-Vaal system (24 *L. aeneus* and 15 *L. kimberleyensis*, hereafter referred to jointly as "Orange-Vaal yellowfish"). The Orange-Vaal samples were collected between January 2004 and October 2006 from ten localities in South Africa (Figure 1; Supporting Information Table S1). These localities represented areas of interest in the upper and lower Orange River, as described previously in Bloomer *et al*. (2007). Fin and muscle samples were stored in 96% ethanol at 4°C. DNA was extracted using the DNeasy Blood and Tissue Extraction Kit (Qiagen). DNeasy extractions were performed following the manufacturer's protocol.

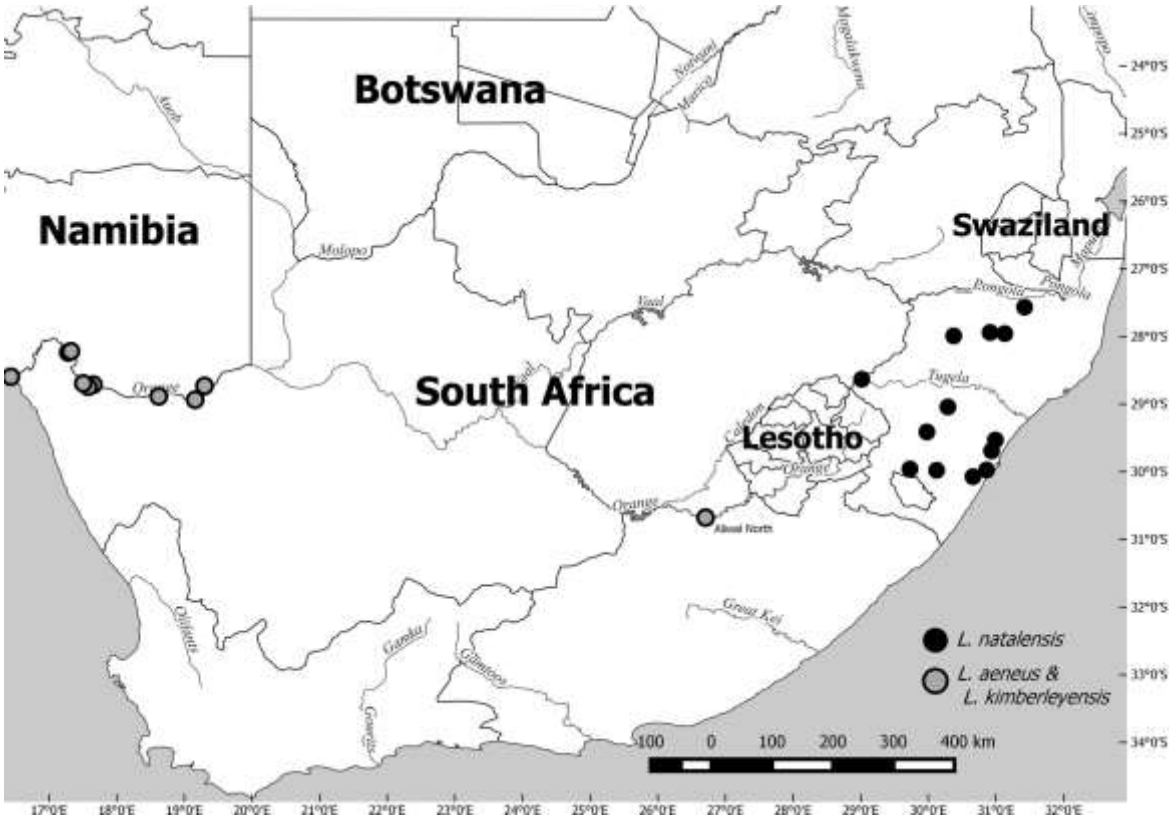

Figure 1. Sampling locations across South Africa showing major river systems. Sampling localities in black correspond to KwaZulu-Natal yellowfish, localities in grey are Orange-Vaal yellowfish. This map was produced using qgis (QGIS Development Team, 2016. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://www.qgis.org/)

A GeneQuant™ *pro* RNA/DNA calculator spectrophotometer (Amersham Biosciences, Freiberg, Germany) and agarose gel electrophoresis were used to assess DNA concentration and quality. Samples were sent to Beijing Genomics Institute Hong Kong Co., Limited (BGI, Hong Kong) for screening by agarose gel electrophoresis and Qubit® 2.0 Fluorometer (Invitrogen) evaluation followed by ddRAD sequencing. Orange-Vaal yellowfish were sequenced in two independent batches, with a separate batch for the KwaZulu-Natal yellowfish (Supporting Information Table S1).

2.2. Library preparation and sequencing

High-quality DNA samples were sent to BGI for library preparation and sequencing. Samples were digested with the restriction enzymes *Mlu*CI and *Nla*III following the double-digest paired-end protocol of Peterson *et al*. (2012). These enzymes were chosen as they were found to cut frequently throughout the genome of a wide range of species (Peterson *et al*., 2012). Each individual was tagged with a unique 4-8 base pair barcode, with five Orange-Vaal samples replicated as controls. These libraries were filtered for fragments between 300-500 bp and combined across individuals. Sequencing (90 bp paired-end) was done in a single lane of an Illumina HiSeq 2000 (Illumina Inc., USA) for the first Orange-Vaal batch, and in a single lane of an Illumina HiSeq 4000 (Illumina Inc., USA) for the other. The resulting reads were screened for poor quality (reads with more than 50% low quality bases i.e. quality value ≤ 5 (*E*)) and demultiplexed at BGI before being returned to us for further analyses. The reads were then trimmed to a consistent length of 80 bp after removing adapter and barcode sequences.

2.3. Bioinformatic discovery of mitochondrial SNPs

We apply two approaches to mining mitochondrial markers (pipeline summarized in Supporting Information File S1). Our first approach initially uses a SNP identification pathway to detect candidate

loci which are then screened using BLAST against a reference genome. CLC GENOMICS WORKBENCH 7.0.4 (CLC Inc., Aarhus, Denmark), FASTQC (Andrews, 2010) and command line searches were used to assess quality, GC-content, and levels of error in the libraries of raw RAD data files. Paired reads were physically joined for SNP identification using a custom script (AST Papadopulos, personal communication), resulting in a single file per individual containing reads of 160 bp. Adapter pollution was removed using command line searches (GREP) with regular expressions (Stobie *et al.* 2018). Identification of homologous sequences and SNP identification were conducted using STACKS 1.44 (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011; Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013). The program PROCESS_RADTAGS was run with the additional flags *-r* (rescue RAD tags), *-c* (clean data) and *-q* (remove low-quality reads). Paired reads were retained for downstream processing. Following the STACKS parameter-testing method of Paris, Stevens, and Catchen (2017) we identified an optimal parameter set of (*-m* 5 *-M* 2 *-n* 2) for SNP identification in paired reads from KwaZulu-Natal yellowfish (Supporting Information Figures S1-S3, Table S3). For conformity, we adopted the same parameters for Orange-Vaal yellowfish.

Once stacks had been built and SNPs identified, reads that had passed through this pipeline were then run through CLC MAIN WORKBENCH 6.9 (CLC Inc., Aarhus, Denmark) local BLASTN (Altschul *et al*. 1997) against the mitogenomes of *Hypselobarbus* (*H. jerdoni* – NC_031587.1), *Labeobarbus* (*L. intermedius* – NC_031531.1; *L.* sp. Kongou – AP011324.1; *L.* sp. Lucien – AP011323.1), *Neolissochilus* (*N. hexagonalepis* – NC_026106.1 and KU553349.1; *N. soroides* – AP011314.1; *N. stracheyi* – NC_031555.1), *Tor* (*T. khudree* – NC_027617.1; *T. mosal mahanadicus* – KU870466.1; *T. putitora* – NC_021755.1 and AP011326.1; *T. sinensis* – NC_022702.1; *T. tambroides* – JX444718.1 and AP011372.1; *T. tor* – KR868704.1), and *Varicorhinus* (*V. maroccanus* – NC_031528.1). These taxa were chosen as they are most closely related to *Labeobarbus* and had at least one mitochondrial genome available at the time of writing. The search allowed low complexity regions to be filtered and used default parameters (word size = 11, match = 2, mismatch = -3, gap existence = 5, gap extension

= 2). Any hits identified by BLAST were screened for significance at a critical *E*-value of 1*E*-20 and for a hit length exceeding 60 base pairs. These reads were verified by using BLASTN through the NCBI web page (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to match against the entire GenBank nucleotide database.

These candidate mitochondrial loci were included on a "whitelist", which excludes all other loci from subsequent analysis, and run again through the POPULATIONS module of STACKS. The results were filtered to exclude loci containing at least one SNP in a heterozygous state as these would be indicative of either sequencing error, merging of loci, contamination or a nuclear encoded sequence of mitochondrial origin (NUMTs). These loci were similarly filtered to exclude duplicated SNPs arising from overlap between the forward and reverse reads of a pair. Output from this whitelisted dataset was exported in FASTA and GENEPOP formats. These files were modified to reflect the haploid nature of these markers. Individual haplotypes were generated by combining SNP alleles across all identified mitochondrial loci, these haplotypes were converted to NEXUS format using PGD SPIDER 2.0.7.2 (Lischer & Excoffier, 2012) and imported into TCS 1.21 (Clement, Posada, & Crandall, 2000) to produce statistical parsimony haplotype networks with 95% connection limits. Collection limit constraints were further relaxed to join all haplogroups. We required samples to possess at least 20 SNPs to be retained for the TCS network construction, as excessive amounts of missing data result in pairwise deletion of informative markers.

