

**Population genomics and phylogeography of South
African *Labeobarbus* (Rüppel, 1835) species**

by

Connor Seamus Stobie

Submitted in partial fulfilment of the requirements for the degree

Philosophiae Doctor

in the

Department of Genetics

Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

April 2018

Supervisor: Prof. Paulette Bloomer

Co-supervisors: Dr Carel J. Oosthuizen & Dr Michael J. Cunningham

To my parents, Cheryl and Bruce.

Declaration

I, the undersigned, declare that the thesis which I hereby submit for the degree *Philosophiae Doctor* at the University of Pretoria, is my own work and has not previously been submitted for a degree at this or any other tertiary institution.

C Stobie

Mr Connor S. Stobie

December 2017

Ethics Statement

The author, whose name appears on the title page of this thesis, has obtained, for the research described in this work, the applicable research ethics approval.

The author declares that he has observed the ethical standards required in terms of the University of Pretoria's Code of ethics for researchers and the Policy guidelines for responsible research.

Population genomics and phylogeography of South African *Labeobarbus* (Rüppel, 1835) species

Connor Seamus Stobie

Department of Genetics

Philosophiae Doctor

Supervisors: Prof. Paulette Bloomer

Molecular Ecology and Evolution Programme

Department of Genetics

University of Pretoria

Dr Carel Oosthuizen

Department of Zoology & Entomology

University of Pretoria

Dr Michael Cunningham

Molecular Ecology and Evolution Programme

Department of Genetics

University of Pretoria

Summary:

The hexaploid freshwater fish genus *Labeobarbus* are distributed across Africa. These large fish are renowned for their use in game and subsistence fishing. Recent taxonomic changes demonstrate the difficulty in delimiting species in this genus, primarily due to factors such as morphological plasticity, intraspecific variation and hybridisation between taxa. This is particularly true for several of the South African species. The KwaZulu-Natal yellowfish (*L. natalensis*) is renowned for intraspecific

variation particularly in mouth morphotype. This wide range of variation has resulted in the species being described as up to 12 different species previously. Recent mitochondrial genetic analyses have suggested that there is substantial variation between populations of *L. natalensis* delimited by drainage system. Elsewhere, the Orange-Vaal smallmouth and largemouth yellowfishes (*L. aeneus* and *L. kimberleyensis*) occur sympatrically and are morphologically, ecologically and behaviourally distinct. However, a number of studies have reported hybrids occurring in the wild, particularly in artificial habitats such as dams. Some studies have even questioned the validity of these species due to their suspected propensity to hybridise and the inability for allozymes to distinguish the species.

In this study I examine the phylogeography of these three South African species using a population genomics approach. I use a reduced-representation Next-Generation Sequencing (NGS) approach known as double-digest Restriction-site Associated DNA (ddRAD) sequencing to partially bypass some of the complications of working with a hexaploid genome. I investigate the genetic structure and phylogeographic patterns and processes influencing these populations.

While exploring the phylogeography of *L. natalensis* I was able to delimit five populations using nuclear Single Nucleotide Polymorphisms (SNPs), contrary to previous studies that identified six geographically separate mitochondrial DNA (mtDNA) lineages. Discordance between these results was re-examined following the identification of a new approach to mine mitochondrial loci from Restriction-site Associated DNA (RAD) data which I showcase as a separate research chapter. My partial mitogenome reconstructions agreed with previous studies in identifying six mitochondrial lineages. This suggests that gene flow between populations together with local fixation of mitochondrial haplotypes may have resulted in the discrepancy between results, which points to a flaw in approaches such as DNA barcoding. I further focus on several methodological aspects, particularly for data filtering and SNP identification, where unexpected results were obtained.

The mitogenome mining approach I have developed is highly useful for RAD studies which seek to verify their nuclear RAD results using mitochondrial markers. I demonstrate that there is no need to independently sequence mitochondrial genes as the marker merely needs to be extracted from

the RAD data. Discordance between mitochondrial and nuclear results can suggest the presence of certain processes occurring such as introgression, as I show here.

Finally, I examine the genetic differentiation of the two species of Orange-Vaal yellowfishes and determine that a few loci potentially under selection and the mitochondrial haplotype show genetic differentiation between the two species, against a background of relative homogeneity. Some evidence of hybridisation and bidirectional introgression was obtained between the two species, with one sample identified as a recent hybrid.

Preface

Members of the genus *Labeobarbus* are large, slow growing fish of the minnow family Cyprinidae distributed in freshwaters across Africa. These fishes are a popular target for game and subsistence fishing, particularly in South Africa where they are known as yellowfish. They are most interesting from a genetics perspective as all members of the genus are hexaploid. The South African members of the genus number seven species, three of which were investigated in this study. Historically, the genus has been the subject of much debate due to the extreme degree of morphological plasticity or intraspecific variation together with the common occurrence of hybridisation.

The KwaZulu-Natal yellowfish, *L. natalensis*, is endemic to the province of KwaZulu-Natal, South Africa. The species occurs widely across drainage systems despite the presence of geographical barriers such as waterfalls. Additionally, this species is known to exhibit a wide range of different morphotypes, and has historically been described as many different species. Initial genetic studies on this species have suggested much diversity between the drainage systems. Identification of these populations would aid in effective conservation management of the species.

The Orange-Vaal smallmouth and largemouth yellowfishes (*L. aeneus* and *L. kimberleyensis*) are co-occurring species distributed across the Orange-Vaal system. The two species fill very different ecological roles in the system, and are morphologically distinct. However, putative hybrids have been reported in the literature, particularly in dams. Previous studies have not been successful in determining whether hybridisation is indeed occurring in riverine habitat. Some studies have even questioned the validity of the two species.

In this study I sought to investigate the phylogeography of these three species of *Labeobarbus* using Next-Generation Sequencing (NGS) markers in a population genomics framework. This enabled me to determine the populations of *L. natalensis* across the distribution, and the patterns and processes acting on this species. I assessed the genetic distinction between the Orange-Vaal yellowfishes using this approach. Finally, I assessed whether intra- and interspecific hybridisation are occurring within and between these species, with a focus on the sympatric Orange-Vaal yellowfishes.

The research chapters in this thesis have been structured as independent research articles. There is therefore a small degree of duplication between some of the results. The style followed is that of the journal *Molecular Ecology*.

Chapter 1 reports on a broad overview of the research in the literature relating to this study. This chapter attempts to place the study within a broader framework. I focus here on the importance of the genus together with difficulties in its taxonomy and the processes contributing to this, with a brief look at the background for each study species. I then move on to discuss the role of phylogeography and conservation using genetic markers followed by advances with genomics. Finally, I provide specific details relating to previous research on the study species and particular scenarios I will examine in this project, leading me to the research questions, hypotheses and predictions.

Chapter 2 examines the phylogeography of *L. natalensis* across the distribution using a reduced-representation genomic approach, double-digest Restriction-site Associated DNA (ddRAD) sequencing. I add additional focus to the methodological constraints and challenges faced during this process, with an emphasis on proper filtering of NGS data. The population genetic structure of the species is assessed and further explored by modelling the population history. Evidence of intraspecific hybridisation is found and briefly discussed. I also provide estimates of several population genetic parameters, and suggestions for further research into effective conservation management of the populations identified in this study. This study has been accepted for publication by the journal *Ecology and Evolution*. My co-authors on this paper are Prof. Paulette Bloomer, Dr Carel Oosthuizen and Dr Michael Cunningham.

Chapter 3 presents a novel approach for mining mitochondrial reads from NGS data, resulting in either candidate SNPs to assess mitochondrial signal or in partial mitogenome reconstructions. This approach is intended to provide researchers with a better framework for fully using their NGS data, as many researchers currently sequence mitochondrial markers independently after NGS. I validate the mitogenome reconstructions using Sanger sequencing of a mitochondrial gene. Finally, I use the mitogenome reconstructions to assess the mitochondrial relationships between the species, finding evidence of mitonuclear discordance and introgression. This chapter will be submitted for publication early in 2018 with my co-authors Prof. Paulette Bloomer, Dr Carel Oosthuizen and Dr Michael Cunningham.

Chapter 4 investigates potential hybridisation and introgression of *L. aeneus* and *L. kimberleyensis* across the upper and lower Orange River. I also determine whether these two sympatric species are genetically differentiated. I make use of both nuclear markers identified in this Chapter as well as mitochondrial markers found in Chapter 3. I demonstrate a technical bias affecting the dataset as a result of restriction enzyme choice, which may be useful for further research. I also test for the effect of loci under selection, the direction of any introgression, and genetic structure within species. This chapter will be submitted for publication in 2018 with my co-authors Prof. Paulette Bloomer, Dr Carel Oosthuizen and Dr Michael Cunningham.

The final chapter or **Synthesis** assesses the results of this study as a whole, particularly with respect to the aims determined at the start of this project and how I achieved them. I examine how I answered my research questions before providing a summary of additional findings of this study primarily focussing on the methodology employed, challenges faced and suggestions for management and future research.

Acknowledgements

“Once there were brook trout in the streams in the mountains. You could see them standing in the amber current where the white edges of their fins wimpled softly in the flow. They smelled of moss in your hand. Polished and muscular and torsional. On their backs were vermiculate patterns that were maps of the world in its becoming. Maps and mazes. Of a thing which could not be put back. Not be made right again. In the deep glens where they lived all things were older than man and they hummed of mystery.”

- Cormac McCarthy, *The Road*

I would like to thank each of my supervisors for the hours of discussion, brainstorming, support, mentorship and fun that came with working closely with them on this project. In particular, I would like to thank Paulette Bloomer for that call out of the blue inviting me to her research group, for being the final word in our discussions, and for always being there with empathy and understanding. Thank you to Carel Oosthuizen for being a partner in crime, always willing to bend an ear and point to the best way to approach something. And thank you to Michael Cunningham, who was always up for an impromptu chat about absolutely anything and from whom I have learned so very much. Thank you to all three of you for the many opportunities you have opened up for me along the way, and the guidance you have shown me on my journey.

Thank you to everyone in MEEP (Molecular Ecology and Evolution Programme) for all of your help in this endeavour. In particular, I would like to thank my wonderful girlfriend Ilkser Kiper for always being supportive, teaching me so much, and generally being super awesome. Thank you to Thierry Hoareau, Arrie Klopper and Kerry Reid for your mentorship, advice, and criticisms. Thank you to Amanda Maswanganye, Sarita Maree, Tonna Mojekwu and Samantha Mynhardt for all of your support and advice. Thank you to the Department of Genetics at large as everyone has been so friendly and helpful during my time here. And thank you to the Department of Genetics at my previous institution, the University of KwaZulu-Natal, for all that they taught me and for guiding me prior to the start of this particular journey. In particular, thank you to Sandi Willows-Munro and Ché Pillay who have both been momentous mentors, and their respective research groups. Thank you also to all collaborators and advisors who hosted me on research visits related to this degree and engaged in astute discussion on how best to deal with the “darned ploidy!”, namely Prof. Vincent Savolainen and his research group at Imperial College London (thank you also for taking me along to see the Queen), Dr Ulrich Schliewen and his research group at Zoologische Staatssammlung München (Munich) and Prof. Bengt Hansson and his research group at Lund University.

Thank you to all of my friends, particularly to Minke, Kisten, Ilkadim, Rush, Brandon, Jody, Colin, Alice, Geoff, Gareth, Eduan, Bevan, Kathleen, Jason, Giancarlo, Grant, Kevin, Ismier and Akshay. You have all been most supportive and understanding during this time.

Thank you to my family for always supporting, loving and encouraging me. Thank you to Cheryl, Bruce, Cait, Mel, Blake, Ali, John, Kieran, Cara, Fintan, Juno and Otis. I could not have done this without your love and support.

Table of Contents

Chapter 1: New tools to solve old challenges: Assessing phylogeography and population genomics with ddRAD sequencing in the hexaploid freshwater fish genus *Labeobarbus*..... 1

1.1 The broad placement and importance of <i>Labeobarbus</i>	1
1.2 Taxonomic difficulties concerning <i>Labeobarbus</i>	2
1.3 <i>Labeobarbus</i> of South Africa and description of the species used in this study	7
1.4 <i>Labeobarbus natalensis</i> , the KwaZulu-Natal yellowfish	10
1.5 The Orange-Vaal yellowfishes, <i>L. aeneus</i> & <i>L. kimberleyensis</i>	13
1.6 Conservation and phylogeography in an era of genetics	14
1.7 Conservation and phylogeography in an era of genomics: The rise of population genomics....	21
1.8 Context of this study	27
1.9 Research questions	30
1.10 Hypotheses and predictions	30
1.11 References	33

Chapter 2: Exploring the phylogeography of a hexaploid freshwater fish by RAD sequencing 43

Abstract.....	43
2.1 Introduction	44
2.2 Materials and methods.....	48
2.2.1 Sample collection and DNA extraction	48
2.2.2 Library preparation and sequencing.....	49
2.2.3 Bioinformatics pipeline, quality control and read mapping	49
2.2.4 SNP discovery.....	50
2.2.5 Population genetic parameters and structure.....	52
2.2.6 Population history.....	53
2.3 Results.....	54
2.3.1 Library preparation and sequencing.....	54
2.3.2 Bioinformatics, mapping and SNP discovery	55
2.3.3 Population structure and variation.....	56
2.3.4 Population history.....	65
2.4 Discussion.....	66
2.4.1 Sequencing and mapping.....	66
2.4.2 Bioinformatics and SNP discovery	68
2.4.3 Population genetic parameters and structure.....	72
2.4.4 Population history.....	75