The program BEAUTI 1.7.4 (Drummond, Suchard, Xie, & Rambaut, 2012) was used to generate XML files for use in BEAST 1.7.4 (Drummond *et al*., 2012) from the NEXUS files for both KwaZulu-Natal yellowfish and Orange-Vaal yellowfish. Default parameters were used. We chose a substitution model of HKY+G (Hasegawa, Kishino, & Yano, 1985) with empirical base frequencies to match our mitogenome mapping approach below. A Yule Process tree prior was used. The length of the chain was set to 10 x 10^6 steps with sampling every 1000 steps. Three replicates were obtained for this analysis then independently viewed in FIGTREE 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/) and

combined using LOGCOMBINER 1.7.4 (http://beast.community/logcombiner). The tree produced from this analysis was annotated using TREEANNOTATOR 1.7.4 (http://beast.bio.ed.ac.uk/TreeAnnotator) and viewed in FIGTREE.

The abundance of organelle DNA is anticipated to result in much greater mitochondrial read depth relative to nuclear loci (Ekblom, Smeds, & Ellegren, 2014). The identification of mitochondrial loci allowed us to test the assumption that mitochondrial loci can be filtered from a RAD dataset by applying an upper read depth threshold. We calculated the number of mitochondrial RAD tags captured and the proportion of mitochondrial loci in a dataset of KwaZulu-Natal yellowfish (Read 1 fragments, -*M* 3, -*n* 2, -*r* 0) as the minimum depth of coverage (-*m*) was increased from 3 to 100.

2.4. Reference-based mitogenome assembly

Our second approach to obtaining mitochondrial data was to directly map sequence reads to a reference mitogenome. Raw reads were filtered for quality, GC-content, adapter pollution, levels of error and degree of overlapping reads as per the previous approach above. Prior to merging of paired reads the data was imported to CLC GENOMICS WORKBENCH where overlapping paired reads were merged. Individual read files (both paired and merged together) were mapped in CLC GENOMICS WORKBENCH against the mitogenome of the East African *L. intermedius*, which is the most closely related reference. We included seven replicated samples (the same sample sequenced twice in separate runs with different barcodes) to assess consistency of results. We performed an *in silico* digestion of the *L. intermedius* reference mitogenome to estimate the expected proportion of the *Labeobarbus* mitochondrial genome covered by our sequence data (Supporting Information Table S2). In this analysis we considered both complete and uniformly incomplete restriction digestion. Complete digestion includes fragments in the 300 – 500 bp size range excluding those with internal restriction sites. Uniformly incomplete digestion includes all fragments in this range irrespective of internal restriction sites. Mappings used default parameters except for fraction length and similarity

which were both set to 0.9. Non-specific mappings were ignored. Consensus mappings required a minimum depth of 3 reads and gaps were filled with Ns. Conflicts were resolved by voting for the most common nucleotide. Consensus sequences were exported in FASTA format.

Consensus sequences were aligned using MAFFT 7.294 (Katoh & Standley, 2013) and imported into MEGA 6.06 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013) for manual checking of the alignment. The alignment was exported in NEXUS format. Statistical parsimony haplotype networks were generated in TCS as above. Model checking was performed in MEGA to identify the optimal nucleotide substitution model for the dataset. The alignment was then exported into BEAUTI to produce the XML file for use in BEAST. The optimal available site heterogeneity model was HKY+G, which we used with empirical base frequencies. We independently ran a functionally-partitioned model-based method for each of the mitochondrial genes and found it gave similar results to the single-model unpartitioned dataset which we present here. We used a strict clock with a cyprinid rate for cytochrome *b* (cyt *b*) of 0.53% per lineage per million years (Dowling, Tibbets, Minckley, & Smith, 2002) which has previously been used for *Labeobarbus* (Tsigenopoulos *et al*., 2010) and other cyprinids (Jeffries *et al*., 2016). We specified a Yule Process tree prior with a chain of 50 x 10^6 MCMC steps, sampling every 5,000 steps. All other parameters were left at their default values. Three replicates were again repeated for this analysis, combined in LOGCOMBINER, annotated in TREEANNOTATOR and viewed in FIGTREE. We produced a phylogram in MRBAYES (Ronquist & Huelsenbeck, 2003) from the same input file but removing the outgroup, with the same substitution model, and 30 x 10^6 MCMC steps. All other parameters were retained at default values. Output was viewed in FIGTREE.

2.5. Validation with Sanger sequencing

Consensus sequences produced from the mitogenome mapping approach were compared against Sanger sequences for the cyt *b* gene. DNA was extracted for Sanger sequencing following the Chelex

protocol (Estoup, Largiader, Perrot, & Chourrout, 1996) for all samples of KwaZulu-Natal yellowfish. We sequenced cyt *b* using custom primers developed for this species (cytb-H – 5'-AGG GCA GGC TAA TTC TAG TG-3' and cytb-L – 5'-GAA CCT TAA TGG CAA GCC TAC G-3'). Reactions contained 1 x PCR buffer, 0.5 U of SuperTherm *Taq* polymerase (Southern Cross Biotechnologies), 1.25 mM $MgCl_2$, 0.1 pmol forward and reverse primers (Whitehead Scientific), 0.06 mM dNTPs (Promega) and 50-100 ng of DNA. PCR conditions were 94°C for 2 minutes, followed by 35 cycles of 94°C for 30 seconds, 55°C for 30 seconds and 72°C for 30 seconds, ending in a final elongation step of 72°C for 15 minutes. Amplicons were precipitated with 4.55 volumes of absolute ethanol, 0.45 volumes of Sabax® water (Adcock-Ingram) and 0.1 volumes of 3 M NaAc. Precipitated products were sequenced in both directions using the PCR primers.

Cycle sequencing was performed using the ABI PRISM BigDye Terminal Cycle Sequencing Ready Reaction Kit version 3.1 (Applied Biosystems) following the manufacturer's protocols. Approximately 50-100 ng of DNA was used with 1 µl of BigDye reaction mix, 1x BigDye Sequencing Buffer and 0.32 pmol of forward or reverse primer. Sequences were produced on an ABI 3130 automated sequencer (Applied Biosystems) and visualised as well as assembled in CLC MAIN WORKBENCH.

Mitogenome mappings were trimmed to the corresponding area of cyt *b* and then aligned with the Sanger sequences in MEGA. Pairwise distances were computed between the ddRAD and Sanger sequence for each sample, including replicates. Sequencing error was calculated as a proportion of nucleotide sites aligned across both approaches.

2.6. Nuclear SNP comparison

We mined nuclear SNP datasets from the RAD data to provide a comparison for our mitochondrial results. This was done using the STACKS pipeline similarly to our initial approach above, but using only Read 1 fragments. We selected the parameters *-m* 3, *-M* 3, *-n* 2 for the KwaZulu-Natal yellowfish and *-m* 3, *-M* 2, *-n* 1 for the Orange-Vaal yellowfish. The parameter *--max_locus_stacks* was set to 7 to

take into account hexaploidy of these species (Stobie *et al*. 2018). The POPULATIONS module was executed without specifying a predefined population, *-r* was set to 0.8, with a minor allele frequency filter to remove singletons (*--min_maf* = 0.03 for KwaZulu-Natal yellowfish and 0.015 for Orange-Vaal yellowfish), and a blacklist of mitochondrial loci found in the same manner as above using CLC MAIN WORKBENCH and BLASTN was specified. This resulted in 665 SNPs for KwaZulu-Natal yellowfish and 984 for Orange-Vaal yellowfish. Output files were edited to produce XML files for BEAST as described above. The same models and parameters were used to assist comparability between results. As for the mitochondrial analysis, three replicate analyses were combined in LOGCOMBINER, and then viewed in FIGTREE. A comparison to determine the effect of retaining mitochondrial SNPs was also performed in the same way, but excluding the blacklist used. This approach used 679 and 988 SNPs respectively.

3. Results

3.1. Raw sequence data

Illumina 90PE sequencing of the first Orange-Vaal batch resulted in around 83.5 million reads over 11 individuals, with an average of 7,587,820 raw reads per individual (Table 1). In contrast, the second Orange-Vaal batch consisted of almost 336 million reads across 28 individuals (four of which were replicated). The average number of reads per sample in this batch was 10,494,963 reads. Filtering of each batch for adapter pollution and quality resulted in a significant drop in read count (Table 1). Using command line searches, we confirmed that all read pairs started with CATG (*Nla*III) and ended with AATT (*Mlu*CI), which is a requirement for STACKS. The Q20 percentage and GC-content of the Orange-Vaal libraries were slightly higher to that observed for KwaZulu-Natal yellowfish (Table 1). In addition, the second Orange-Vaal batch also has higher GC-content and better base quality scores than either of the other two libraries. Further investigation of this difference showed that the prevalence of repetitive elements may partially account for this difference.

Table 1 Summary statistics from initial analysis of RAD sequencing data

Library	Raw reads	Filtered reads	Q20 (%)	GC (%)
KwaZulu-Natal	152,941,120	113,883,800	97.10 – 97.95	38.5 – 40.8
Orange-Vaal 1	83,466,022	79,133,404	97.76 – 98.15	39 – 40
Orange-Vaal 2	335,838,822	335,325,196	98.40 – 98.85	42 – 44
Total	572,245,964	528,342,400		

3.2. STACKS*-based mitochondrial SNP discovery*

We found 102 putative mitochondrial SNPs in the KwaZulu-Natal yellowfish dataset (1.70%) and 98 putative mitochondrial SNPs in the Orange-Vaal yellowfish dataset (0.08%). The mitochondrial location of these SNPs was supported by BLASTN searches. A number of these SNP loci (12 and 13, respectively from each dataset) included heterozygous individuals and were removed. This is potentially indicative of either sequencing error, erroneous merging of paralogous loci, contamination or the presence of NUMTs. Duplicated SNPs originating from overlapping sequences were also identified and removed (14 and 19, respectively). After filtering we retained 77 mitochondrial SNPs, out of a total of 5,991 SNPs (1.29%) in KwaZulu-Natal yellowfish, and 66 mitochondrial SNPs out of 117,183 in total (0.06%) from the Orange-Vaal yellowfishes.