2.5 Conclusion	76
2.6 Acknowledgements.....	77
2.7 References	77
2.8 Supporting Information	84
Chapter 3: Finding stories in noise: Mitochondrial portraits from RAD data	94
Abstract.....	94
3.1 Introduction	95
3.2 Materials and Methods.....	98
3.2.1 Sample collection and DNA extraction	98
3.2.2 Library preparation and sequencing.....	100
3.2.3 Bioinformatics and mitochondrial SNP discovery.....	100
3.2.4 Mitogenome assembly.....	102
3.2.5 Comparison with Sanger sequencing.....	103
3.3 Results.....	105
3.3.1 Library preparation and sequencing.....	105
3.3.2 Bioinformatics and mitochondrial SNP discovery.....	105
3.3.3 Mitogenome assembly.....	108
3.3.4 Comparison with Sanger sequencing.....	115
3.4 Discussion.....	115
3.4.1 Sequencing and mapping.....	115
3.4.2 STACKS-based mitochondrial SNP identification	116
3.4.3 Mitogenome mapping SNP identification.....	117
3.4.4 Sanger sequencing validation	121
3.5 Conclusion	121
3.6 Acknowledgements.....	123
3.7 References	123
3.8 Supporting Information	127
Chapter 4: Investigating hybridisation and introgression of the Orange-Vaal yellowfish species <i>Labeobarbus aeneus</i> and <i>L. kimberleyensis</i>	135
Abstract.....	135
4.1 Introduction	136
4.2 Materials and methods.....	139
4.2.1 Sample collection and DNA extraction	139
4.2.2 Library preparation and sequencing.....	140

4.2.3 Bioinformatics pipeline and quality control.....	141
4.2.4 SNP discovery.....	142
4.2.5 Mitogenome read mapping	144
4.2.6 Population genetic parameters, structure, and assessment of nuclear hybridisation.	144
4.3 Results.....	146
4.3.1 Library preparation, sequencing and bioinformatics.....	146
4.3.2 SNP discovery.....	147
4.3.3 Mitogenome read mapping	147
4.3.4 Population genetic parameters, structure and assessment of nuclear hybridisation..	148
4.4 Discussion.....	152
4.4.1 Library preparation, sequencing and bioinformatics.....	152
4.4.2 SNP discovery.....	154
4.4.3 Mitogenome read mapping	155
4.4.4 Population genetic parameters, structure and assessment of hybridisation.....	157
4.5 Conclusion: Hybridisation and introgression of Orange-Vaal yellowfishes	160
4.6 References	162
4.7 Supporting Information	166
Chapter 5: Synthesis.....	177
5.1 Aims.....	177
5.2 Observations from this study.....	178
5.3 RAD sequencing	180
5.4 Data filtering and SNP identification.....	184
5.5 Assessing population structure and demography	186
5.6 Conservation and management.....	187
5.7 Future directions.....	188
5.8 Conclusion.....	189
5.9 References	189

List of Tables

Table 1.1 Species-specific microsatellite loci identified from repeat-enriched libraries for <i>Labeobarbus natalensis</i>	20
Table 1.2 Intron and exon markers tested which yielded useable sequences in the pilot study. The target size indicated is based on closely-related genera in Cyprinidae	20
Table 2.1 Summary information from initial analysis of RAD sequencing data	54
Table 2.2 Mapping of RAD sequencing reads to the <i>C. carpio</i> nuclear (GCF_000951615.1) genome	55
Table 2.3 Summary of the SNP identification process and filtering steps to the final SNP dataset	56
Table 2.4 Summary data produced by the POPULATIONS program in STACKS for variant positions (top) and all positions (bottom). Populations are: Umf = Umfolozi; Tug = Tugela; Mko = Mkomas; Mbo = Mbokodweni; Umg = Umgeni. Summary data were also calculated for all individuals as a single population (Total) and for the northern (North) and southern (South) lineages identified in this study, excluding the potentially admixed Umgeni lineage	64
Table 2.5 Pairwise F_{ST} values between the five populations identified in this study	65
Table 3.1 Summary statistics from initial analysis of RAD sequencing data	105
Table 4.1 SNP identification process summary and filtering steps to obtain the final SNP dataset	147
Table 4.2 Summary statistics (from POPULATIONS) for variant positions (top) and all positions (bottom) after the potentially misidentified sample was removed from the analysis, reducing SNPs from 577 to 497	151
Table 4.3 Results of the MIGRATE-N assessment of migration between species after removing the misidentified sample 083A. Each column indicates the percentile (2.5%, 25.0%, 75.0% or 97.5%), mode, mean or median of the posterior distribution of the parameters θ and M	152

List of Figures

- Fig. 1.1** Proposed process by which hexaploidy arose in *Labeobarbus* as per Oellermann & Skelton (1990) with additional specifics from Yang *et al.* (2015). Each block is a loose representation of a haploid genome (n), with the colour indicating ancestral groups 4
- Fig. 1.2** Six different mouth morphotypes of the KwaZulu-Natal yellowfish, *L. natalensis*, demonstrating the high degree of intraspecific mouth morphotype polymorphism (figure reproduced from Vreven *et al.* 2016, where it was in turn sourced from Crass 1964) 10
- Fig. 1.3** Sampling locations from a previous study analysis utilising mitochondrial control region data to infer six populations of *L. natalensis* across the distribution (Bloomer *et al.* In prep). Populations are colour-coded. The inset highlights places the study region (orange) within South Africa. This map was generated using QGIS (QGIS Development Team, 2016. QGIS Geographic Information System. Open Source Geospatial Foundation Project. <http://www.qgis.org/>) and overlays from the National Freshwater Ecosystem Priority Areas (NFEPA) project (Nel *et al.* 2011) 29
- Fig. 2.1** Estimated number of research articles using RAD sequencing by year, indicating near-exponential growth. Number of research articles was estimated by the number of Google Scholar hits for “Restriction-site associated DNA” 46
- Fig. 2.2** Distribution of samples in this study across KwaZulu-Natal with reference to a map of South Africa (top left). River names are indicated at each sampling site. The colour of each sampling site corresponds to the putative population identified in this study. The shape of the symbol indicates an association to either the northern or the southern lineage. This map was produced using QGIS (QGIS Development Team, 2016. QGIS Geographic Information System. Open Source Geospatial Foundation Project. <http://www.qgis.org/>) and the National Freshwater Ecosystem Priority Areas (NFEPA) project (Nel *et al.* 2011) 48
- Fig. 2.3** Box-and-whisker plot showing the distribution of the mean depth of coverage across all 23 samples (plus two replicates) as the value for the minimum number of reads ($-m$) is varied from 3 to 10, as per Paris *et al.* (2017). Whiskers here indicate the maximum and minimum values across the samples. Paris *et al.* (2017) recommend a depth of coverage of $>25x$, which would indicate that $-m = 5$ is most suitable for this dataset. The default parameter set was $-m 5 -M 3 -n 2$ 51
- Fig. 2.4** STRUCTURE analysis with $K = 2-5$ using 723 high-quality filtered loci. Each individual (indicated as columns along the X-axis) is probabilistically assigned (probability of assignment q on the Y-axis) to one of the inferred genetic clusters. $K = 2$ was recovered as having the most support with the Evanno method (Evanno *et al.* 2005). CLUMPP was used to produce this representation from 20 replicates 57
- Fig. 2.5** Hierarchical STRUCTURE results for the northern and southern lineages. Each individual (indicated as columns along the X-axis) is probabilistically assigned (probability of assignment q on the Y-axis) to one of the inferred genetic clusters. A – primary run of STRUCTURE for $K=2$ on which the hierarchical STRUCTURE was based. B – northern lineage hierarchical STRUCTURE for $K=2$ and $K=3$ showing the results for minor clusters (4/20 replicates and 6/20 replicates respectively) only when location is specified as prior. C – southern lineage hierarchical STRUCTURE for $K=2$ showing the major cluster (14/20 replicates) with location as prior 58
- Fig. 2.6** A – FINERADSTRUCTURE co-ancestry matrix, indicating pairwise co-ancestry between individuals. The lower diagonal indicates raw copy numbers whereas the upper diagonal shows aggregated copy numbers. Individuals clustering into populations are indicated by clustering in the accompanying tree and along the diagonal of the plot. The sample labelled ‘L’ is the individual

from the Lions River, whereas the samples marked 'Mz' are the two individuals from the Mzimkhulu. B – MCMC pairwise comparison heat plot. Pairwise co-ancestry less than 0.6 are not shown (white blocks). Populations are broadly indicated by squares clustering together along the diagonal. Intra-population divergence is observed within the Mkomaas and Mzimkhulu lineages, as shown by pairwise comparisons not in blue (1.0 estimated co-ancestry). This is particularly obvious between the Mkomaas and Mzimkhulu Rivers (within the Mkomaas population) 60

Fig. 2.7 Principal Components Analysis plots indicating the distribution of individuals into populations according to the first three principal components identified in FINERADSTRUCTURE, accounting for 70.8% of the eigenvalues. Component 1 splits Umfolozi from the rest. Component 2 splits Tugela (and by extension Umgeni) from all others. Component 3 isolates Mbokodweni from the other populations. A – The first two components which split samples into all five observed populations. The southern lineage (Mkomaas and Mbokodweni) clusters closely together. B – Components 1 and 3 split the samples into five populations. Subdivision within the five major populations is most apparent. C – Components 2 and 3 split the samples into five populations, although one sample of the Mkomaas population (KNU018) clusters closely with the Umfolozi population 62

Fig. 2.8 Factorial Correspondence Analysis plot showing the clustering of individuals into the five populations according to the first three components, accounting for 25.86% of the variance in the data 63

Fig. 2.9 Venn diagram illustrating the association of alleles between four populations (Mkomaas, Mbokodweni, Tugela and Umfolozi) after samples from the potentially admixed Umgeni population are removed, resulting in 720 diallelic loci. Values in italics indicate private alleles for each of the four populations, whereas the value in bold indicates alleles found in all populations 64

Fig. 2.10 Scenario determined as most likely by DIYABC scenario testing using the logistic regression approach. Divergence times are indicated on the right. Divergence points are not drawn to scale. The node *t0* was fixed in time as the oldest point, whereas all other nodes were allowed to vary in relation to one another 66

Fig. 3.1 Sampling locations across South Africa showing major river systems. Sampling localities in black correspond to *L. natalensis*, localities in grey are *L. aeneus* and *L. kimberleyensis*. This map was produced using QGIS (QGIS Development Team, 2016. QGIS Geographic Information System. Open Source Geospatial Foundation Project. <http://www.qgis.org/>) 99

Fig. 3.2 Haplotype network produced from 77 polymorphic mitochondrial SNPs identified using the STACKS pipeline in *L. natalensis*. Boxes indicate 95% connection limits (three steps). A maximum of nine connecting steps were allowed 107

Fig. 3.3 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 four connecting steps were allowed 108

Fig. 3.4 Haplotype network produced from consensus mitochondrial genome sequences found by mapping read data of 25 *L. natalensis* samples against the *L. intermedius* reference mitogenome. Two replicate pairs are included and indicated by white-bordered circles with shades of purple indicating pairs 110

Fig. 3.5 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 white-bordered circles with shades of purple indicating the pairs 111
- Fig. 3.6** Bayesian constant coalescent chronogram based on partial mitogenome reconstructions for 69 *Labeobarbus* samples (including seven replicate samples indicated by paired purple circles). 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 *L. natalensis* populations match those described in Chapter 2. 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 113
- Fig. 3.7** 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 circles. Nodal probability support values equal to or exceeding 0.95 are indicated by a *. The outgroup was trimmed from the tree following rooting 114
- Fig. 4.1** Sampling localities of *L. aeneus* and *L. kimberleyensis* across South Africa. The number associated with each sampling site indicates how many samples were sequenced from that locality 140
- Fig. 4.2** Mean depth of coverage across all samples (including replicates) as the minimum read depth (-m) was varied as per Paris *et al.* (2017). Whiskers indicate the maximum and minimum values. The default parameter set was -m 5 -M 2 -n 1. The optimal value of -m yielding mean depth of coverage >25x (Paris *et al.* 2017) was found to be -m = 8 143
- Fig. 4.3** Results of various methods of species assignment across the 39 samples in this study. A: Morphological assignment given in the field when the sample was collected. B: Mitogenome assignment based on the results of the mapping approach presented in Chapter 3. C: STRUCTURE plot showing $K = 2$ for all 39 samples across the distribution utilising the final dataset of 577 SNPs. Each individual is probabilistically assigned (probability of assignment q) to one of the two species. This representation was produced using CLUMPP over 20 replicates. D: GENECLASS2 probability of assignment of each sample utilising 577 SNPs and using the morphological assignment as the default population assignment. E: GENECLASS2 probability of assignment using the nuclear genetic STRUCTURE assignment as the default population assignment 149
- Fig. 4.4** Factorial Correspondence Analysis indicating the clustering of samples into species and libraries based on the first three components, accounting for 24.57% of the variance between samples. The two sequencing libraries are indicated. One sample from library 1 with the highest sequencing depth clustered with the samples of library 2 (green arrow) 150

List of Supporting Information

Table S2.1 Sampling locations of *L. natalensis* used in this study. Each of the samples listed yielded DNA of a quality satisfactory for RAD sequencing. Sample numbers match those indicated in the STRUCTURE results 84

Table S2.2 Statistics generated as $-n$ is varied from $-n = -M - 1$ to $-n = -M + 1$ using the script COUNT_FIXED_SNPS.PY and the populations module of STACKS with a SNP sample representation cut-off of 80% ($-r = 0.8$), as per Paris *et al.* (2017). I followed Paris *et al.*'s (2017) recommendation in selecting the parameter set which yielded the greatest number of polymorphic loci ($-n = 0$) 85

Fig. S2.1 SNP density plot for 85-bp Read 1 fragments after initial processing in STACKS using the parameter set $-m\ 5\ -M\ 2\ -n\ 1$. The initial uniform distribution suggests low levels of sequencing error, whereas the slight increase in SNPs towards the end of the loci is indicative of sequencing error. Trimming to a length of 80-bp (indicated by the dashed line) removes most of these errors 86