Haplotype networks obtained using mitochondrial SNPs from the STACKS pipeline (Figures 2-3) show a general division into species and catchment-associated populations. We recovered the previously identified populations of KwaZulu-Natal yellowfish (Stobie *et al*., 2018), as well as a novel split within the southernmost population, with separate haplogroups in the Mkomaas and Umzimkhulu River catchments (Figure 2). The Mkomaas haplogroup was the most divergent from other lineages, while the Umzimkhulu lineage had the highest divergence among haplotypes within a haplogroup. Haplotypes from the Orange-Vaal yellowfishes split into two major haplogroups, roughly

corresponding to each species (Figure 3), but with some 'misplaced' individuals, where assignment to an mtDNA lineage conflicted with phenotypic identification of the fish.

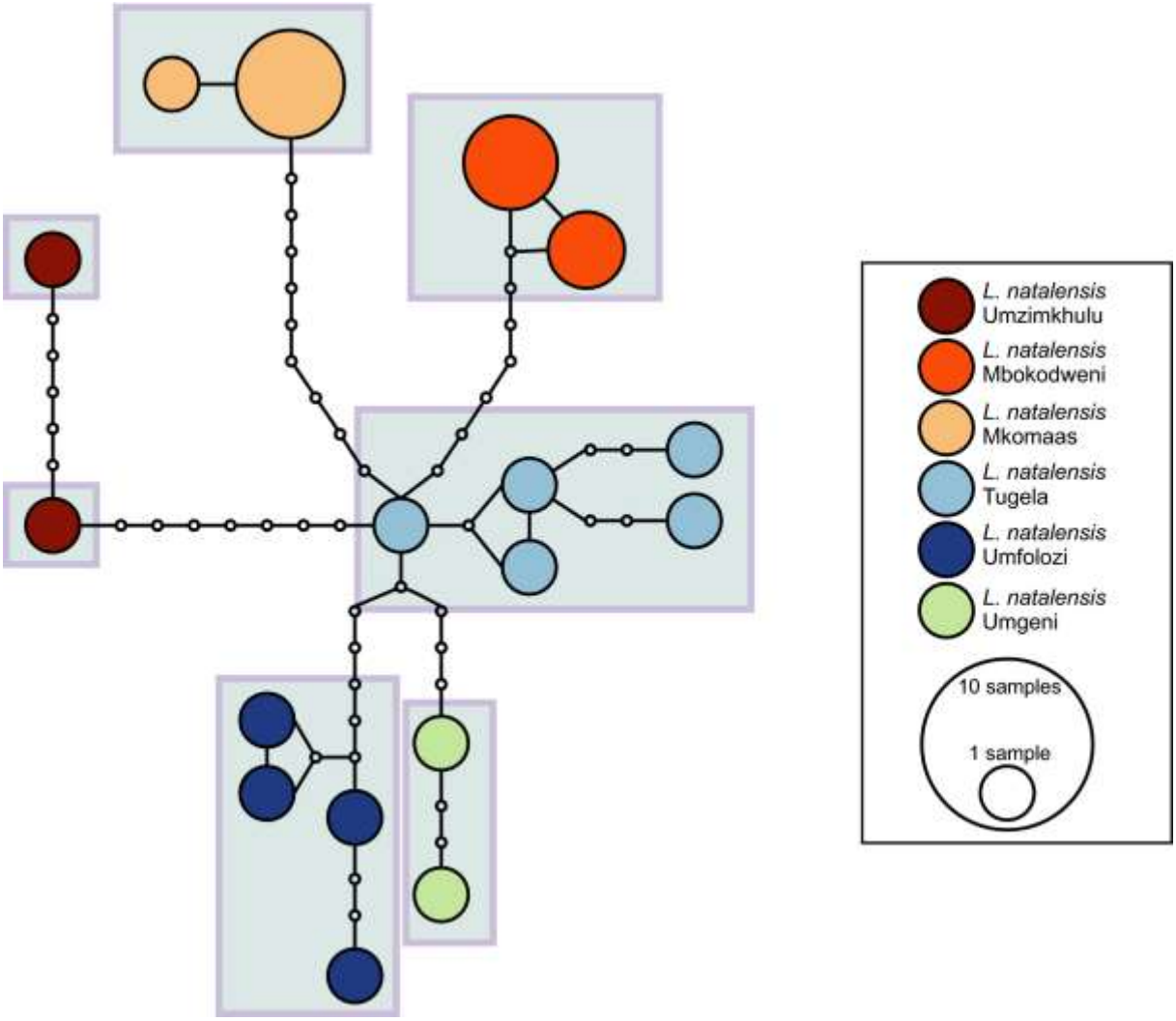

Figure 2: Haplotype network produced from 77 polymorphic mitochondrial SNPs identified using the STACKS pipeline in KwaZulu-Natal yellowfish. Boxes indicate 95% connection limits (three steps).

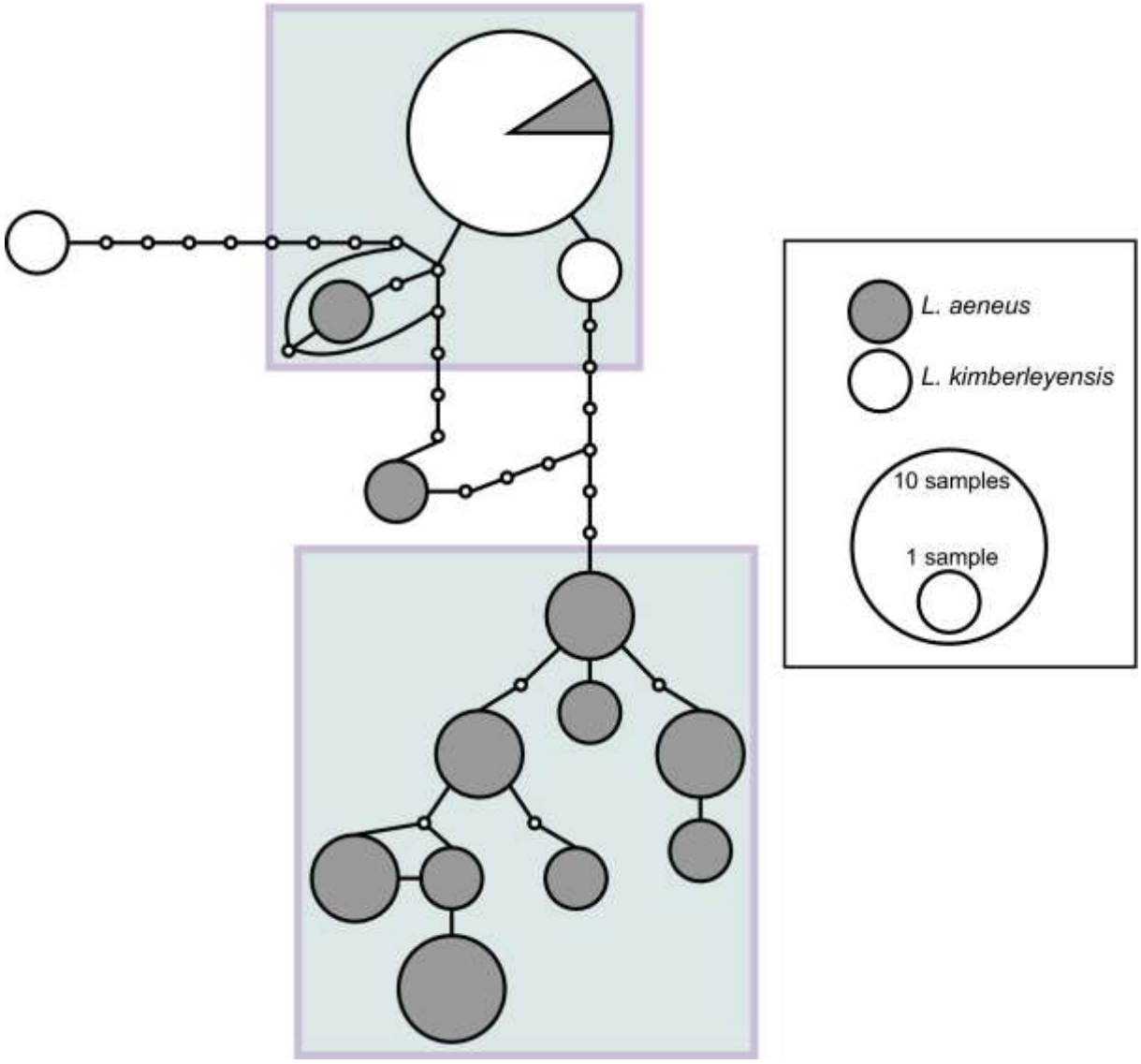

Figure 3: Haplotype network from 66 polymorphic mitochondrial SNPs identified using the STACKS pipeline in Orange-Vaal yellowfish. Boxes indicate 95% connection limits (three steps). A maximum of ten connecting steps were required to join all haplotypes.

The Yule Process phylogenetic trees (Supporting Information Figure S4) produced based on the STACKS-based approach show contrasting results between the KwaZulu-Natal yellowfish and the Orange-Vaal yellowfish. The KwaZulu-Natal yellowfish tree shows structuring by populations broadly matching those identified using nuclear SNPs (Stobie *et al*. 2018). The one exception here is the separation of the Umzimkhulu lineage as seen in the haplotype network. The Orange-Vaal yellowfish showed no substructure below the species level, and three samples which were morphologically

identified as Orange-Vaal smallmouth yellowfish clustered with high support in the Orange-Vaal largemouth side of the tree. Some nodes received low support, likely due to the relatively few markers used in this approach. However, most major divergences were well-supported.

The number of mitochondrial RAD tags retained in datasets declined as the depth of coverage requirement is increased (Supporting Information Figure S5A). This appears to be a fairly uniform loss (although the greatest difference in this dataset is between *-m* = 3 and *-m* = 5) until only 20 of the 72 mitochondrial tags are retained at *-m* = 100. The proportion of mitochondrial tags increases over this scale (Supporting Information Figure S5B) from 0.009% at *-m* = 3 to 5.95% high at *-m* = 90 before declining slightly again at *-m* = 100.

3.3. Reference-based mitogenome assembly

In total, 467,848 reads were mapped to the *L. intermedius* mitogenome across all individuals (~0.1% of all reads). *In silico* restriction digestion of the *L. intermedius* mitogenome gave an expected 80% coverage, if uniformly incomplete restriction digestion is prevalent in the data. This seems a fair estimate as the average mitogenome coverage per individual was found to be 66% (range of 44-83%) indicating a high level of partial digestion. This is supported by the presence of internal restriction sites within 9.54% mapped reads. Chimaeric coverage, combining mapped reads across all individuals, yielded 98% of the mitogenome. Imposing a minimum depth of three reads at a site reduced the average coverage to 54% (range = 15-76%) and yielded 536 SNPs for downstream analyses. Estimates incorporating the possibility of complete digestion yielded an expected coverage of only 22%.

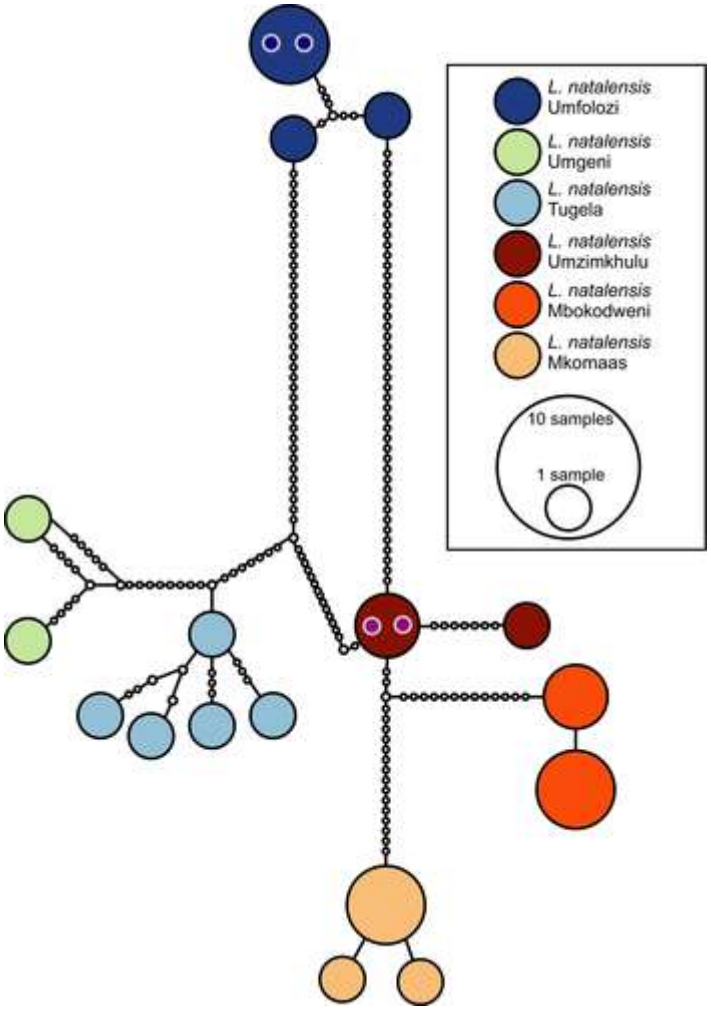

Figure 4: Haplotype network produced from consensus mitochondrial genome sequences found by mapping read data of 25 KwaZulu-Natal yellowfish samples against the *L. intermedius* reference mitogenome. Two replicate pairs are included and indicated by grey-bordered filled dots with shades of purple indicating pairs.

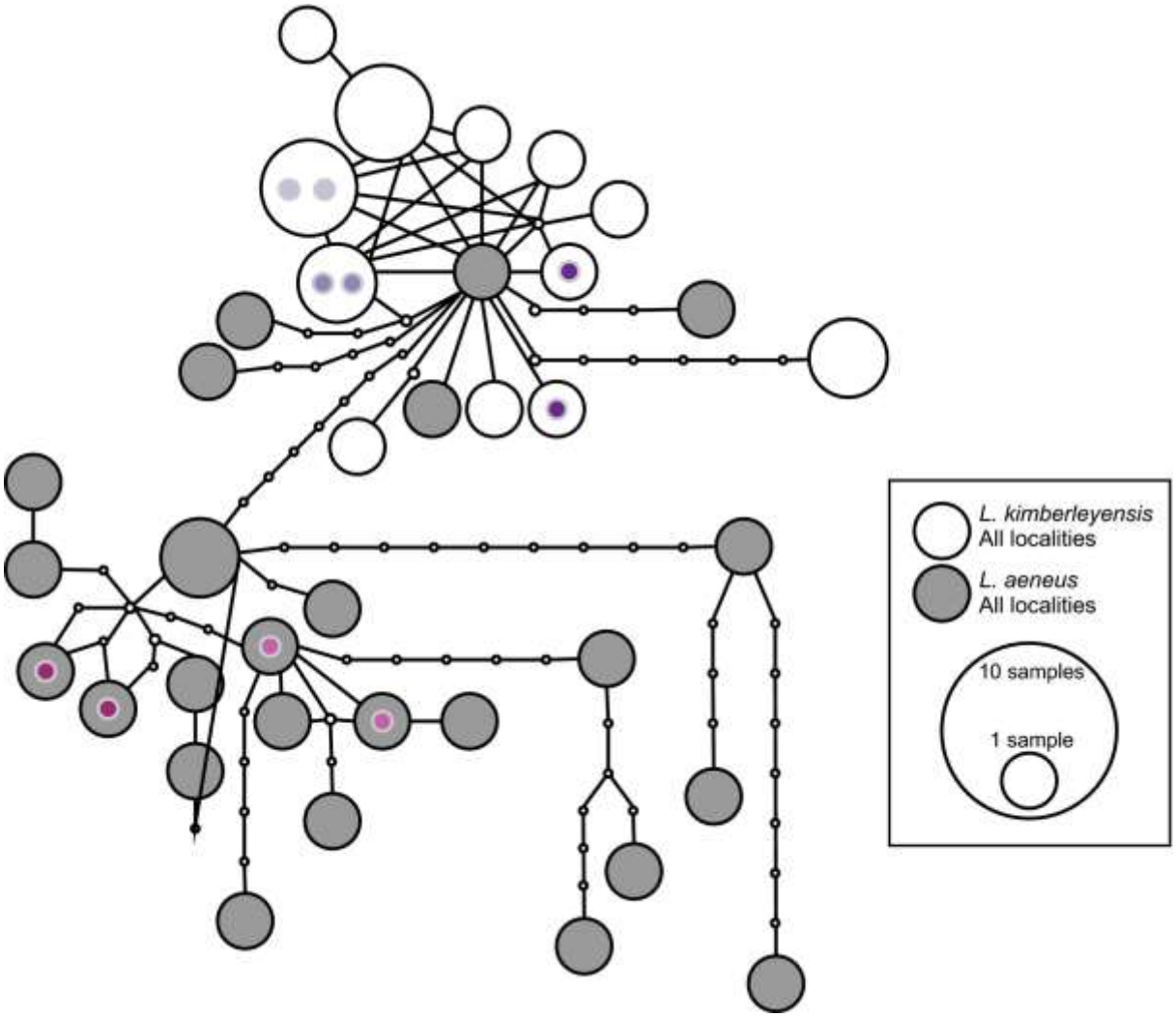

Figure 5: Haplotype network for the 44 Orange-Vaal yellowfish produced using consensus mitochondrial genome sequences found by mapping read data against *L. intermedius*. Five replicate pairs are included and indicated by grey-bordered dots with shades of purple indicating the pairs.

Haplotype networks from mitogenome mapping of KwaZulu-Natal and Orange-Vaal yellowfish (Figures 4-5) were consistent with previous studies (Bloomer *et al*., 2007; Stobie *et al*., 2018) and with the STACKS-based approach of haplotype inference from filtered mitochondrial SNPs. However, the mapping approach yielded finer resolution of haplotype differences and greater divergence among populations, as more data is generated by mapping reads directly to a reference genome. This also suggests that the mapping approach retained many SNPs that were excluded by our STACKS SNP discovery pipeline - as shown by the difference in total mitochondrial SNPs found between approaches. For KwaZulu-Natal yellowfish, haplogroups were separated by a large number of

mutations with relatively few mutations separating haplotypes within populations (Figure 4). Both replicates from this dataset yielded identical sequences. The situation for the Orange-Vaal yellowfishes is more complex, with lower diversity overall, forming two to three clusters, broadly split into the two species (Figure 5). However, five individuals morphologically assigned to Orange-Vaal smallmouth yellowfish carried haplotypes from the Orange-Vaal largemouth yellowfish haplogroup. Three of these samples matched those from the STACKS-based SNP approach - the remaining two dropped out of this analysis due to excess levels of missing data. Two of the five replicated samples from this dataset yielded identical sequences, with haplotypes from the other three replicates differing by one to two nucleotide differences. Minor differences among haplotypes from the same biological sample were expected, as the sequences used to generate these results are not identical, due to variation in coverage and depth among replicates.