Fig. S2.2 Quality scores for each position of the 85-base pair Read 1 sequences of every sample used in this study prior to quality filtering or trimming. Quality scores were obtained using FASTQC (Andrews 2010) and are shown using a box-and-whisker plot where the whiskers indicate the 90th and 10th percentiles of the distribution. The quality throughout the positions is high, although this deteriorates towards the end of the reads 87

Fig. S2.3 Quality scores for each position of the 80-base pair Read 1 sequences of every sample used in this study after trimming and quality filtering in PROCESS_RADTAGS. Quality scores were obtained using FASTQC (Andrews 2010) and are shown using a box-and-whisker plot where the whiskers indicate the 90th and 10th percentiles of the distribution. The quality is still somewhat diminished towards the end of the reads, but is improved compared to the results prior to trimming 88

Fig. S2.4 Parameter optimisation of the STACKS parameters $-m$ and $-M$ measured by assessing the total number of assembled loci, polymorphic loci and SNPs identified as the default parameter set ($m5M3n2$) is changed by the variable on the x-axis as per Paris *et al.* (2017). The distribution of results across the 23 samples (plus two replicates) is given by a box-and-whisker distribution where the whiskers indicate maximum and minimum values. The trend for $-M$ approaches a plateau at $-M = 1$ 89

Fig. S2.5 Parameter optimisation of the STACKS parameters $-m$ and $-M$ measured by assessing the total number of assembled loci, polymorphic loci and SNPs identified as the default parameter set ($m5M3n2$) is changed by the variable on the x-axis as per Paris *et al.* (2017). The total values for the entire population (requiring a SNP to be present in at least 80% of samples as per Paris *et al.* (2017)) are indicated. The trend for $-M$ again approaches a plateau at $-M = 1$ 90

Fig. S2.6 All evolutionary pathways tested in DIYABC leading to the current five observed populations. Populations are abbreviated: Mbo = Mbokodweni; Mko = Mkomaas; Umf = Umfolozi; Umg = Umgeni; Tug = Tugela. The scenarios were tested in batches A, B, C, D and E. Scenarios with the most support using the logistic regression approach for each batch are indicated in bold. A: Basic scenarios of a polytomy were tested. B: Simple evolutionary ladder-like splits resulting in the five populations were tested, keeping the first diverged population as one of the two outgroups identified in A. C: A split into two lineages which later diverged into the current five populations was tested. Note that node ages were allowed to vary, resulting in only the ten scenarios tested. D: All of the most supported scenarios from the previous batches were

tested against each other. The fifth scenario was most supported, although scenarios 3 and 7 received very similar levels of support	91
Fig. S2.7 Results of running the HDPLLOT script (McKinney <i>et al.</i> 2017) on the current dataset with no threshold for the number of samples a locus must appear in to be called ($-r = 0$). Due to the sparsity of samples, only a certain number of values are valid for heterozygosity (H). The y-axis indicates deviation of both alleles of a diallelic locus from 1:1. The primary cloud (indicated by the rectangle) indicates non-paralogous loci, with duplicated loci separating above this cloud and diverged duplicates clustering to the right of the plot. Setting thresholds of $H < 0.8$ and $-7 < D < 12$ yielded 463 of 16,893 SNPs which were identified as potentially paralogous loci	92
Fig. S2.8 Distribution of the final set of SNPs across all individuals used in this study due to the requirement for SNPs to be present in at least 60% ($=14$) of all samples	93
Table S3.1 Detailed sampling information for all samples of <i>L. aeneus</i> , <i>L. kimberleyensis</i> and <i>L. natalensis</i> used in this study	127
Table S3.2 <i>In silico</i> restriction digestion of the <i>Labeobarbus intermedius</i> mitogenome with the restriction enzymes <i>Mlu</i> CI and <i>Nla</i> III to predict expected coverage of this approach in South African <i>Labeobarbus</i> . Gap size is calculated by comparing the value of 'Below' for that position to the value of 'Above' for the previous position. If Below is greater, the gap is equal to Below - Above	128
Table S3.3 Final step of parameter optimisation as per Paris <i>et al.</i> (2017) for <i>L. natalensis</i> where statistics are generated for different values of $-n$ as it is varied from the default parameter set of $-m\ 5\ -M\ 3\ -n\ 2$. These statistics are generated using the script COUNT_FIXED_SNPS.PY as well as the populations module of STACKS where SNPs are called only if they occur in at least 80% of samples ($-r\ 0.8$). Although Paris <i>et al.</i> (2017) suggest selecting the dataset with the greatest number of polymorphic loci ($-n = 1$ here), I opted to select a value of $-n = -M = 2$ due to some of the technical aspects of the data (such as prevalence of overlapping paired reads, adapter pollution, and many paralogous loci as a result of polyploidy)	131
Fig. S3.1 Mean depth of coverage across all 23 samples of <i>L. natalensis</i> (plus two replicates) as the minimum number of reads required to call a stack ($-m$) is varied. This methodology is as per Paris <i>et al.</i> (2017). The default parameter set used was $-m\ 5\ -M\ 3\ -n\ 2$. Whiskers here indicate the 10 th and 90 th percentiles	132
Fig. S3.2 Parameter optimisation protocol for <i>L. natalensis</i> as per Paris <i>et al.</i> (2017) showing the number of assembled loci, polymorphic loci, and SNPs as different values are tested for $-m$ and $-M$ from the default parameter set of $-m\ 5\ -M\ 3\ -n\ 2$. The distribution given by the box-and-whisker plots illustrate the results across all 23 samples (plus two replicates), where the whiskers indicate minimum and maximum values. The trend for $-M$ begins to reach a plateau at $-M = 2$	133
Fig. S3.3 Extension of the parameter optimisation protocol by Paris <i>et al.</i> (2017) for <i>L. natalensis</i> where the number of assembled loci, polymorphic loci, and SNPs are recorded at the population level after enforcing a threshold that loci have to occur within at least 80% of the samples in the	

population. Different values of $-m$ and $-M$ were tested in this way using a default parameter set of $-m$ 5 $-M$ 3 $-n$ 2. A plateau was again observed for $-M$ starting at $-M = 2$	134
Table S4.1 Sampling information for <i>L. aeneus</i> and <i>L. kimberleyensis</i> used in this study. Species assignment was based on morphological identification. The division of samples into libraries for sequencing (Orange-Vaal 1 / Orange-Vaal 2) is also indicated. Replicated samples are indicated by <i>rep</i> in parentheses following the sample name. The allocated sample number (#) refers to the number in STRUCTURE plots and other results. All samples were collected by I. Roger Bills (SAIAB)	166
Table S4.2 Optimisation of the STACKS parameter $-n$ using the method of Paris <i>et al.</i> (2017). Parameter sets for $-n = -M$, $-n = -M - 1$ and $-n = -M + 1$ were tested for the number of assembled loci, polymorphic loci and SNPs across all samples as well as at the population level (numbers in bold) using the script COUNT_FIXED_SNPS.PY and POPULATIONS respectively. All replicates were included, and the “80% rule” was specified for either all samples as a single population or within at least a single population when both populations were defined. Numbers in italics indicate the criteria for which the optimal dataset is chosen according to Paris <i>et al.</i> (2017)	167
Table S4.3 Results of a BLAST analysis of two of the outlier loci found to be potentially under selection. These SNPs were found to belong to the same gene, which is composed of multiple copies of a monomeric unit approximately 588 bp long and which appears in multiple sites with varying repeats across Cyprinidae. Due to this and the fact that a loss of genetic structure between species was observed after filtering the four outlier loci, it appears that this gene is important for differentiating between the two Orange-Vaal yellowfish species	168
Fig. S4.1 SNP density spectrum showing the distribution of SNPs across all assembled loci (read 1 fragments only) in STACKS prior to any filtering or trimming using default parameters $-m$ 5 $-M$ 2 $-n$ 1. The first four positions will always record an absence of SNPs as they correspond to the first restriction enzyme. An elevated number of SNPs was recorded at every third position after site 4, which I argue is due to an in-frame genic bias caused by the ‘ATG’ in restriction enzyme recognition sequence (see main text). A uniform density across locus positions is indicative of low levels of sequencing error. To remove sequencing errors, I employed a size selection cut-off of 80 base pairs (magenta line)	169
Fig. S4.2 Quality scores for each position of all reads for each sample within the first Orange-Vaal yellowfish library as per FASTQC (Andrews 2010). Whiskers here indicate 10 th and 90 th percentiles. The quality here was superior to that of <i>L. natalensis</i> (Chapter 2) but inferior to the second library (Fig. S4.3)	170
Fig. S4.3 Quality scores for each position of all reads for each sample within the second Orange-Vaal yellowfish library as per FASTQC (Andrews 2010). Whiskers indicate 10 th and 90 th percentiles. Quality was near-perfect for most reads across all samples, with little variance across sites	171
Fig. S4.4 Parameter optimisation for STACKS parameters $-M$ and $-n$ as per the method of Paris <i>et al.</i> (2017). Numbers of assembled loci, polymorphic loci and SNPs were assessed as values of $-m$ and $-M$ were varied from a default of $-m$ 5 $-M$ 2 $-n$ 1. All samples and replicates were used in this assessment	172
Fig. S4.5 Parameter optimisation for STACKS parameters $-M$ and $-n$ at the population level as per the method of Paris <i>et al.</i> (2017). Numbers of assembled loci, polymorphic loci and SNPs were assessed as values of $-m$ and $-M$ was varied from a default of $-m$ 5 $-M$ 2 $-n$ 1. All samples and replicates were used in this assessment. I used Paris <i>et al.</i> (2017)’s “80% rule” for both samples as a single population (1 pop) or for that cut-off within at least one of two defined populations (2	

pops). The two sets reported similar results, although the latter possessed a comparative abundance of SNPs and loci due to the relaxation of constraints 173

Fig. S4.6 STRUCTURE plots generated for each of the additional datasets tested in this study. Where not specified otherwise, all parameters in both STACKS and STRUCTURE are the same as for the primary analysis in the main study. Each plot was produced using CLUMPP from 10 replicates. A: primary dataset with four F_{ST} outlier loci removed, showing a complete loss of genetic differentiation between the two species. B: Additional dataset using the $-r = 0.5$ cut-off across all samples as a single population (50% of all samples have to possess a SNP for it to be called). C: Dataset from B after F_{ST} outlier loci have been removed. This plot now mistakenly groups samples from the second library together with the *L. kimberleyensis* population, presumably due to a loss of genetic differentiation and library bias (see main text). D: Second additional dataset where the $-r = 0.5$ cut-off is set for at least one of the two populations (50% of the samples within at least one of two populations are required to possess a SNP for it to be called). E: The same dataset as D but with F_{ST} outlier loci removed. Here little difference between results is observed. F: Third additional dataset where only samples from library 2 are run through STACKS with all the same parameters except $-r = 0.8$ in both populations. G: The same dataset as F but after F_{ST} outlier loci have been removed. Again a loss of genetic differentiation is observed 174

Fig. S4.7 MIGRATE-N output indicating the posterior distributions for the parameters mutation-scaled effective population size (θ) and mutation scaled immigration rate (M) between species after removing sample 083A. Population 1 refers to *L. aeneus* whereas population 2 is *L. kimberleyensis*. Associated summary statistics are indicated in Table 4 175

Fig. S4.8 Results of an initial screening for paralogous loci using HDPLLOT (McKinney *et al.* 2017) with the final parameter set prior to filters ($-r = 0$). The parameter H represents the proportion of heterozygotes whereas D is the deviation of alleles at a diallelic locus from a 1:1 ratio. It appears as if the clouds of duplicated loci and non-paralogous loci are directly adjacent to each other, and it therefore appears difficult to correctly isolate paralogous SNPs in this case. The relatively few samples and sequencing coverage undoubtedly contributed towards this result 176

Chapter 1: New tools to solve old challenges: Assessing phylogeography and population genomics with ddRAD sequencing in the hexaploid freshwater fish genus *Labeobarbus*

1.1 The broad placement and importance of Labeobarbus

Freshwater systems cover approximately 0.8% of the Earth's surface, and yet contain an astonishing ~6% of all described species (Dudgeon *et al.* 2006; Abell *et al.* 2008). Globalisation, pollution, and other effects of human occupation have put these systems under severe threat and this makes the species occurring in freshwater ecosystems some of the most vulnerable to extinction (Ricciardi & Rasmussen 1999; Abell *et al.* 2008). Freshwater systems also feature a remarkably high degree of endemism (Abell *et al.* 2008). Surprisingly, even though freshwater systems are species rich and feature high levels of endemism that are under severe threat, little research has gone into conservation initiatives centred on freshwater systems (Abell *et al.* 2008). Difficulties with species concepts and taxonomic classifications complicate research studies about these species, and insufficient sampling has been a major problem in applying conservation strategies to freshwater fish (Abell *et al.* 2008). This is particularly obvious for large families such as Cyprinidae.

The family Cyprinidae is the largest freshwater fish family, with over 3,100 species recognised worldwide (Eschmeyer & Fong 2017), representing approximately 80% of all freshwater fish in temperate zones (Naran *et al.* 2007a). Currently it is believed that this family originated from Asia in the Tertiary Era, later spreading to Europe, Africa and North America (Fouché 2009). This distribution must have taken place through freshwater bodies as the continents developed, since species within this family are stenohaline and cannot tolerate changes in salt concentrations (Durand *et al.* 2002; Fouché 2009).

In southern Africa there are approximately 210-220 indigenous freshwater fish species, of which about 80 are cyprinids (Skelton 2001; Skelton 2010). Relatively little is known about the evolutionary history and divergence of South African endemic freshwater fish (Fouché 2009). One of the most

popular and sought after genera found in South African waters is *Labeobarbus* (Rüppell 1835) or yellowfishes as they are known locally.

1.2 Taxonomic difficulties concerning *Labeobarbus*

Historically, *Labeobarbus* has been lumped in the “monstrous aggregation” of the genus *Barbus* (Myers 1960: 213). *Barbus* grouped together a wide range of fish and consisted of more than 800 species across multiple continents in a wide variety of environs (Berrebi 1995; Berrebi & Valiushok 1998; Borisov *et al.* 2012). A unique aspect of this genus is how it was split into three major lineages according to the different ploidy levels identified (Berrebi 1995). Diploid (2n) barbs have approximately 50 chromosomes, with tetraploids (4n) at 100 and hexaploids (6n) with 150 (Oellermann & Skelton 1990). However, following calls for taxonomic revision of *Barbus* (Durand *et al.* 2002; reviewed in Vreven *et al.* 2016), the African hexaploid lineage was recently grouped in its own genus of *Labeobarbus* (Oellermann & Skelton 1990; Golubtsov & Krysanov 1993; Doadrio, 1994; Berrebi *et al.* 1996; Skelton 2001) which currently includes 125 recognised species (Vreven *et al.* 2016).

In the last few years *Labeobarbus* was added to the new tribe Torini (Karaman 1971), together with *Carasobarbus* (6n), *Neolissochilus* (4n) and *Tor* (4n), among others (Yang *et al.* 2015). The mahseers *Neolissochilus* and *Tor* are distributed across southern and Southeast Asia, with the himris or *Carasobarbus* found in the Middle East and North Africa and *Labeobarbus* restricted to Africa (Yang *et al.* 2015; Vreven *et al.* 2016). An exploration of the taxa contributing towards the allopolyploidy event by cloning paralogs of *RAG1* exon 3 have suggested that an ancestral tetraploid Torini (maternal) hybridised with a diploid *Cyprinion* (paternal), resulting in the current hexaploid lineage (Fig. 1.1) (Yang *et al.* 2015). An additional round of autopolyploidisation may also have occurred after this hybridisation event (Oellermann & Skelton 1990). This event is thought to have

occurred so long ago that the genus has now become functionally diploid through mutation (Naran *et al.* 2007b). The genus *Cyprinion* is distributed primarily across West Asia, which together with the distribution of the members of Torini and the differing ploidy levels of the genera within the tribe suggests that the advent of hexaploidy occurred prior to the divergence of *Carasobarbus* and *Labeobarbus* from the rest of the tribe, i.e. before the tribe extended into Africa (Naran *et al.* 2007b; Yang *et al.* 2015). This has allowed a tentative dating of the advent of hexaploidy to the Tortonian Epoch, Late Miocene (11.6-7.25 mya) due to the presence of a land bridge the lineage may have used to colonise Africa (Tsigenopoulos *et al.* 2010). A recent *Labeobarbus* phylogeny using mitochondrial data appears to support this radiation out of Asia into northern Africa followed by further radiations south (Tsigenopoulos *et al.* 2010). This path of radiation has been suggested previously for other freshwater fish invasions towards southern Africa (Bowmaker *et al.* 1978; Swartz 2005).

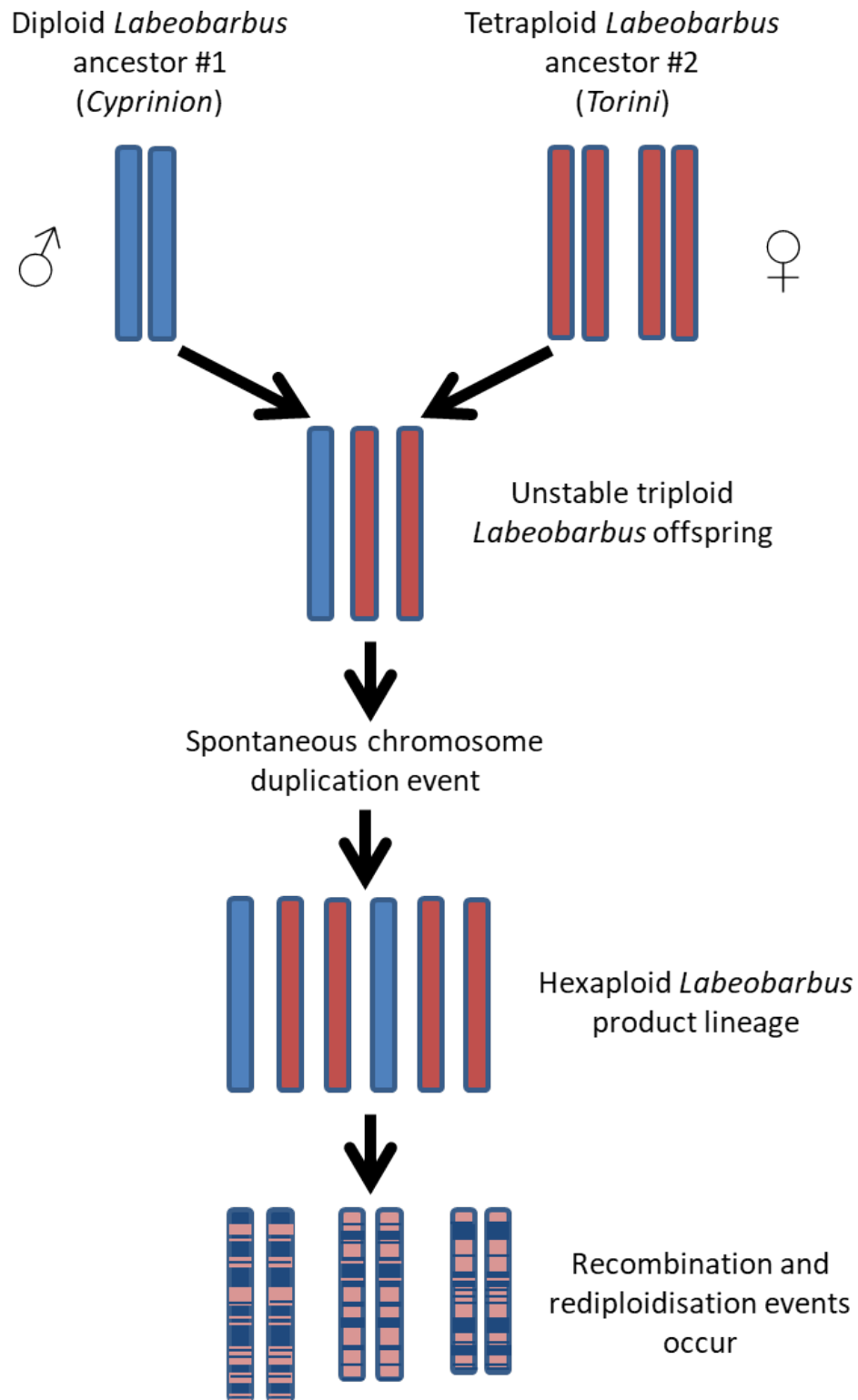


Fig. 1.1 Proposed process by which hexaploidy arose in *Labeobarbus* as per Oellermann & Skelton (1990) with additional specifics from Yang *et al.* (2015). Each block is a loose representation of a haploid genome (n), with the colour indicating ancestral groups.

At a morphological level, studies differentiating large cyprinids have been complicated by morphological plasticity within the family (Machordom & Doadrio 2001; Doadrio *et al.* 2002). This is demonstrated by another key change in the taxonomy of recent years, where the genus *Varicorhinus*, which possessed purportedly distinctive morphological characteristics, was synonymised with *Labeobarbus* (Tsigenopoulos *et al.* 2010; Berrebi *et al.* 2014; Yang *et al.* 2015; Vreven *et al.* 2016). Following several calls to investigate the taxonomy of *Varicorhinus* (Berrebi & Valiushok 1998; Tsigenopoulos *et al.* 2010; Levin 2012) it was found that despite the general occurrence of *Labeobarbus*-like mouth-types (i.e. rubberlips) and *Varicorhinus*-like mouth types (i.e. keratinous with a sharp cutting edge) within the genus to which they were attributed, the discovery of intermediate phenotypes and phenotypes occurring in the “wrong” genus, together with the finding that these phenotypes are not limited to these genera, led to the suggestion that *Labeobarbus* and *Varicorhinus* represented the same genus (reviewed in Vreven *et al.* 2016). In some cases, sympatric species described as *Labeobarbus* and *Varicorhinus* respectively were later discovered to represent different morphotypes of the same (or similar) species, such as *Barbus brucii* and *V. brucii*, which were described by Boulenger (1907) as distinct genera and later found to be synonyms of *B. marequensis* (now *L. marequensis*) (Groenewald 1958; Jubb 1963). This is thought to have occurred due to morphological plasticity and intraspecific phenotype polymorphism within the genus, whereby individuals which develop an extreme scraping mouth typical of the type species for *Varicorhinus* were (mistakenly) grouped accordingly (Vreven *et al.* 2016). This notion was further supported by independent molecular analyses assessing ploidy (Oellermann 1989; Oellermann & Skelton 1990) and mitochondrial paraphyly of the two genera (Tsigenopoulos *et al.* 2010; Levin *et al.* 2013; Yang *et al.* 2015).

Another example demonstrating the extent of confusion morphological plasticity can create, as well as the complication of separating morphological plasticity from genetic divergence is seen in the

Lake Tana (Ethiopia) species-flock of about 15 species (Nagelkerke & Sibbing 2000), formerly lumped together as a single species (*L. intermedius*) with a high degree of plasticity (Banister 1973; Mina *et al.* 1996 a,b; Mina *et al.* 2001). Although some studies have suggested that plasticity may result in the wide range of morphotypes, others argue that this plasticity would not occur at very early stages in the development of fish species (Berrebi & Valiushok 1998; de Graaf *et al.* 2010). Hence early differentiation is indicative of genetic divergence, not plasticity (Berrebi & Valiushok 1998; de Graaf *et al.* 2010). Differentiation at later periods in development may therefore be an indication of plasticity. In the case of the Lake Tana species-flock, it appears that there is sufficient evidence for genetic divergence between at least some of the proposed species (reviewed briefly in de Graaf *et al.* 2010), although this is still contested (Mina *et al.* 2012).

An extension on the theme of morphological plasticity is the occurrence of “tobi koi” or “shoot carp” (Nakamura & Kasahara 1955; Nakamura & Kasahara 1956; Nakamura & Kasahara 1961; Wohlfarth 1977), whereby carp that are grown under certain conditions of stock density or with limited food develop into two distinct morphotypes. This appears to be a specialised form of growth depensation, whereby variability in terms of growth is observed within a population, typically resulting in “large” and “small” subpopulations (Ricker 1958; Magnuson 1962). One of these morphotypes develops at a more rapid rate than the “normal” morphotype, and may cannibalise its peers (Nakamura & Kasahara 1955; Shirahata 1964; Vilizzi & Walker 1999). This could be a potential explanation for some of the intraspecific morphological polymorphism exhibited by some *Labeobarbus* species. Although it has been suggested as a possible explanation for overlapping gut lengths in two sympatric species (Eccles 1986), to my knowledge no evidence of this phenomenon has been documented in this genus.

An alternative hypothesis to that of morphological plasticity is hybridisation, whereby two genetically distinct populations interbreed, possibly resulting in introgression between these populations in one or both directions (Rhymer & Simberloff 1996). Introgression is here defined as

gene flow between populations through hybridisation and backcrossing (Rhymer & Simberloff 1996). Hybridisation reportedly occurs relatively frequently between species within Cyprinidae (Berrebi 1995; Chenuil *et al.* 1999) and more specifically *Labeobarbus* (reviewed in Vreven *et al.* 2016). However, determination of the extent of hybridisation and introgression remains a challenging task and often requires a combination of nuclear and mitochondrial markers for effective assessment (Rhymer & Simberloff 1996). Mitochondrial markers can be useful for indicating the direction of introgression, but as a uniparentally inherited, single recombinational unit have limited potential to detect signals of hybridisation and may not show an accurate representation of introgression due to events such as mitogenome capture (Rhymer & Simberloff 1996; Ballard & Whitlock 2004). In some cases, morphological features can be used to assess the extent of hybridisation (Vreven *et al.* 2016).

A final complicating factor resulting in divergence between morphotypes and ultimately speciation is ecological speciation, whereby divergent selection acts on a particular trait in differing environments, resulting in reproductive isolation (Schluter 2001). It is thought that co-occurring morphotypes may specialize into different niches which are then reinforced by selective pressures, resulting in physical isolation and eventually speciation (Bernatchez *et al.* 1996). The divergence observed between several *Labeobarbus* species which occur in sympatry, such as between *L. aeneus* and *L. kimberleyensis* (Skelton 2001), may be suggestive of this phenomenon. However, identifying species demonstrating ecological speciation is no trivial task (Schluter 2001).

1.3 *Labeobarbus* of South Africa and description of the species used in this study

There are currently seven recognised species of yellowfish found in South Africa (Vreven *et al.* 2016). These are *L. aeneus*, *L. kimberleyensis*, *L. marequensis*, *L. natalensis*, *L. polylepis*, *L. seeberi* and *L. nelspruitensis*, which was until recently placed in *Varicorhinus*. Several other tetraploid and diploid

genera found in South Africa are also occasionally considered together with yellowfish such as the Berg-Breede River whitefish, *Pseudobarbus capensis*, but these will not be discussed in this study. Yellowfish are primarily riverine species, with a preference for the larger rivers as seen by the fact that they appear in all the larger sub-Saharan rivers (Skelton & Bills 2008), however they may also occur in dams. South African yellowfish are migratory, travelling up rivers to spawn in gravel beds. Yellowfish develop slowly and have relatively long life-spans, particularly for the larger species (Mulder 1973; Skelton 2001). This means that young yellowfish are highly vulnerable to introduced predatory species, including several species of American bass, *Micropterus* spp., a major reason for population declines across South Africa (Skelton & Bills 2008).