The Bayesian Yule Process chronogram (Figure 6), produced from the mapped mitogenome sequences, splits these South African yellowfish into KwaZulu-Natal yellowfish and Orange-Vaal lineages first, with further subdivision into the six populations of KwaZulu-Natal yellowfish and the two Orange-Vaal species, Orange-Vaal smallmouth and largemouth yellowfish. We did not observe any further clustering by locality for the Orange-Vaal yellowfishes, which was surprising given the geographic distance between the upper Orange sampling site near Aliwal North and the remaining sampling sites in the lower Orange river, more than 1000 km downstream. The lower Orange sampling sites are also located downstream of Augrabies Falls, a well-known biogeographic barrier (Skelton, 1986). Most internal nodes are supported by high posterior probabilities. The use of a cyt *b* divergence rate allowed us to tentatively date divergence between the species used in this analysis. We estimated that divergence between the ancestors of *L. intermedius* and the South African lineage occurred around 5.85 mya, with KwaZulu-Natal yellowfish splitting from the Orange-Vaal yellowfish 1.71 mya and subsequent division between Orange-Vaal yellowfish occurring around 1.43 mya. Populations within KwaZulu-Natal yellowfish diverged between 0.36-1.56 mya. This calibration

suggests more recent coalescence within the Orange-Vaal yellowfish lineages – around 0.46 mya within Orange-Vaal largemouth yellowfish and 0.28 mya within Orange-Vaal smallmouth yellowfish.

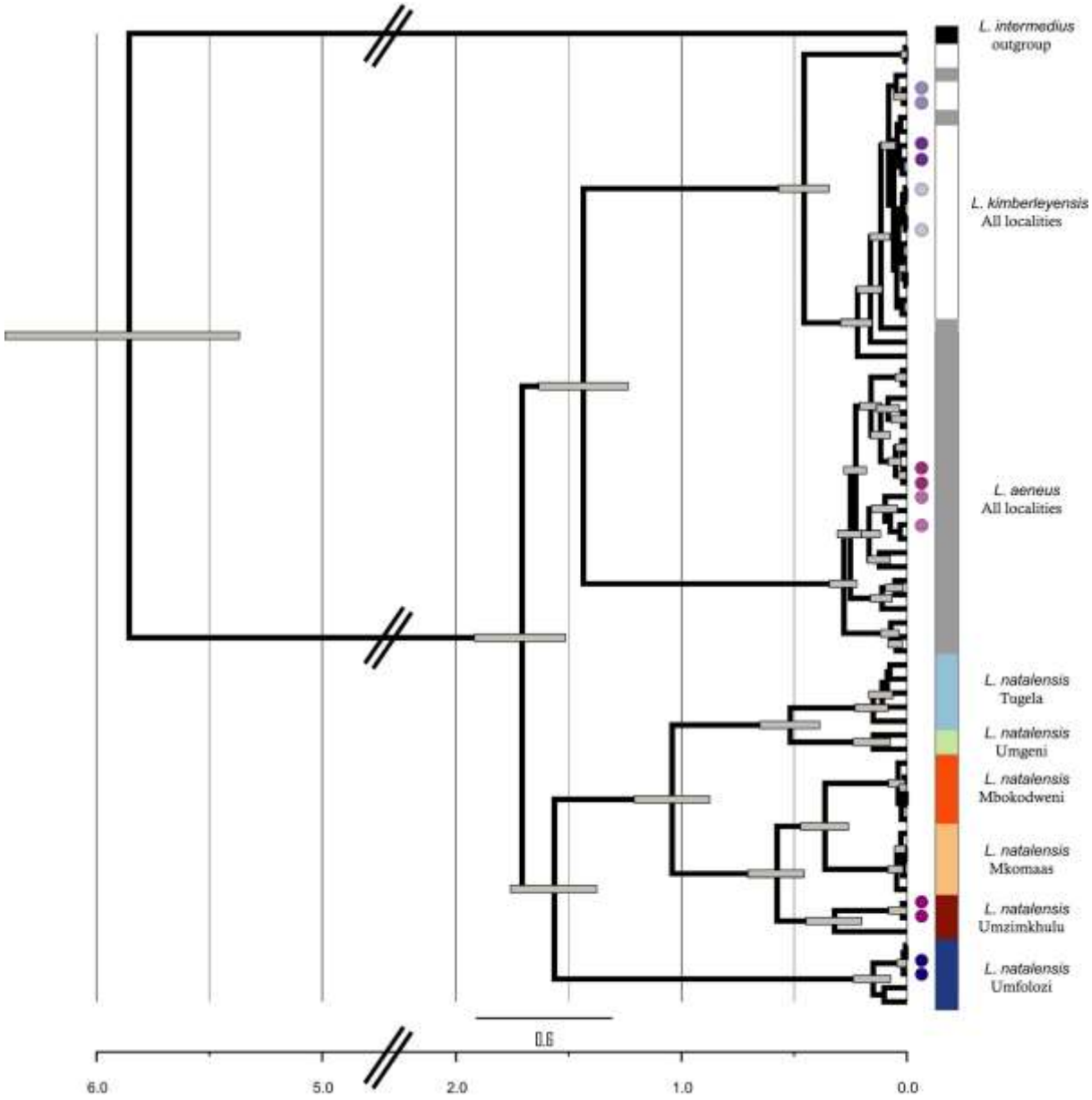

Figure 6: Bayesian Yule Process chronogram based on partial mitogenome reconstructions for 69 *Labeobarbus* samples (including seven replicate samples indicated by paired purple dots). The reference mitogenome of the East African *L. intermedius* was used as outgroup. Populations and species identified in this analysis are indicated by the coloured bar on the side. Colours of the KwaZulu-Natal yellowfish populations match those described in Stobie *et al*. (2018). Grey bars indicate 95% highest posterior densities (HPD) intervals for nodes with posterior probability support of 0.95 or higher. The grid scale indicates divergence dates in millions of years.

The Bayesian phylogram produced in MRBAYES from these data (Figure 7) supported our previous findings with each species forming a monophyletic cluster, and clear substructure within KwaZulu-Natal yellowfish. Most internal nodes on the tree were strongly supported. All replicates either clustered together or clustered within groups with low posterior probability support for separation.

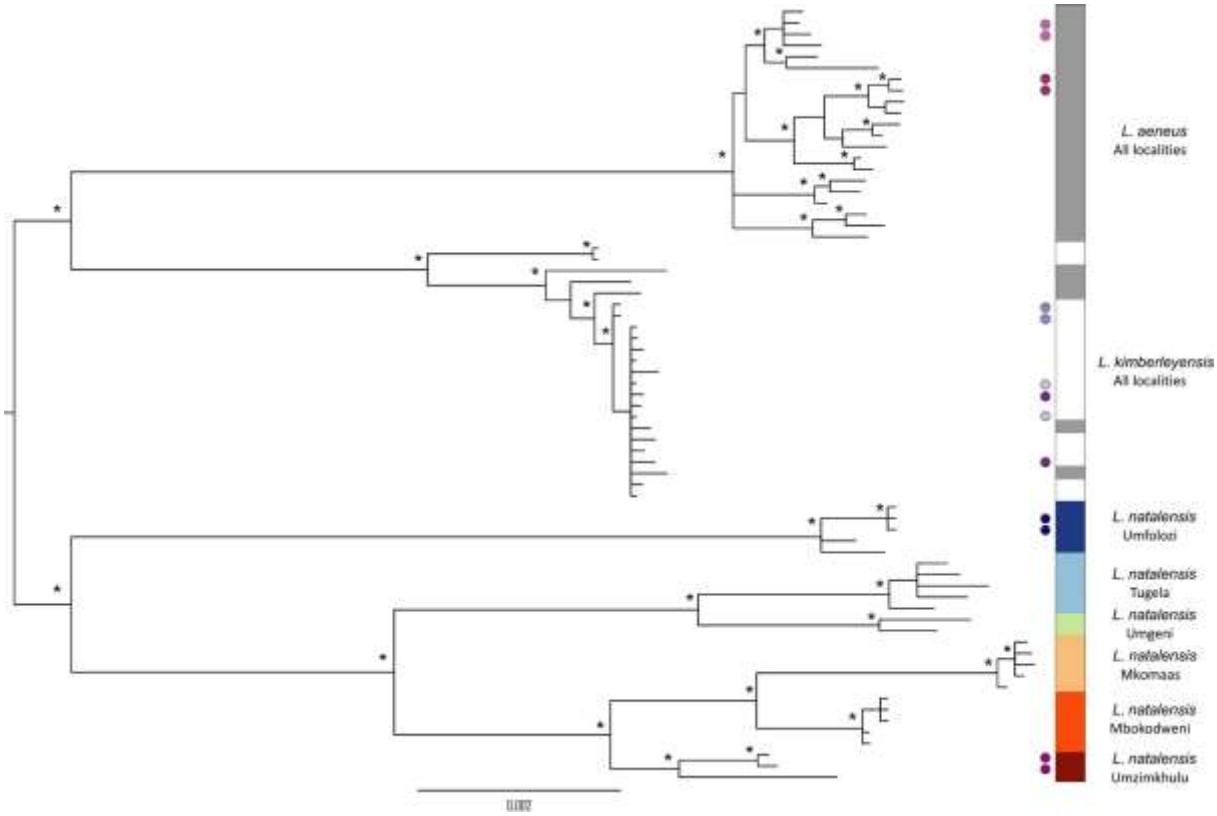

Figure 7: Phylogram produced in MRBAYES from partial mitogenome reconstructions. Different yellowfish populations identified in this study are indicated by the coloured bar. Replicate pairs are represented by paired purple dots. Nodal probability support values equal to or exceeding 0.95 are indicated by a *.

3.4. Comparison with Sanger sequencing

Sequences were obtained for all specimens of KwaZulu-Natal yellowfish (23 samples) for 928 bp of cyt *b*. This allowed us to compare 4,919 base pairs that were obtained in both the Sanger sequencing and mitogenome mapping approaches in 27 pairwise comparisons due to our inclusion of replicated samples. Of these 4,919 bases, we found only two errors, which occurred in the same individual from

the same read mapping, giving an estimated error rate of 0.04% across all comparisons with a maximum error rate of 0.68% within that individual.

3.5. Nuclear SNP phylogeny

The 665 nuclear SNPs for KwaZulu-Natal yellowfish and 984 SNPs for Orange-Vaal yellowfish were used to produce a pair of phylogenetic trees (Supporting Information Figure S6). These trees demonstrated very similar results to those obtained using the mined mitochondrial data, but with several key differences - namely that the Umzimkhulu lineage of KwaZulu-Natal yellowfish was subsumed within the Mkomaas lineage and is no longer distinct as has also been shown elsewhere (Stobie *et al*. 2018), and only one sample of Orange-Vaal yellowfish now appears to be introgressed. All other samples which had been identified to have an incongruent mitochondrial assignment compared to their morphological assignment seem to have a nuclear genotype supporting the morphological assignment. Mito-nuclear discordance appears to be occurring within these species.

The comparison of phylogenies produced from nuclear datasets to determine the effect of retaining mitochondrial SNPs showed subtle differences between trees, particularly for the KwaZulu-Natal yellowfish (Supporting Information Figure S7). The Umzimkhulu lineage which was incorporated into the Mkomaas using only nuclear SNPs was again found to be completely distinct once just 14 mitochondrial SNPs were retained in this analysis. The support values across the KwaZulu-Natal tree were also improved.

4. Discussion

In this study we have presented two approaches to mine mtDNA markers from RAD sequencing data. These informative mitochondrial polymorphisms are typically excluded from population genomic

analyses by filters for Hardy-Weinberg expectations of individual heterozygosity or by high coverage thresholds. RAD sequencing of two additional species of *Labeobarbus* has improved our understanding of diversity and speciation in this genus, with mitochondrial haplotype reconstruction and analysis adding insights that were not apparent from nuclear sequences alone, in particular when there is mito-nuclear discordance. Furthermore, our results show that it is more effective to extract mtDNA markers from NGS data than to re-sequence one or more mitochondrial markers, such as cyt *b*. Finally, we verified mitogenome data by comparing these to Sanger sequences of cyt *b* and contrasted nuclear and mitochondrial signal.

4.1. Sequencing and mapping

Previously we reported a higher than expected coverage of the nuclear genome from mapping RAD sequencing reads to the *Cyprinus carpio* nuclear genome (Stobie *et al*., 2018). In the present study, we found coverage of the mitochondrial genome was only slightly greater than expected from *in silico* digestion of the *L. intermedius* mitogenome assuming uniformly incomplete restriction digestion. The *in silico* digestion is based on a single reference sequence, but by using consensus across samples we were able to reconstruct a near-complete *Labeobarbus* mitogenome for each species. This increase in consensus coverage arises through point mutations that introduce the restriction enzyme recognition sites into new genomic regions while other mutations erase restriction sites elsewhere. The proportion of mitochondrial to nuclear reads estimated from our data is comparable to that found in studies of birds (Oswald, Overcast, Mauck, Andersen, & Smith, 2017) and far exceeds the proportion of mitochondrial DNA within the total genome. The average mitogenome mapping coverage is lower than the estimate produced by *in silico* digestion assuming incomplete enzyme digestion. The sequence coverage here falls between that expected for complete

digestion and uniformly incomplete digestion. The presence of targeted restriction sites within some sequence reads confirmed some degree of partial digestion.

Sample quality and library preparation influence the depth and coverage of genomic sequencing data but these artefacts have less impact on mitogenome mapping. The second Orange-Vaal yellowfish dataset included a greater quantity of data, giving greater depth and coverage than the other two batches combined. This may influence whole genome SNP comparisons, but due to the abundance of the comparatively small mitochondrial genome, this discrepancy appeared to have minimal impact here. This is shown by our comparable coverage across species and samples. Variation in depth and coverage across samples is unlikely to drastically influence biological signal due to the stochasticity from which uneven sequencing arose (Eaton *et al*., 2017). Repetitive elements are abundant in the nuclear dataset and may have affected mean GC content in this library. Variation in GC content and quality among batches of samples could also be due to unrecognised differences in library preparation. Differences between libraries can occur using RAD sequencing (Andrews *et al*., 2016), however we have accounted for this by the strict filtering criteria used in our analyses and do not anticipate this having any impact on our results.

4.2. STACKS-based mitochondrial SNP identification

We identified a relatively large proportion of mitochondrial SNPs through our STACKS-based pipeline. Mitochondrial SNPs comprised 1.29% of all SNPs identified in KwaZulu-Natal yellowfish and 0.06% of all SNPs in the Orange-Vaal yellowfish. The difference in the proportion of mitochondrial SNPs between species reflects variation in coverage and depth among sequence libraries, in particular the greater depth of the second batch of Orange-Vaal yellowfish data, resulting in many more SNPs passing the stringent filters in our pipeline.

The size of the mitochondrial genome led us to expect fewer mitochondrial SNPs, relative to the complete genome. In general, the higher diversity of mitochondrial sequences would make these less likely to be recognised as homologues. Also, mitochondrial SNP genotypes would be called as alternate homozygotes within different individuals, leading to a discrepancy between observed and expected heterozygosity. However, the comparative abundance of mitogenome reads appears to have resulted in high coverage across samples which may not share nuclear SNPs due to sequencing depth limitations. These markers are therefore likely to pass through a STACKS pipeline, although, given the rapid sorting of mtDNA, these markers may subsequently be excluded as F_{ST} outliers (Gleason & Burton, 2016; Stobie *et al.*, 2018).

Our assessment of mitochondrial read prevalence as minimum required coverage is increased provides evidence that although mitochondrial RAD tags are generally present at a higher coverage than nuclear markers (as shown by the increase in proportion of mtDNA tags as *-m* increases), merely using a high read coverage criteria will not result in the exclusion of all mtDNA tags. We demonstrate this by the decline in number of mtDNA tags as *-m* is increased. Although the sample space to find mtDNA tags is reduced, some tags are also lost. This suggests that studies intending to remove or isolate mtDNA loci should employ additional methods as we have here.

The mitochondrial SNPs passing through the STACKS pipeline provided variation exceeding that expected from resequencing a single mitochondrial gene. Haplotype networks and Yule Process phylogenetic trees produced from these mitochondrial SNPs split populations of each species, although some Orange-Vaal haplogroups clustered in a different group to that expected on morphologically assignment. This could result from incorrect morphological assignment of some individuals, from mitochondrial introgression, or from sharing of ancestral polymorphisms. Hybridization is believed to occur between these two species (Gaigher, 1976; Eccles, 1986; van Vuuren *et al.*, 1989; Mulder *et al.*, 1990), therefore we believe that mitochondrial introgression is a likely explanation. Our nuclear comparison adds an additional level of complexity here by showing

that all but one of these samples possesses nuclear genotypes concordant with the morphological assignment. This by itself demonstrates the practicality of this approach in identifying mito-nuclear discordance. These results should allow us to unravel the complex interplay between these sympatric species.

4.3. Organismal inferences from mitogenome mapping

Haplotype networks produced from the mapping approach reveal similar results but with more variability and greater resolution than the STACKS-based approach. Comparison of mitochondrial analyses with results from genome-wide SNP analyses also contributed to our interpretation of population structure in *Labeobarbus*. Results from mitochondrial haplotypes and nuclear SNPs (Stobie *et al.*, 2018) are concordant in splitting KwaZulu-Natal yellowfish into groups by drainage system. We observed relatively few mutations separating samples within populations of KwaZulu-Natal yellowfish. Yellowfish from the Umzimkhulu system carry a distinct mitochondrial lineage, as suggested in a previous study of mitochondrial control region sequences (Bloomer *et al*., 2007). These individuals were not distinguished from the Mkomaas population using nuclear RAD SNPs either here or in a previous study (Stobie *et al*., 2018). This may be indicative of a recent divergence that is not yet evident from sorting of nuclear polymorphisms, or irregular gene flow between populations which has allowed fixation of mitochondrial haplotypes despite sharing of nuclear alleles (Stobie *et al*., 2018). Alternatively, because we did not sample the southernmost limit of the distribution, the Mtamvuna River, it is possible that a population in this area possesses the distinct mtDNA lineage. Samples from the adjacent Umzimkhulu system may reflect mitochondrial introgression from this unsampled lineage. Despite this, allowing just 14 mitochondrial loci to be retained in the phylogeny produced from nuclear SNPs resulted in the clear split between these lineages, as well as improved support across the KwaZulu-Natal yellowfish tree. This suggests that filtering of mitochondrial SNPs from “nuclear” SNP datasets may be important in certain scenarios.

Samples of KwaZulu-Natal yellowfish from the Umgeni drainage carried a distinct haplogroup, which contrasts with our previous inference, based on analyses of nuclear RAD data, that this population is an admixture between Northern and Southern lineages (Stobie *et al*., 2018). Additionally, a single sample from the upper Umgeni system (Lions River) differed from those elsewhere in this system and could not be distinguished genetically from fish in the Tugela system, both in our mitochondrial and nuclear RAD analyses. This indicates a translocation event from the Mooi River (Tugela population) into the upper Lions River (Umgeni system), as a result of the Mooi-Mgeni Transfer Scheme (MMTS) (Hunter, 2009). The MMTS has been operational since 2003, with some earlier infrequent transfers (Hunter, 2009). Overall the results from the Umgeni suggest archaic admixture, a period of isolation, sufficient for the development of a unique mitochondrial lineage, and more recent human facilitated translocation via an interbasin transfer scheme. Although our lower Umgeni samples do not show evidence of recent admixture, these samples were collected by members of the Yellowfish Working Group during 2003-2006, therefore we do not know the current extent of introgression into the Umgeni population.

Haplotype reconstruction from SNPs and analyses based on mitogenome mapping revealed five samples of Orange-Vaal yellowfish clustering into a haplogroup that differed from that expected from morphological assignment. These individuals were morphologically classified as Orange-Vaal smallmouth yellowfish but in this analysis their mitochondrial haplotypes were found to cluster with those from Orange-Vaal largemouth yellowfish. This could be due to biological reasons we are currently investigating.