In terms of conservation status, the IUCN has listed *L. seeberi*, *L. kimberleyensis* and *L. nelspruitensis* as Near-Threatened (Barkhuisen 2017; Impson *et al.* 2017; Roux *et al.* 2017), with the other four species being comparatively widespread and abundant. Despite this, all yellowfish species are in need of effective conservation management as each species faces particular (primarily anthropogenic) threats. The main threats to yellowfish in South Africa are predation by alien fish species, poor water quality due to pollution, over-fishing by subsistence anglers and hybridisation due to poorly considered translocation initiatives (Skelton & Bills 2008). Alien fish species are considered to be the largest threat to freshwater biodiversity (Swartz *et al.* 2008). Southern Africa has been estimated to have over 20 different species of alien freshwater fish invaders (de Moor & Bruton 1988; Skelton 2010).

In South Africa, the Federation of South African Flyfishers (FOSAF) has embarked on an initiative to promote the conservation of all species of yellowfish, and has also actively advertised them as a prized fly-fishing and catch-and-release fish to promote tourism (Fouché 2009). A mention must be made of the Yellowfish Working Group (YWG), a non-profit organisation developed from the 1994 FOSAF conference. The goal of this group is to promote the long-term conservation needs of yellowfish as flagship species for freshwater systems in order to improve the conservation of South

African aquatic systems. Over the years, actions and publicity employed by the YWG for yellowfish have resulted in numerous positive interventions, as well as increasing awareness of this genus and potential problems with South African freshwater systems (Skelton & Bills 2008).

South African species of *Labeobarbus* appear to exhibit substantial morphological plasticity as discussed above, particularly in mouth type in species such as *L. aeneus*, *L. marequensis* and *L. natalensis* (Fig. 1.2) (reviewed in Vreven *et al.* 2016). Large, rubbery lips are used for finding food in pebbles, and thinner, keratinised lips are used to pry food off of rocks (Skelton 2001; Skelton & Bills 2008). Some morphological characteristics such as this may be adaptations to particular microhabitats or feeding niches for the same species (Fouché 2009). For example, the YWG has reported two co-occurring but ecologically distinct species (*L. aeneus* and *L. kimberleyensis*) sharing morphological characteristics, such as mouth shape (YWG 2013). This may be explained by the fact that they both occur in the same general habitats and different microhabitats (deeper waters for *L. kimberleyensis*), but may have, in some instances, been forced to specialise to the same microhabitat (Moyle & Baltz 1985; Fouché 2009). This is exemplified by a simple common-garden experiment by Groenewald (1958), where a rubber-lipped specimen of *L. aeneus* was taken from its natural habitat and raised in a muddy pond, later developing 'normal' or keratinous lips.

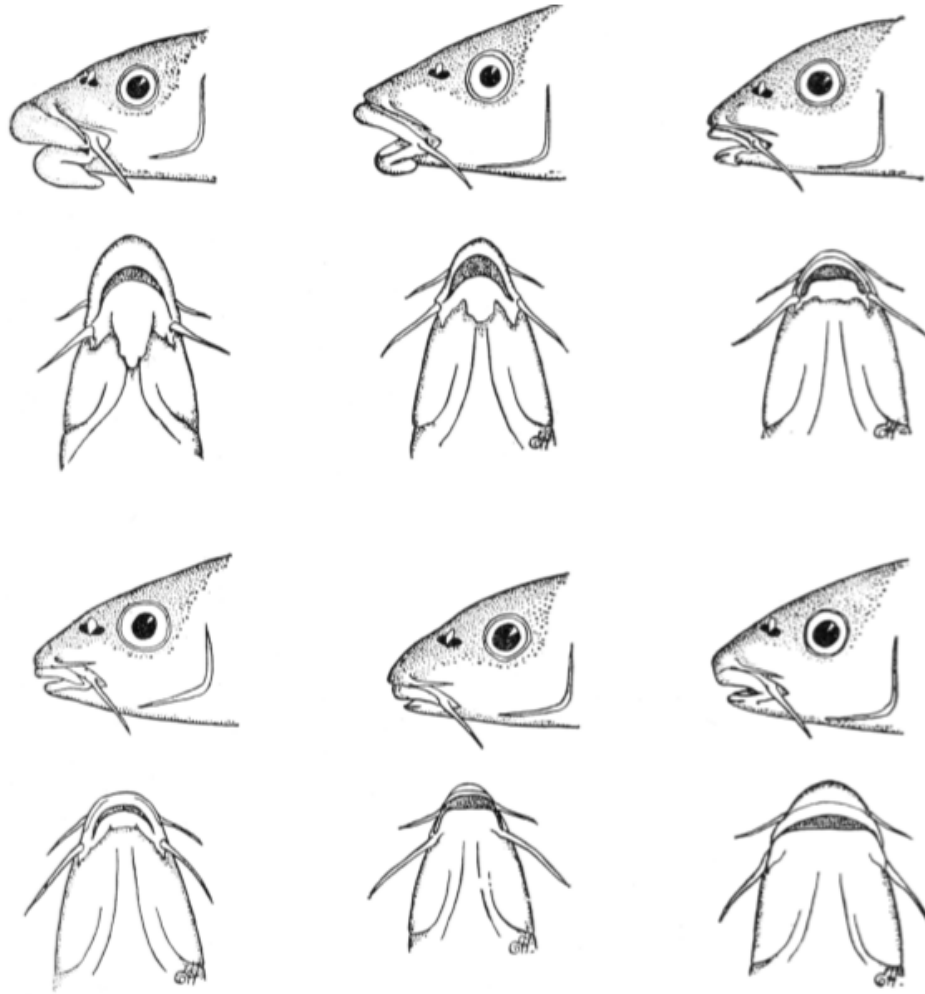


Fig. 1.2 Six different mouth morphotypes of the KwaZulu-Natal yellowfish, *L. natalensis*, demonstrating the high degree of intraspecific mouth morphotype polymorphism (figure reproduced from Vreven *et al.* 2016, where it was in turn sourced from Crass 1964).

1.4 *Labeobarbus natalensis*, the KwaZulu-Natal yellowfish

Labeobarbus natalensis (de Castelnau, 1861), or the KwaZulu-Natal yellowfish is endemic to KwaZulu-Natal, South Africa. The species has played a significant role in human history, and bones of the fish have been found in caves used as shelters by humans in the north Thukela basin dating back 4,400 years (Mazel 1989). The drainage systems in KwaZulu-Natal are ancient, having reached their current configuration approximately 5.1 mya after tectonic uplift in this region (Partridge & Maud 2000; Rivers-Moore *et al.* 2007). This has resulted in deep riverine valleys which are difficult to

access by land and even more difficult for freshwater species to travel between rivers that do not have some form of drainage connection. In addition, a large number of physical barriers such as waterfalls exist in these river systems, making freshwater fish movement difficult or unidirectional. It is thought that this species and four others (*L. polylepis*, *L. seeberi*, *L. aeneus* and *L. kimberleyensis*) arose from a common ancestor that invaded the Orange River basin approximately 2-3 mya (Skelton 2001; Bloomer *et al.* 2007). Due to the difficulty of access to these rivers and the stenohaline nature of this fish, it is difficult to understand how *L. natalensis* has colonised the KwaZulu-Natal rivers so successfully, or what pathways it took to achieve this. This is exemplified by the presence of the species above some waterfalls (such as across the elevated Midlands area of central KwaZulu-Natal) but not others (such as the upper Mzimkhulu River), which is thought to provide evidence of a coastal dispersal between river systems prior to tectonic upheaval giving rise to some of these waterfalls about 1 mya (Crass 1964).

Currently, *L. natalensis* is well distributed throughout the KwaZulu-Natal drainage systems, from the coast up to an altitude of 1500 m and greater (Crass 1964; Karssing 2008). The habitats it occurs in include pools, rapids and the deeper waters of the larger rivers (Crass 1964; Karssing 2008). These fish move downstream during winter in search of warmer water and deeper pools to avoid predation (Crass 1964; Karssing 2008). Adults and juveniles migrate upstream in August/September and can often be seen jumping at this time of the year, attempting to overcome obstacles such as small waterfalls (Crass 1964; Karssing 2008). Most *Labeobarbus* travel and forage in schools, and *L. natalensis* is no exception (Crass 1964; Skelton & Bills 2008). This species is highly polymorphic (as shown by the mouth morphotypes in Fig. 1.2) and this has resulted in a confusing history of at least 12 historical synonyms (Crass 1964; Cambray *et al.* 2017).

Labeobarbus natalensis is relatively abundant in KwaZulu-Natal and has been listed as a species of Least Concern on the IUCN Red Data assessment list for South Africa (Cambray *et al.* 2017). However, the species is thought to be in decline (Karssing 2008). Most of the rivers in KwaZulu-Natal

are in a reasonable ecological state to support the fish, although its largest threats would appear to be the water quality near the Pietermaritzburg/Durban area and predation by alien fish species such as bass (*Micropterus* spp.) and carp (*Cyprinus carpio* Linnaeus, 1758) (Karssing 2008). Another potential threat would be hybridisation with other yellowfish species, namely *L. aeneus*, which was introduced into the Thukela catchment by the Thukela-Vaal Interbasin Water Transfer (IBT) scheme (Karssing 2008). Initial genetic research has shown that the two species are distinct and that *L. natalensis* displays considerable genetic divergence among KwaZulu-Natal river systems - potentially a result of isolation (Bloomer *et al.* 2007; Karssing 2008; Bloomer *et al.* In prep). Additional processes such as philopatry of this riverine species or size-assortative mating as seen in cichlids (Schliewen *et al.* 2001) may have accelerated this process. A further threat, given the apparent divergence between populations of *L. natalensis*, would be hybridisation within the species between different populations due to interbasin water transfers.

A number of protected areas for systematic conservation of freshwater river systems in KwaZulu-Natal are being planned and the YWG has suggested that detailed molecular analyses should be conducted for *L. natalensis*. This is important to determine the genetic and morphological structuring of this species so that those populations with the greatest need of conservation efforts may be identified and included (Karssing 2008). Due to the success of the YWG with yellowfish conservation and awareness together with the popularity of *L. natalensis*, there is also a very real risk that they may be artificially moved beyond their natural range into other areas of South Africa for use as game fish (Karssing 2008). Careful monitoring is necessary to ensure that this does not occur, as *L. natalensis* may then become an alien species which could threaten the diversity in other rivers.

1.5 The Orange-Vaal yellowfishes, *L. aeneus* & *L. kimberleyensis*

Perhaps the most popular endemic fish for freshwater game fishing in South Africa would be the co-occurring Orange-Vaal yellowfishes, *L. aeneus* (Burchell 1822) and *L. kimberleyensis* (Gilchrist & Thompson 1913) (Gaigher 1976; van Vuuren *et al.* 1989; Skelton 2001). The two species are quite different ecologically, behaviourally and biologically. *Labeobarbus aeneus* is an abundant omnivorous, generalist feeder, whereas *L. kimberleyensis* is piscivorous and, as an apex predator within the system, comparatively rare (Mulder 1973; Eccles 1986). Several morphological features distinguish these species, such as colour, eye position and mouth type (Skelton *et al.* 2001). Although it can be difficult to tell the two species apart based solely upon morphological characters (Mulder *et al.* 1990), previous research has found a clear separation based on morphological variation between the two species (Bloomer *et al.* 2007). Additionally, the length of the gut (Kruger & Mulder 1973; Eccles 1986), ideal habitat and range of occupation of the species differ (O'Brien *et al.* 2013). This may be attributed to the difference in ecological role and hence diet of the two species (Mulder 1973; O'Brien *et al.* 2013). Both species develop relatively slowly, although *L. kimberleyensis* is particularly slow with 6-8 years to maturity for males and females respectively (Mulder 1973; Skelton 2001).

Despite all of these differences, there have been a number of reports in recent years of intermediate morphotypes being recorded, particularly in dam environments (Gaigher 1976; Eccles 1986; van Vuuren *et al.* 1989; Mulder *et al.* 1990). It is unclear whether these intermediates are a result of hybridisation or morphological plasticity/developmental changes such as the divergence of morphotypes as seen in “shoot carp” (Eccles 1986). Due to the presence of these intermediates, predominantly in dam or artificial environments, hybridisation may be occurring due to forced habitat sharing as a result of lack of suitable breeding habitat for each species, resulting in anthropogenic breakdown of reproductive isolation (Rhymer & Simberloff 1996; Bloomer *et al.* 2007). A previous study using morphometrics, allozymes and mitochondrial DNA (mtDNA) sequences

for samples across the distribution in natural riverine habitat determined a distinct separation of samples into the different species by morphometrics, but also found “species-diagnostic” allozymes were no longer diagnostic, and allele-sharing of mitochondrial markers (Bloomer *et al.* 2007). These results are indicative of either introgression, ancestral polymorphism between two recently-diverged species, or a developmental aspect to differentiation (Bloomer *et al.* 2007). However, that report also detected significant differences between the upper and lower Orange-Vaal River for *L. aeneus* (Bloomer *et al.* 2007). These regions are separated by Augrabies Falls, which has been shown previously to be a biogeographic barrier separating historically distinct fish faunas (Skelton 1986).

1.6 Conservation and phylogeography in an era of genetics

Conservation efforts for yellowfish species have been viewed as important due to the threats the genus faces as well as its importance to both subsistence and recreational anglers (Skelton & Bills 2008). In addition, yellowfish are a good freshwater health indicator - they require good quality water with relatively few pollutants or alien species (Skelton & Bills 2008). Due to their value as an indicator of water quality and their popularity amongst anglers, yellowfish are being successfully used as “flagship” species for freshwater systems (Skelton & Bills 2008). There is therefore a need to clarify the divergence within endemic *L. natalensis* and to implement conservation methods both for its own continued survival and to ensure that it does not become an alien invader itself. Although the need for conservation intervention may not seem pressing, the fact that *Labeobarbus* is a group of flagship species which is used to educate people about freshwater systems makes it pivotal that conservation plans are put into place as soon as possible as a form of “conservation ethics” (Skelton 2010). Additionally, potential hybridisation of the two Orange-Vaal yellowfishes is worrying and needs assessment. Introgression between the species could potentially lead to extirpation of the rarer *L. kimberleyensis*, depending on directionality of any gene flow (Rhymer & Simberloff 1996).