The species that was selected for mapping, *L. intermedius*, is somewhat distantly related to South African yellowfish, as expected given geographic separation and the history of southward radiation of this genus throughout the African continent (Tsigenopoulos *et al*., 2010). Consequently, we observed a long branch connecting it to the three South African *Labeobarbus* in our Bayesian constant coalescent chronogram. We tentatively dated the node of the common ancestors between *L.*

intermedius and the South African lineage as occurring around 5.9 mya, which fits with the proposed theory of invasion of Africa by early *Labeobarbus* in the Late Miocene (Tsigenopoulos *et al*., 2010). The same node based on cyt *b* divergence alone was previously estimated at 4.23 mya (Tsigenopoulos *et al*., 2010). The different node ages may be due to differences between using partial mitogenome reconstructions as opposed to a single gene. Divergence within the South African lineage occurs from around 1.71 mya, which is close to the proposed invasion of the ancient Orange River system by the common ancestor of South African smallscale yellowfishes 2-3 mya (Skelton, 2001) and is also concordant with the previous work using cyt *b* (Tsigenopoulos *et al*., 2010).

4.4. Sanger sequencing validation

Comparison of our mitogenome mapping approach with Sanger sequences revealed only two discrepancies, giving an estimated error rate of 0.04%. This shows that some mistakes are incorporated in the mapping approach, but these are relatively rare. The success of this approach is also influenced by divergence between the source organism and reference. We found greater error when we used *N. hexagonalepis* as our reference mitogenome for mapping (data not shown). However, the use of strict mapping parameters coupled with visual inspection of inconsistencies should remove most problematic areas from a mitogenome mapping. This allows one to extract useful mitochondrial sequences from data that are typically filtered to exclude all but the nuclear genome.

5. Conclusion

In this paper we have presented a novel approach to mining RAD data for mitochondrial sequences that can be isolated and analysed independently, potentially validating nuclear RAD SNP results or identifying genomic processes responsible for discordance, such as introgression. We identified these

markers using two different methods: a rigorous SNP discovery pipeline and direct mapping of sequence reads to a reference mitogenome. Both methods involved basic initial processing of reads for adaptor pollution and quality control. Our SNP discovery pipeline involved additional filters aimed at recognising common allelic variants within individuals and populations. Mitochondrial SNPs were confirmed using BLAST. Our mapping approach uses more of the raw data giving high coverage, averaging 54% of the mitogenome per sample and 98% coverage for a chimaeric consensus across southern African *Labeobarbus*, and finer resolution of mitochondrial haplotype variation than the SNP pipeline. However, the success of this approach depends on a suitable mitogenome reference. Finally, this approach also appears robust to variation in genome sequence quality and coverage.

Acknowledgements

We would like to thank the following colleagues for contributing to the sampling in this study: R. Bills, M. Nkosi, N. Rivers-Moore, J. Craigie, H. Plank, H.E. Filter, T. Wilkinson, R. Arderne, J. Wakelin and A. Howell; a special word of gratitude to our late colleague R. Karssing. We would also like to thank V. Savolainen, B. Hansson, U.K. Schliewen, E.J. Vreven, A.S.T. Papadopulos, I.E. Kiper and their research groups for insightful discussions about this project. Thank you to J. Day, R. Ogden and W.A. Cresko for their extremely helpful comments on the thesis from which this paper arose. This project was funded by the South African National Research Foundation (NRF), and the University of Pretoria's Genomics Research Institute (GRI). The financial assistance of the National Research Foundation (NRF; Innovation doctoral scholarship awarded to CSS and Incentive funding to PB, grant number 77240) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389-3402. doi: 10.1093/nar/25.17.3389

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81-92. doi: 10.1038/nrg.2015.28

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Retrieved from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, *22*(11), 3179-3190. doi: 10.1111/mec.12276

Avise, J. C. (1986). Mitochondrial DNA and the evolutionary genetics of higher animals. *Philosophical Transactions of the Royal Society London B*, *312*(1154), 325-342. Retrieved from: http://www.jstor.org/stable/2396333

Avise, J. C. (2009). Phylogeography: retrospect and prospect. *Journal of Biogeography*, *36*(1), 3-15. doi: 10.1111/j.1365-2699.2008.02032.x

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., ... & Saunders, N. C. (1987). Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, *18*(1), 489-522. doi: 10.1146/annurev.es.18.110187.002421

Avise, J. C., Bermingham, E., Kessler, L. G., & Saunders, N. C. (1984). Characterization of mitochondrial DNA variability in a hybrid swarm between subspecies of bluegill sunfish (*Lepomis macrochirus*). *Evolution*, *38*(5), 931-941. doi: 10.1111/j.1558-5646.1984.tb00364.x

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10), e3376. doi: 10.1371/journal.pone.0003376

Ballard, J. W. O., & Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Molecular Ecology*, *13*(4), 729-744. doi: 10.1046/j.1365-294X.2003.02063.x

Barkhuizen, L. M. (2017). *Labeobarbus kimberleyensis*. The IUCN Red List of Threatened Species 2017: e.T63292A100166441. Downloaded on **07 December 2017**. doi: 10.2305/IUCN.UK.2017-3.RLTS.T63292A100166441.en

Bermingham, E., & Avise, J. C. (1986). Molecular zoogeography of freshwater fishes in the southeastern United States. *Genetics*, *113*(4), 939-965.

Bermingham, E., & Moritz, C. (1998). Comparative phylogeography: concepts and applications. *Molecular Ecology*, *7*(4), 367-369. doi: 10.1046/j.1365-294x.1998.00424.x

Bloomer, P., Bills, I. R., van der Bank, F. H., Villet, M. H., Jones, N., & Walsh, G. (2007). *Multidisciplinary investigation of differentiation and potential hybridisation between two yellowfish species* Labeobarbus kimberleyensis *and* L. aeneus *from the Orange-Vaal system. Follow-up study 2004-2007*. Yellowfish Working Group Report (pp. 1-67). Johannesburg, South Africa: Federation of South African Flyfishers and AngloGold Ashanti Ltd.

Brito, P. H., & Edwards, S. V. (2009). Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, *135*(3), 439-455. doi: 10.1007/s10709-008-9293-3

Brown, W. M., George, M., & Wilson, A. C. (1979). Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences*, *76*(4), 1967-1971. doi: 10.1073/pnas.76.4.1967

Catchen, J. M., Amores, A., Hohenlohe, P. A., Cresko, W. A., & Postlethwait, J. H. (2011). Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, *1*(3), 171-182. doi: 10.1534/g3.111.000240

Catchen, J. M., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular ecology*, *22*(11), 3124-3140. doi: 10.1111/mec.12354

Clement, M., Posada, D., & Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Molecular ecology*, *9*(10), 1657-1659. doi: 10.1046/j.1365-294x.2000.01020.x

Cruaud, A., Gautier, M., Galan, M., Foucaud, J., Sauné, L., Genson, G., ... & Rasplus, J. Y. (2014). Empirical assessment of RAD sequencing for interspecific phylogeny. *Molecular Biology and Evolution*, *31*(5), 1272-1274. doi: 10.1093/molbev/msu063

DaCosta, J. M., & Sorenson, M. D. (2014). Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE*, *9*(9), e106713. doi: 10.1371/journal.pone.0106713

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013). Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology*, *22*(11), 3151-3164. doi: 10.1111/mec.12084

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499-510. doi: 10.1038/nrg3012

Dowling, T. E., Tibbets, C. A., Minckley, W. L., & Smith, G. R. (2002). Evolutionary relationships of the plagopterins (Teleostei: Cyprinidae) from cytochrome *b* sequences. *Copeia*, *2002*(3), 665-678. doi: 10.1643/0045-8511(2002)002[0665:EROTPT]2.0.CO;2

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, *29*(8), 1969-1973. doi: 10.1093/molbev/mss075

Eaton, D. A., Spriggs, E. L., Park, B., & Donoghue, M. J. (2017). Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic biology*, *66*(3), 399-412. doi: 10.1093/sysbio/syw092

Eccles, D. H. (1986). Development of the gut in the South African cyprinid fish *Barbus aeneus* (Burchell). *African Zoology*, *21*(2), 165-169. Retrieved from: http://hdl.handle.net/10520/AJA00445096_1008

Edwards, S., & Bensch, S. (2009). Looking forwards or looking backwards in avian phylogeography? A comment on Zink and Barrowclough 2008. *Molecular Ecology*, *18*(14), 2930-2933. doi: 10.1111/j.1365-294X.2009.04270.x

Ekblom, R., Smeds, L., & Ellegren, H. (2016). Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC Genomics*, *15*(467). doi: 10.1186/1471-2164-15-467

Estoup, A., Largiader, C. R., Perrot, E., Chourrout, D. (1996) Rapid one-tube DNA extraction for reliable PCR detection of fish polymorphic markers and transgenes. *Molecular Marine Biology and Biotechnology*, *5*, 295-298.

Funk, D. J., & Omland, K. E. (2003). Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 397-423. doi: 10.1146/annurev.ecolsys.34.011802.132421

Gaigher, I. G. (1976). The reproduction of *Barbus cf. kimberleyensis* (Pisces, Cyprinidae) in the Hardap Dam, South West Africa. *African Zoology*, *11*(1), 97-110. doi: 10.1080/00445096.1976.11447519

Gleason, L. U., & Burton, R. S. (2016). Genomic evidence for ecological divergence against a background of population homogeneity in the marine snail *Chlorostoma funebralis*. *Molecular Ecology*, *25*(15), 3557-3573. doi: 10.1111/mec.13703

Hasegawa, M., Kishino, H., & Yano, T. A. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, *22*(2), 160-174. doi: 10.1007/BF02101694

Herrera, S., Watanabe, H., & Shank, T. M. (2015). Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Molecular Ecology*, *24*(3), 673-689. doi: 10.1111/mec.13054

Hunter, A. M. S. (2010). *A review of the fluvial geomorphology monitoring of the receiving streams of the Mooi-Mgeni River Transfer Scheme Phase 1* (Masters of Environment and Development thesis). University of KwaZulu-Natal.