There is therefore an urgent need to verify whether hybridisation is occurring in the Orange-Vaal system, and to quantify the extent of any introgression between the species, for informed conservation planning. This methodology can then be applied more broadly to other scenarios of potential hybridisation, such as between *L. natalensis* and *L. aeneus* in the north of KwaZulu-Natal or more broadly across other taxa.

Identifying the genetic diversity of populations within a species across its geographical range is paramount for making informed conservation decisions (Palumbi 2003; Reitzel *et al.* 2013). Conservation efforts are often hindered by a lack of accurate data concerning the distributions of biodiversity, information relating to species in a system and the pathways these species may use (Smith & Bermingham 2005). Studies on the biogeography and phylogeography of a species, or group of species, have been used to overcome these gaps in the available knowledge in order to aid conservation efforts (Smith & Bermingham 2005). Biogeographic and phylogeographic analyses are used to identify the impact local and regional processes have on the divergence of a species or a group of species (Smith & Bermingham 2005). This allows for the identification of areas with high species richness, endemism, or other factors which may be informative for the implementation of conservation efforts (Moritz 2002; Luikart *et al.* 2003; Smith & Bermingham 2005).

Freshwater fish serve as excellent models in phylogeographic and biogeographic analyses for a number of reasons – primarily that most, by definition, cannot disperse via salt water (although see Durand *et al.* 2002) and their movement is therefore dependant on the connectivity between bodies of freshwater (Unmack 1999; Durand *et al.* 2002; Smith & Bermingham 2005; Azuma *et al.* 2008). Phylogeographic and biogeographic studies of freshwater fish should therefore be able to identify chronological and geographic trends that match historical events (Bermingham & Martin 1998; Sivasundar *et al.* 2001). Although many studies have used Cyprinidae in phylogeographic analyses (Zardoya & Doadrio 1998, 1999; Tsigenopoulos & Berrebi 2000; Machordom & Doadrio 2001; Durand *et al.* 2002), comparatively few comprehensively sample African taxa (Durand *et al.* 2002).

However, the technique has recently become popular within South African taxa (Swartz *et al.* 2007; Swartz *et al.* 2009; Chakona *et al.* 2013; Swartz *et al.* 2014; Chakona *et al.* 2015; Chakona & Skelton 2017; van der Walt *et al.* 2017).

Phylogeography uses a multidisciplinary approach by combining ecology, population genetics, phylogenetics and historical geography (Avice 2000). Genetic variation is traditionally estimated by using mitochondrial or nuclear DNA molecular markers. The haplotypes observed are used to determine a phylogenetic model representing how a particular lineage may have diverged over time to produce these observed haplotypes (Templeton 2004). Geographical information is then added to this model to generate phylogeographic patterns (Templeton 2004). This approach uses both temporal and spatial elements to obtain accurate information (Avice 2000; Avice 2009). When seeking to determine phylogenetic information, different fields may be utilised depending on the evolutionary scope. At a macro-evolutionary scale where large differences between very different taxonomic groups are being examined, phylogenetic trees would be appropriate (Avice 2009). At a micro-evolutionary level where small differences between very similar taxonomic groups are being examined, such as between related individuals in a family, population genetics and pedigrees may be most appropriate (Avice 2000; Avice 2009). Between the macro- and micro-evolutionary levels, phylogeography may be most appropriate as this field shares aspects from both the macro- and micro-evolutionary scales and may be used flexibly to favour one side of this spectrum more than the other (Avice 2000; Avice 2009). Phylogeographic studies should be conducted over as wide a range as possible for the distribution of a particular species, to enable the assessment of spatial structuring between populations (Sivasundar *et al.* 2001).

Mitochondrial DNA markers are generally favoured as they trace genetic inheritance through the maternal line (a single parent) and have no or rare occurrence of recombination in comparison to nuclear genes (Wilson *et al.* 1985; Birky 2001; Ballard & Whitlock 2004). Mitochondrial DNA also generally mutates at a faster rate than single copy nuclear DNA, which, compounded by the lower

effective population size, provides the potential for a greater amount of variation which would be useful for phylogenetic or phylogeographic analysis (Moritz *et al.* 1987; Igea *et al.* 2010). It has been suggested that sex-biased gene flow and greater frequencies of mitochondrial fixation may cause discrepancies between relationships obtained from mitochondrial and nuclear markers (Beebee 2005; Pinho *et al.* 2009), and between mitochondrial gene trees and the “true” species tree (Maddison 1997; Naylor & Brown 1998; Nichols 2001; Ballard & Whitlock 2004). Empirical evidence shows that if only part of a species range is analysed with mtDNA markers, it is likely that a shallow phylogeny lacking structure will result (Sivasundar *et al.* 2001). Studies using mtDNA markers over large geographic ranges for freshwater fish generally reveal deep phylogenetic divergences between populations (Avice 2000; Sivasundar *et al.* 2001). The number of mtDNA polymorphisms is generally correlated with the effective size of the female population of a species, so that large population sizes may result in high degrees of polymorphism (Sivasundar *et al.* 2001). In addition, due to the lower effective population size in mtDNA markers, there is a greater variance in allele frequencies due to genetic drift (but see Ballard & Whitlock 2004 for a discussion of associated assumptions). Genetic drift would tend to reduce the variation within populations, but could potentially increase the variation between populations, which here may indicate genetic structure between lineages. The increased mutational rates mentioned above may also lead to a greater incidence of fixed mutations within populations. Mitochondrial markers have been used in several phylogeographic studies on southern African fish species (Waters & Cambray 1997; Wishart *et al.* 2006; Swartz *et al.* 2007; Tsigenopoulos *et al.* 2010).

Nuclear markers are also highly utilised in phylogeographic analyses. A major asset nuclear markers boast is that each nuclear locus on a separate chromosome or linkage group may be treated as an independent marker for analysis (Prychitko & Moore 2000), whereas all mtDNA markers are linked (Naylor & Brown 1998; Rokas *et al.* 2003). Unfortunately, there is also a much greater likelihood of recombination events occurring in nuclear genetic markers (Creer 2007).

Recombination events complicate phylogenetic analyses (Hare 2001) as many datasets have insufficient power to detect recombination events (Schierup & Hein 2000). Recombination events suggest that more than one phylogenetic tree is needed to reflect the relationship between the different segments of the DNA sequence; i.e. these segments have different phylogenetic histories. Methods which take recombination into account often do not capture data which may be found in more basic analyses, such as pairwise distances (Schierup & Hein 2000). Some authors argue that recombination events also provide further information, which may be used to trace phylogenetic history (Schierup & Hein 2000; Creer 2007) and failing to account for recombination may also introduce significant artefacts to an analysis (Schierup & Hein 2000). Indels and heterozygous introns with differing lengths may also provide additional information for phylogenetic analysis (Creer 2007).

Exons, which are the coding regions of DNA, are relatively conserved and have a low mutational rate when compared with mtDNA or non-coding regions of the nuclear genome (Palumbi 1995; Ballard & Whitlock 2004). This makes them ideal markers for identifying ancient divergences between lineages and piecing together deeper branching events in phylogenetic trees. This also means that they are not optimal for separating very closely related species which recently diverged or for phylogeographic analysis, as they often do not accumulate enough mutational differences to provide resolution at these taxonomic levels (Chojnowski *et al.* 2008; Igea *et al.* 2010).

Introns (non-coding DNA found between exons) mutate at a much faster rate than the highly-conserved exonic regions and are therefore more variable (Chojnowski *et al.* 2008). This rate of mutation is still slower than that of mtDNA – the average intron is estimated to mutate at roughly a quarter of the average rate of the mtDNA in animals (Prychitko & Moore 1997). Hence, introns are more useful than exons for providing phylogenetic data for more closely related species or at a phylogeographic level. Gene trees recovered using intron molecular data have been found to be highly concordant with species trees (Creer 2007). Introns are rarely used in fish studies, with most

studies using universal primers which amplify across a wide number of species (Li *et al.* 2010) (but see studies such as Chow & Hazama 1998; Quattro *et al.* 2001; Hassan *et al.* 2002).

A complete phylogeographic analysis would likely use a combination of different types of molecular markers providing resolution at different levels (Prychitko & Moore 1997; Bermingham & Moritz 1998; Creer 2007; Brito & Edwards 2009; Li *et al.* 2010). Exon markers could be used to identify the more ancient divergences between lineages, intron markers could trace more recent divergences between taxa, and finally mtDNA would provide an independent dataset for more recent divergences. In addition to this, multiple nuclear markers could be used concurrently in the analysis to improve the resolution, provided that each marker is independent (Sang 2002; Rokas *et al.* 2003; Willows-Munro *et al.* 2005; Brito & Edwards 2009; Li *et al.* 2010; Curto *et al.* 2012). The ultimate aim of this form of combined approach would be congruence between datasets, indicating the recovery of an accurately resolved phylogeny (Creer 2007; Brito & Edwards 2009; Li *et al.* 2010).

In the pilot study preceding this study, my research team attempted this approach by designing a panel of nuclear and mitochondrial markers for optimisation and analysis in *L. natalensis*, with the intention of combining this dataset with previously generated mitochondrial control region sequences to infer phylogeography of the species. Because the results of this pilot study do not form part of the primary research of this thesis, I expand on the findings of this study briefly here. I assessed the use of candidate loci for microsatellites, exons, introns and a mitochondrial locus. Twelve sets of microsatellite markers were designed using sequence data obtained after creating repeat-enriched libraries with the FIASCO protocol (Zane *et al.* 2002). I also tested 19 additional microsatellite markers designed for acanthopterygian fish species (Gotoh *et al.* 2013). Of the microsatellites, only two of the custom markers were found to be polymorphic (Table 1.1). I selected 44 intron and three exon markers for optimisation on twelve representative samples across the distribution, but found that only three introns and one exon amplified successfully (Table 1.2). An

additional eight microsatellites, four introns and one exon showed promising results but required further optimisation.

Table 1.1 Species-specific microsatellite loci identified from repeat-enriched libraries for *Labeobarbus natalensis*.

Locus	Dye	Primer Sequence (5'-3')	Repeat Motif	Allele Range (bp)	Annealing Temperature
<i>LabeoMS01</i>	PET	F: CCGCTCAGATCATCAGCT T R: GTGGTCCAGCACAGTTCA GCT	(CA) ₁₂	248-300	64°C
<i>LabeoMS02</i>	6-FAM	F: AGAGAGCTGGATCTGAGCTAC R: GAGATGAGATCAGTGATAGT	(CA) ₉	273-295	63°C

Table 1.2 Intron and exon markers tested which yielded useable sequences in the pilot study. The target size indicated is based on closely-related genera in Cyprinidae.

Marker	Type	Primer Sequence (5'-3')	Source	Target size	Annealing Temperature
<i>S7</i>	Intron 1	S7RPEX1F: TGGCCTCTTCCTTGGCCGTC S7RPEX2R: AACTCGTCTGGCTTTTCGCC	(Chow & Hazama 1998)	806-841 bp (Xuzhen <i>et al.</i> 2002)	62°C
<i>12775 (slc7a14a)</i>	Intron 4	F: AGAACTCACTGTGGGCTTCAACAACG R: GGATGGAGTGTGTTGGGACTCTTGG	P. Bloomer	Unknown	48°C
<i>Am2B3</i>	Intron 3	F: TGGAAACCGAAACATTGTGAAC R: CCCATCCAGTCATTCTGATCC	(Hassan <i>et al.</i> 2002)	Unknown	58°C
<i>RAG1</i>	Exon 3	RAG1Tel_F: GCTTTCACCAAGTTTGARTGGCAGCC RAG1Tel_R: TGGCGCGCCACACAGGCTTCATYTG	T. Hoareau	Unknown	54°C

Both the microsatellite and exon/intron analyses were complicated greatly by hexaploidy - one of the polymorphic microsatellites identified gave genotypes typical of a tetraploid locus, whereas intron *S7* yielded at least two paralogous sequences in the amplification product and the exon marker *RAG1* produced sequences with a lot of background noise which may have also contained a paralog. This phenomenon is expected due to the allopolyploid origins of hexaploidy in this genus and means that unless paralog-specific primers are designed, universal primers will amplify a

mixture of all gene copies (Yang *et al.* 2015). This suggests that even though functional diploidisation may be underway in *Labeobarbus* (Naran *et al.* 2007b), a large proportion of the genome appears to retain a high level of sequence similarity. After cloning to confirm the presence of a paralog in S7 together with difficulty experienced in successfully amplifying, sequencing and then analysing these complicated data, I determined that another approach to obtain nuclear markers would be preferable. Useable nuclear sequence data could have been obtained for the markers optimised; however it likely would have required extensive cloning at the population level which is prohibitively expensive. It should be noted that amplification of the mitochondrial gene *cyt b* was successful across most samples.

1.7 Conservation and phylogeography in an era of genomics: The rise of population genomics

A new option for genetic studies is next-generation sequencing (NGS). A number of new technologies have emerged (e.g. Roche 454, Illumina GA and ABI SOLiD) which can generate several fold more data than traditional “first-generation” Sanger sequencing, and which are relatively affordable (Schuster 2007; von Bubnoff 2008; Harismendy *et al.* 2009; Metzker 2009). In addition to the sequencing technologies themselves, a number of different NGS methods have been developed to aid the development of genome-wide genetic markers. Several of these approaches are useable on species for which no reference genome is available – a great advantage over previous challenges in developing nuclear and mitochondrial markers (Davey *et al.* 2011; Hohenlohe *et al.* 2011). The sheer number of markers which may be identified by this process and screened in populations greatly increases the accuracy of phylogenetic inferences which can be made. These methods also combine marker discovery and genotyping into a single step, streamlining the process. Genome-wide marker analysis is highly useful for conservation studies as it has very powerful screening properties for identifying hybrids (Hohenlohe *et al.* 2013).

Many of these NGS methods make use of restriction enzymes to reduce the representation of the genome under investigation (Davey *et al.* 2011). Restriction enzyme-based methods are thought to provide a relatively unbiased set of markers across the entire genome (Davey *et al.* 2011). The wide number of different restriction enzymes may be used to customize a particular project – by selecting a different enzyme, researchers can choose to work more on a particular area in the genome such as biasing for GC-rich areas (Amores *et al.* 2011; Arnold *et al.* 2013; DaCosta & Sorenson 2014; Gonen *et al.* 2015), or avoiding highly repetitive regions of the genome (Davey *et al.* 2011). The selection of a restriction enzyme is dependent on the number of fragments desired, and, to some extent, the species under study (Davey *et al.* 2011; Hohenlohe *et al.* 2011; Peterson *et al.* 2012).

Genome-wide marker development methods use the same general approach: multiple DNA samples are digested with the selected restriction enzyme/s; the resulting fragments undergo a selective amplification process; and sequencing of the selected fragments occurs using a form of NGS (Davey *et al.* 2011). Variants identified through this method are considered to be potential Single Nucleotide Polymorphism (SNP) markers (Davey *et al.* 2011). Depending on the downstream analysis programs, it may be important to select only one SNP per fragment to reduce the complications of linkage (Reitzel *et al.* 2013). Diallelic SNPs are often used to reduce the complexity of an analysis (Reitzel *et al.* 2013). Sample DNA used in genome-wide marker development must be of a high quality and high yield, free of contamination of RNA, protein, or other DNA (Davey *et al.* 2011). Ideally, the same amount of DNA must be presented for each individual in these analyses, so that markers are not lost due to uneven coverage between samples (Davey *et al.* 2011). A rough coverage of 8-10x has been suggested for maximum-likelihood analyses (Hohenlohe *et al.* 2011), although recent papers suggest a bare minimum of 10x coverage with a preferred minimum of 20-25x (Paris *et al.* 2017; Rochette & Catchen 2017).

No reference genome is available for *L. natalensis*, or, indeed, for any *Labeobarbus*. I therefore opted to focus on one particular reduced-representation NGS approach using restriction enzymes, Restriction-site Associated DNA (RAD) sequencing (Miller *et al.* 2007; Baird *et al.* 2008), as a NGS methodology for this study. This method reduces the genomic data targeted by NGS so that high coverage can be obtained for selected markers (Davey *et al.* 2011). RAD sequencing has been used extensively in the literature for a wide variety of applications, particularly in fish. These include: to identify population divergence (Larson *et al.* 2014; Boehm *et al.* 2015; Ferchaud & Hansen 2016), QTL and linkage mapping (Houston *et al.* 2012; Gagnaire *et al.* 2013; Brieuc *et al.* 2014; Henning *et al.* 2014), assessments of hybridisation (Lamer *et al.* 2014; Pujolar *et al.* 2014; Hand *et al.* 2015), investigating genome evolution (Brawand *et al.* 2014; Kai *et al.* 2014; Waples *et al.* 2016), phylogeography (Reitzel *et al.* 2013; Macher *et al.* 2015) and, more recently, phylogenetics (Wagner *et al.* 2013; Gonen *et al.* 2015). The high power of this form of analysis may be particularly useful in phylogeographic studies of a given species across its distribution, as this should identify trends in the data that may otherwise be missed in an intraspecific analysis (Davey *et al.* 2011). An example of this is found in Reitzel *et al.* (2013) where traditional genetic markers such as microsatellites were inconclusive for providing phylogeographic information for an anemone but NGS methods were able to produce highly resolved genetic structure of these populations.

RAD sequencing starts with restriction digestion of sample DNA (1-2 ng per sample) (Davey *et al.* 2011). Adaptors are added to mark the fragments belonging to different individuals, and pooled together (Davey *et al.* 2011). The fragments are sheared, size selected, and sequenced, resulting in high coverage for these restriction sites (Davey *et al.* 2011). Paired-end sequencing is popularly used in the literature and provides further information which may be used to identify SNPs or anchors for markers in tandem with previously identified genome data (Davey *et al.* 2011). Although the procedures involved for these different methods are relatively simple, they may produce highly variable results between species, and the time required for some aspects of a method may be

demanding (Davey *et al.* 2011). Several attempts may be needed to obtain a functional library, which may be prepared by a company instead (Davey *et al.* 2011). This procedure produces so many putative loci that a number of highly rigorous screening conditions should be put in place to optimise the SNPs produced (Hohenlohe *et al.* 2011).

RAD sequencing has a number of potential issues which have been identified. In particular restriction fragments which are too short or too long are not optimal for Illumina sequencing, causing a large number of fragments to be discarded and resulting in decreased read depth (Davey *et al.* 2013). PCR GC-bias may also affect the read depth, and has been recorded in other sequencing methods (Davey *et al.* 2013).

Several modified versions of the RAD sequencing protocol exist, including double-digest RAD (ddRAD) sequencing (Peterson *et al.* 2012). This modification of RAD sequencing greatly reduces random shearing events and provides more fine-scale selection for specific genomic fragments (Peterson *et al.* 2012). Double-digest RAD sequencing is intended to provide greater precision and repeatability (Peterson *et al.* 2012). This method is less expensive and time-intensive than traditional RAD sequencing (Peterson *et al.* 2012). Two digestion enzymes are used instead of one – the same common-cutting enzyme normally applied in traditional RAD sequencing as well as a rare-cutting enzyme (Peterson *et al.* 2012). In combination, this results in fewer but well-represented genomic intervals which are ideally returned in similar sequence read proportions from different individuals (Peterson *et al.* 2012). However, the inherent bias introduced in RAD sequencing may be even more pronounced in ddRAD (Arnold *et al.* 2013; DaCosta & Sorenson 2014). An allele frequency estimation bias is introduced through RAD sequencing by mutations occurring in the restriction sites resulting in certain reads not being retained in some individuals, leading to incorrect population inferences (Arnold *et al.* 2013). This bias is compounded in ddRAD sequencing due to the requirement for two restriction sites, and the possibility that mutations in novel restriction sites may also interfere during the size selection process (Arnold *et al.* 2013). Indels of sufficient length may also eliminate alleles

from the size-selection range in ddRAD (DaCosta & Sorenson 2014). In addition, a genealogical bias for missing haplotypes (Arnold *et al.* 2013) and overrepresentation for particular genomic regions specific to the restriction enzymes used, particularly for ddRAD sequencing (DaCosta & Sorenson 2014) have been identified. These caveats which result in differences in read depth across loci are often combatted by removing loci of low coverage from the analysis (Andrews *et al.* 2016).

A number of genetic analysis tools which can simulate RAD markers *in silico* have been developed (e.g. Lepais & Weir 2014). A good first step in planning a RAD sequencing project is to digest the genome of your organism, if available, or that of a closely related organism, to determine the cut site frequency of potential enzymes. This allows tailoring of the expected marker count and coverage for the particular study. Unfortunately, this first step was not possible in this project as no reference genome is available for any of the hexaploid members of the tribe. Analyses generally involve performing quality control at various stages throughout a pipeline, and then examining the data for biological relationships (Davey *et al.* 2013). Examples of these analysis programs include STACKS (Catchen *et al.* 2011; Catchen *et al.* 2013), RAPID (Willing *et al.* 2011), RADTOOLS (Baxter *et al.* 2011), PYRAD (Eaton 2014) and dDOCENT (Puritz *et al.* 2015). Of the various programs that are available, STACKS is arguably the most popular, and has been supplemented with two detailed workflows to the SNP identification process (Paris *et al.* 2017; Rochette & Catchen 2017). STACKS can even detect paralogous sequences that have resulted from genome duplication events, and flag them for removal from the analysis (Catchen *et al.* 2011). This is done by calculating the degree of heterozygosity observed and comparing this to expected levels, as well as by identifying more than two alleles for a particular locus (Catchen *et al.* 2011).

Genetic analyses for *Labeobarbus* are complicated because of polyploidy. All *Labeobarbus* are hexaploid (as discussed above) and most genetic analyses are designed for use in diploid organisms. The pilot study previously showed that the hexaploid genome often results in difficulties resolving genotypes using microsatellites and other traditional nuclear markers, and I predicted that this

would extend to the analysis of SNPs as well. The issue of polyploidy appears to be relatively seldom addressed in the literature, with a large body of papers appearing to prefer analysing markers with an easier mode of transmission or expression than nuclear markers, such as the mitochondrial genome (which would still be single copy) or the transcriptome (Everett *et al.* 2011). Selection of either of these markers would effectively remove many of the complications of working with a polyploid. Several studies have tackled this problem, particularly in tetraploid fish, and multiple strategies appear to be effective. One strategy used for trout is the removal of all markers which have an observed heterozygosity greater than 0.5 (Hohenlohe *et al.* 2011; Hohenlohe *et al.* 2013). This strategy has been widely adopted in the literature, even outside of studies on polyploids, due to its utility in identifying potential paralogs. Additional approaches which have been utilised include haploid genotyping (Sanchez *et al.* 2009; Hecht *et al.* 2013; Limborg *et al.* 2014; Larson *et al.* 2016; Waples *et al.* 2016), comparing sequencing depths (Dou *et al.* 2012; Lighten *et al.* 2014), and others (reviewed in McKinney *et al.* 2017). In addition, the new HDPLLOT approach uses a combination of assessing heterozygosity levels together with depth of allelic ratios to differentiate both recent and more ancient duplicates from non-duplicated loci (McKinney *et al.* 2017).

The research field of population genomics has been empowered by recent NGS technologies and methodologies resulting in quick and easy access to thousands of markers across a target genome. This discipline uses genomics to obtain a large number of markers across a genome in an attempt to understand evolutionary processes resulting in variation within and between populations in a manner similar to population genetics (Black *et al.* 2001; Luikart *et al.* 2003). An important distinction of population genomics is that it allows for differentiation of locus-specific and genome-wide effects, which is critical in separating different kinds of evolutionary processes and especially for gaining insight into microevolution (Black *et al.* 2001; Luikart *et al.* 2003). Population genomics assumes that neutral loci across the genome will be subjected to the same evolutionary forces and demography, whereas loci under selection will often react in a different way and will therefore be

identified as “outlier” loci explaining additional variation (Luikart *et al.* 2003). This means that identifying outlier loci is paramount in population genomics approaches, both for identifying the neutral demographic history and to investigate potentially adaptive loci (Luikart *et al.* 2003). Population genomics approaches therefore comprise four primary steps: 1) sampling across many individuals in each relevant population; 2) genotyping these samples across a large number of loci, typically using NGS technologies; 3) performing statistical tests to determine outliers which may potentially be under selection; 4) calculate evolutionary and demographic parameters after removing outliers, and investigating outliers independently (Luikart *et al.* 2003).

1.8 Context of this study

Initial phylogeographic studies on *Labeobarbus* using the mitochondrial control region have revealed that there are substantial differences between *L. natalensis* populations throughout its distribution (Bloomer *et al.* 2007; Bloomer *et al.* In prep). The *L. natalensis* lineage was found to cluster into six primary mitochondrial haplotypes representing different major river systems – the Umfolozi, Tugela, Umgeni, Mbokodweni, Mkomaas and Mzimkhulu systems (Fig. 1.3) (Bloomer *et al.* In prep). The findings indicate isolation and limited gene flow within as well as between drainage systems (Bloomer *et al.* In prep). Because this study was conducted using only the mitochondrial control region, the question remains whether this pattern would be supported by other markers, particularly the nuclear genome. Morphological and genetic diversity within and among populations could be explained by morphological plasticity, developmental divergence such as “shoot carp”, hybridisation, local specialisation to different niches, assortative mating, selection, migration (primarily unidirectional due to the presence of barriers), or genetic drift over time. These processes need to be investigated to determine which are at work here. In this study I extend the work conducted in the earlier study (Bloomer *et al.* 2007; Bloomer *et al.* In prep) by intensive sampling of

L. natalensis throughout its distributional range, using the NGS technique of ddRAD sequencing and performing a range of genetic analyses in the population genomics framework to determine the phylogeographic patterns and processes affecting these patterns for *L. natalensis* within KwaZulu-Natal.

Additionally, several previous studies (van Vuuren *et al.* 1989; Bloomer & Naran 2006; Bloomer *et al.* 2007) have investigated reports of hybridisation occurring between the ecologically and biologically divergent species *L. aeneus* and *L. kimberleyensis* across South Africa. As discussed above, these studies were not able to categorically determine whether hybridisation and introgression were occurring between these species, potentially due to complications caused by very recent introgression, or by ancestral polymorphisms between these recently-diverged species. Furthermore, some studies have even questioned the validity of these species due to similarities between them (Mulder *et al.* 1990). I therefore aimed to investigate the genetic divergence between these species, as well as whether hybridisation or introgression is occurring between these species by using SNPs produced from ddRAD sequencing. The power of this genomics approach should allow me to differentiate between the competing hypotheses explaining the previous genetic results.