Jeffries, D. L., Copp, G. H., Lawson Handley, L., Olsén, K. H., Sayer, C. D., & Hänfling, B. (2016). Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Molecular Ecology*, *25*(13), 2997-3018. doi: 10.1111/mec.13613

Jezkova, T., Riddle, B. R., Card, D. C., Schield, D. R., Eckstut, M. E., & Castoe, T. A. (2015). Genetic consequences of postglacial range expansion in two codistributed rodents (genus *Dipodomys*) depend on ecology and genetic locus. *Molecular Ecology*, *24*(1), 83-97. doi: 10.1111/mec.13012

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772-780. doi: 10.1093/molbev/mst010

Levin, D. A. (1983). Polyploidy and novelty in flowering plants. *The American Naturalist*, *122*(1), 1-25. doi: 10.1086/284115

Lischer, H. E., & Excoffier, L. (2011). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298-299. doi: 10.1093/bioinformatics/btr642

Macher, J. N., Rozenberg, A., Pauls, S. U., Tollrian, R., Wagner, R., & Leese, F. (2015). Assessing the phylogeographic history of the montane caddisfly *Thremma gallicum* using mitochondrial and restriction-site-associated DNA (RAD) markers. *Ecology and Evolution*, *5*(3), 648-662. doi: 10.1002/ece3.1366

Marrano, A., Birolo, G., Prazzoli, M. L., Lorenzi, S., Valle, G., & Grando, M. S. (2017). SNP-discovery by RAD-sequencing in a germplasm collection of wild and cultivated grapevines (*V. vinifera* L.). *PloS one*, *12*(1), e0170655. doi: 10.1371/journal.pone.0170655

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, *17*(2), 240-248. doi: 10.1101/gr.5681207

Moura, A. E., Kenny, J. G., Chaudhuri, R. R., Hughes, M. A., Reisinger, R. R., De Bruyn, P. J. N., ... & Hoelzel, A. R. (2015). Phylogenomics of the killer whale indicates ecotype divergence in sympatry. *Heredity*, *114*(1), 48-55. doi: 10.1038/hdy.2014.67

Mulder, P. F. S., van Vuuren, N. G., Ferreira, J. T., & van der Bank, F. H. (1990). A preliminary biochemical comparison of two species of the genus *Barbus* from the Vaal River system. *Water S. A.*, *16*(3), 147-150.

Oellermann, L. K., & Skelton, P. H. (1990). Hexaploidy in yellowfish species (*Barbus*, Pisces, Cyprinidae) from southern Africa. *Journal of Fish Biology*, *37*(1), 105-115. doi: 10.1111/j.1095-8649.1990.tb05932.x

Oswald, J. A., Overcast, I., Mauck, W. M., Andersen, M. J., & Smith, B. T. (2017). Isolation with asymmetric gene flow during the nonsynchronous divergence of dry forest birds. *Molecular Ecology*, *26*(5), 1386-1400. doi: 10.1111/mec.14013

Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S. C., Boisselier, M. C., & Samadi, S. (2015). Use of RAD sequencing for delimiting species. *Heredity*, *114*(5), 450-459. doi: doi:10.1038/hdy.2014.105

Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution*, *8*(10), 1360-1373. doi: 10.1111/2041-210X.12775

Perry, K. D., Pederson, S. M., & Baxter, S. W. (2017). Genome-wide SNP discovery in field and laboratory colonies of Australian *Plutella* species. *bioRxiv*, 141606. doi: 10.1101/141606

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS ONE*, *7*(5), e37135. doi: 10.1371/journal.pone.0037135

Puckett, E. E., Etter, P. D., Johnson, E. A., & Eggert, L. S. (2015). Phylogeographic analyses of American black bears (*Ursus americanus*) suggest four glacial refugia and complex patterns of postglacial admixture. *Molecular Biology and Evolution*, *32*(9), 2338-2350. doi: 10.1093/molbev/msv114

Pujolar, J. M., Jacobsen, M. W., Als, T. D., Frydenberg, J., Munch, K., Jónsson, B., ... & Hansen, M. M. (2014). Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology*, *23*(10), 2514-2528. doi: 10.1111/mec.12753

QGIS Development Team (2016). Quantum GIS Geographic Information System. Open Source Geospatial Foundation Project. Retrieved from: http://www.qgis.org/en/site/

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572-1574. doi: 10.1093/bioinformatics/btg180

Skelton, P. H. (1986). Fish of the Orange-Vaal system. In B. R. Davies & K. F. Walker (Eds.), *The ecology of river systems* (pp. 143-162). Dordrecht, The Netherlands: Dr W. Junk Publishers.

Skelton, P. H. (2001). *A complete guide to the freshwater fishes of southern Africa*. Cape Town, South Africa: Struik.

Skelton, P. H., & Bills, I. R. (2008). An introduction to African yellowfish and to this report. In N. D. Impson, I. R. Bills, & L. Wolhuter (Eds.), *Technical report on the state of yellowfishes in South Africa* (pp. 1-14) (WRC Report No. KV 212/08). Pretoria, South Africa: Water Research Commission.

Stobie, C. S., Oosthuizen, C. J., Cunningham, M. J., & Bloomer, P. (2018). Exploring the phylogeography of a hexaploid freshwater fish by RAD sequencing. *Ecology and Evolution*, *8*(4), 2326-2342. doi: 10.1002/ece3.3821

Streicher, J. W., Devitt, T. J., Goldberg, C. S., Malone, J. H., Blackmon, H., & Fujita, M. K. (2014). Diversification and asymmetrical gene flow across time and space: lineage sorting and hybridization in polytypic barking frogs. *Molecular Ecology*, *23*(13), 3273-3291. doi: 10.1111/mec.12814

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725-2729. doi: 10.1093/molbev/mst197

Terraneo, T. I., Arrigoni, R., Benzoni, F., Forsman, Z. H., & Berumen, M. L. (2018). The complete mitochondrial genome of *Porites harrisoni* (Cnidaria: Scleractinia) obtained using next-generation sequencing. *Mitochondrial DNA Part B*, *3*(1), 286-287. doi: 10.1080/23802359.2018.1443852

Toews, D. P., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*(16), 3907-3930. doi: 10.1111/j.1365-294X.2012.05664.x

Truong, H. T., Ramos, A. M., Yalcin, F., de Ruiter, M., van der Poel, H. J., Huvenaars, K. H., ... & van Eijk, M. J. (2012). Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS ONE*, *7*(5), e37565. doi: 10.1371/journal.pone.0037565

Tsigenopoulos, C. S., Kasapidis, P., & Berrebi, P. (2010). Phylogenetic relationships of hexaploid large-sized barbs (genus *Labeobarbus*, Cyprinidae) based on mtDNA data. *Molecular Phylogenetics and Evolution*, *56*(2), 851-856. doi: 10.1016/j.ympev.2010.02.006

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, *18*(7), 411. doi: 10.1038/nrg.2017.26

van Vuuren, N. G., Mulder, P. F. S., Ferreira, J. T., & van der Bank, F. H. (1989). The identification of hybrids of *Barbus aeneus* X *B. kimberleyensis* and *Labeo capensis* X *L. umbratus* in Hardap Dam, SWA/Namibia. *Madoqua*, *16*(1), 27-34. Retrieved from: http://hdl.handle.net/10520/AJA10115498_497

Vreven, E. J., Musschoot, T., Snoeks, J., & Schliewen, U. K. (2016). The African hexaploid Torini (Cypriniformes: Cyprinidae): review of a tumultuous history. *Zoological Journal of the Linnean Society*, *177*(2), 231-305. doi: 10.1111/zoj.12366

Wallace, D. C. (2008). Mitochondria as chi. *Genetics*, *179*(2), 727-735. doi: 10.1534/genetics.104.91769

Wilson, A. C., Cann, R. L., Carr, S. M., George, M., Gyllensten, U. B., Helm-Bychowski, K. M., ... & Stoneking, M. (1985). Mitochondrial DNA and two perspectives on evolutionary genetics. *Biological Journal of the Linnean Society*, *26*(4), 375-400. doi: 10.1111/j.1095-8312.1985.tb02048.x

Yang, L., Sado, T., Hirt, M. V., Pasco-Viel, E., Arunachalam, M., Li, J., ... & Miya, M. (2015). Phylogeny and polyploidy: resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). *Molecular Phylogenetics and Evolution*, *85*, 97-116. doi: 10.1016/j.ympev.2015.01.014

Zhan, S. H., Glick, L., Tsigenopoulos, C. S., Otto, S. P., & Mayrose, I. (2014). Comparative analysis reveals that polyploidy does not decelerate diversification in fish. *Journal of Evolutionary Biology*, *27*(2), 391-403. doi: 10.1111/jeb.12308

Zhang, D. X., & Hewitt, G. M. (2003). Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, *12*(3), 563-584. doi: 10.1046/j.1365-294X.2003.01773.x

Data Accessibility

RAD analyses:

Raw sequence data for the Orange-Vaal yellowfish, quality control output, filtering regular expressions designed for this study, and processed input files for the final datasets (FASTA, GENEPOP, NEXUS, STACKS output files and BEAST XML files) are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.75fh336. Raw sequence data for the KwaZulu-Natal yellowfish is available at https://doi.org/10.5061/dryad.g00b7.

Sanger sequencing:

Cyt *b* sequences have been made available on GenBank (MH936322-MH936344).

Author Contributions

M.J.C. and C.S.S. designed this stage of the research project and contributed to conceptual design of the project and bioinformatics analyses. P.B. obtained research funding. P.B., C.J.O. and M.J.C. supervised the research progress. C.S.S. performed research, developed custom scripts, analysed and interpreted the data and led the writing of the manuscript. All authors contributed to the final draft of the manuscript.