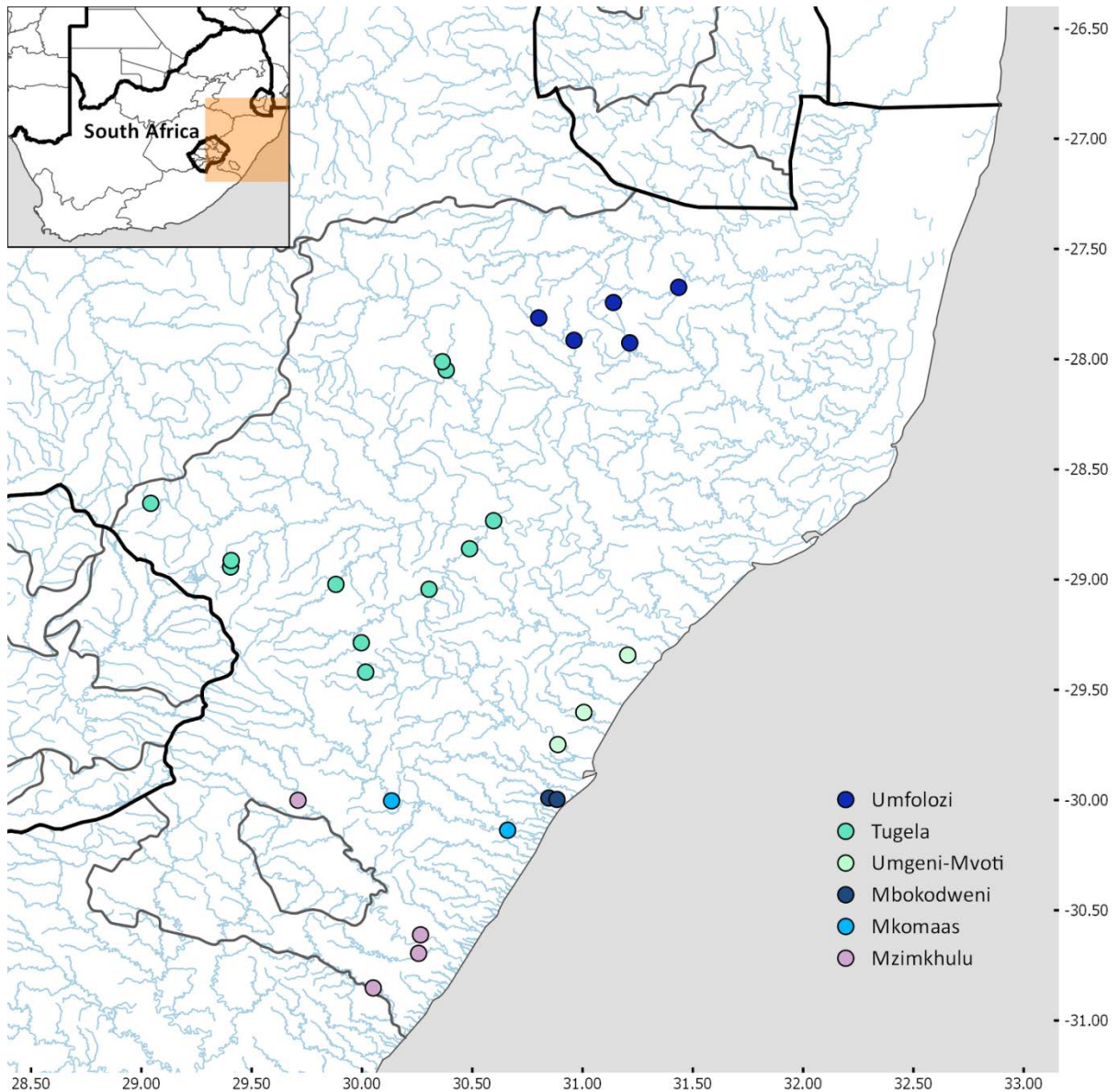


Fig. 1.3 Sampling locations from a previous study using mitochondrial control region data to infer six populations of *L. natalensis* across the distribution (Bloomer *et al.* In prep). Populations are colour-coded. The inset highlights the study region (orange) within South Africa. This map was generated using QGIS (QGIS Development Team, 2016. QGIS Geographic Information System. Open Source Geospatial Foundation Project. <http://www.qgis.org/>) and overlays from the National Freshwater Ecosystem Priority Areas (NFEPA) project (Nel *et al.* 2011).

While optimising and exploring the data produced from the ddRAD sequencing protocol I was fortunate to encounter a number of unexpected technical aspects, one of which led to the third chapter in this thesis where I explore the utility of RAD sequencing data for data mining and

construction of mitogenomes. I utilise these partial mitogenome reconstructions to critically examine the phylogeny of the study species and the relation between the identified populations, as well as revealing signals of intra- and interspecific hybridisation. I believe that this and several of the other novel insights and innovations together with the ability to adequately answer my research questions fully justifies the use of these approaches, despite the difficulties involved in pioneering research utilising NGS technologies.

1.9 Research questions

The major research question for this study therefore was:

- What are the phylogeographic patterns and processes acting on these three species of South African yellowfish?

Which was further refined to the following research questions:

1. How many populations of *L. natalensis* occur and how are they delimited?
2. Can mitochondrial genetic signal be isolated from NGS RAD sequencing data for independent analysis of the research questions?
3. Are *L. aeneus* and *L. kimberleyensis* genetically distinct?
4. Is hybridisation and/or introgression occurring between and within these species?

1.10 Hypotheses and predictions

1. The study on the phylogeography of *L. natalensis* was extensive but primarily focussed on assessing population structure and delimitation using ddRAD sequencing markers in a

population genomics framework (Chapter 2). This was supplemented with mitogenome markers mined from the ddRAD sequencing dataset and used to further assess the mitochondrial lineages of *L. natalensis* and place them within a phylogenetic framework (Chapter 3).

Hypothesis 1: Different drainage basins where *L. natalensis* occurs represent isolated populations with limited gene flow.

Prediction 1: Previous research using the mitochondrial control region marker identified six lineages by drainage basin: the Umfolozi, Tugela, Umgeni, Mbokodweni, Mkomaas and Mzimkhulu (Bloomer *et al.* 2007; Bloomer *et al.* In prep). I therefore predict that there are likely to be six populations of *L. natalensis* across the distribution, separated into these drainage basins.

2. This aim arose after investigating the initial ddRAD sequencing datasets. Mining for mitochondrial markers represents a valuable new use for existing RAD sequencing datasets, and was explored in Chapter 3.

Hypothesis 2: Mitochondrial markers can be mined from RAD sequencing datasets to answer biological questions.

Prediction 2: Despite mitochondrial reads usually being treated as noise to be filtered from an analysis, I predict that they can be isolated and extracted using bioinformatics techniques.

3. Genetic distinctiveness of the sympatric Orange-Vaal yellowfish was investigated in the final research chapter on hybridisation and introgression occurring between these species (Chapter 4). This was supplemented by mitochondrial haplotypes acquired from data mining of the ddRAD sequencing dataset (Chapter 3).

Hypothesis 3: The Orange-Vaal yellowfish are genetically distinct species.

Prediction 3: The Orange-Vaal yellowfish *L. aeneus* and *L. kimberleyensis* form distinct reproductive communities with separate gene pools. Their behavioural, ecological, and morphological divergence suggests this, as do some previous initial genetic results using mtDNA (Bloomer & Naran 2006; Bloomer *et al.* 2007) and allozymes (Mulder *et al.* 1990).

4. Intraspecific hybridisation and introgression occurring within *L. natalensis* was assessed in the phylogeographic analysis using nuclear markers (Chapter 2) and supplemented by the mitogenome mining analysis (Chapter 3). Interspecific hybridisation and introgression between *L. aeneus* and *L. kimberleyensis* was assessed using nuclear SNPs in the final research chapter (Chapter 4). There was no *a priori* reason to exclude interspecific hybridisation between *L. natalensis* and *L. aeneus*, however this was not a priority of this study. My analyses may however be informative of this because the samples of *L. natalensis* are taken in an area impacted by river transfer schemes and also in other diverse areas not impacted by these transfers. The phylogenetic framework of the mitogenome chapter (Chapter 3) may allow for a brief comment on this, but it is not a research focus and would benefit from future research.

Hypothesis 4a: No intraspecific hybridisation or introgression is anticipated in *L. natalensis*.

Prediction 4a: The different drainage systems of KwaZulu-Natal are envisioned to be isolated with little opportunity for gene flow.

Hypothesis 4b: Interspecific hybridisation occurs between *L. aeneus* and *L. kimberleyensis*.

Prediction 4b: Evidence of gene flow between *L. aeneus* and *L. kimberleyensis* is expected.

1.11 References

- Abell R, Thieme ML, Revenga C, *et al.* (2008) Freshwater ecoregions of the world: A new map of biogeographic units for freshwater biodiversity conservation. *BioScience* **58**, 403-414.
- Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics* **188**, 799-808.
- Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* **17**, 81-92.
- Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology* **22**, 3179-3190.
- Avice JC (2000) *Phylogeography: The History and Formation of Species*. Harvard University Press.
- Avice JC (2009) Phylogeography: retrospect and prospect. *Journal of Biogeography* **36**, 3-15.
- Azuma Y, Kumazawa Y, Miya M, Mabuchi K, Nishida M (2008) Mitogenomic evaluation of the historical biogeography of cichlids toward reliable dating of teleostean divergences. *BMC Evolutionary Biology* **8**, 215.
- Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376.
- Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology* **13**, 729-744.
- Banister K (1973) A revision of the large *Barbus* (Pisces, Cyprinidae) of east and central Africa. Studies on African Cyprinidae. Part II. *Bulletin of the British Museum (Natural History), Zoology* **26**, 1-148.
- Barkhuizen LM (2017) *Labeobarbus kimberleyensis*. The IUCN Red List of Threatened Species 2017: e.T63292A100166441. Downloaded on **07 December 2017**. <<http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T63292A100166441.en>>
- Baxter SW, Davey JW, Johnston JS, *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* **6**, 1-11.
- Beebe TJ (2005) Conservation genetics of amphibians. *Heredity* **95**, 423-427.
- Bermingham E, Martin AP (1998) Comparative mtDNA phylogeography of neotropical freshwater fishes: testing shared history to infer the evolutionary landscape of lower Central America. *Molecular Ecology* **7**, 499-517.
- Bermingham E, Moritz C (1998) Phylogeography of endemic toads and post-Pliocene persistence of the Brazilian Atlantic Forest. *Molecular Ecology* **7**, 367-369.
- Bernatchez L, Vuorinen JA, Bodaly RA, Dodson JJ (1996) Genetic evidence for reproductive isolation and multiple origins of sympatric trophic ecotypes of whitefish (*Coregonus*). *Evolution* **50**, 624-635.
- Berrebi P (1995) Speciation of the genus *Barbus* in the north Mediterranean basin: Recent advances from biochemical genetics. *Biological Conservation* **72**, 237-249.
- Berrebi P, Chenuil A, Kotlík P, Machordom A, Tsigenopoulos CS (2014) Disentangling the evolutionary history of the genus *Barbus sensu lato*, a twenty years adventure. In: *Professor Carlos Almaça (1934–2010)–Estado da arte em áreas científicas do seu interesse*. (eds. Alves M, Cartaxana A, Correia A, Lopes L), pp. 29-55. Museu Nacional de História Natural e da Ciência, Lisboa.
- Berrebi P, Kottelat M, Skelton PH, Rab P (1996) Systematics of *Barbus*: state of the art and heuristic comments. *Folia Zoologica* **45**, 5-12.
- Berrebi P, Valiushok D (1998) Genetic divergence among morphotypes of Lake Tana (Ethiopia) barbs. *Biological Journal of the Linnean Society* **64**, 369-384.

- Birky Jr CW (2001) The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annual Review of Genetics* **35**, 125-148.
- Black WCI, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. *Annual Review of Entomology* **46**, 441-469.
- Bloomer P, Bills IR, van der Bank FH, *et al.* (2007) Multidisciplinary investigation of differences and potential hybridization between two yellowfish species *Labeobarbus kimberleyensis* and *L. aeneus* from the Orange-Vaal system. Follow-up study 2004-2007. Yellowfish Working Group.
- Bloomer P, Naran D (2006) Identification of conservation units of two yellowfish species *Labeobarbus kimberleyensis* and *L. aeneus*. In: *Orange-Vaal yellowfish genetics pilot study report for the YWG and AngloGold Ashanti Ltd*, pp. 1-27.
- Bloomer P, Stobie CS, Reid K, Oosthuizen CJ (In prep) Divergent mitochondrial lineages in the KwaZulu-Natal yellowfish (*Labeobarbus natalensis*) across ancient river systems. Target journal: *BMC Evolutionary Biology*.
- Boehm J, Waldman J, Robinson JD, Hickerson MJ (2015) Population genomics reveals seahorses (*Hippocampus erectus*) of the western mid-Atlantic Coast to be residents rather than vagrants. *PLoS ONE* **10**, e0116219.
- Borisov VB, Shkil FN, Abdissa B, Smirnov SV (2012) Development of the cranium in the large African hexaploid barb *Labeobarbus* (= *Barbus*) *intermedius* (Cyprinidae; Teleostei). *Journal of Ichthyology* **52**, 838-860.
- Boulenger GA (1907) On a small collection of fishes made in the eastern watershed of the Transvaal by Capt. G.E. Bruce. *Proceedings of the Zoological Society of London* **77**, 307-311.
- Bowmaker AP, Jackson PBN, Jubb RA (1978) Freshwater fishes. In: *Biogeography and Ecology of Southern Africa* (ed. Werger MJA), pp. 1181-1230. Springer Netherlands.
- Brawand D, Wagner CE, Li YI, *et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513**, 375-381.
- Brieuc MSO, Waters CD, Seeb JE, Naish KA (2014) A dense linkage map for Chinook salmon (*Oncorhynchus tshawytscha*) reveals variable chromosomal divergence after an ancestral whole genome duplication event. *G3: Genes, Genomes, Genetics* **4**, 447-460.
- Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica* **135**, 439-455.
- Cambray JA, Bills R, Chakona A, Coetzer W, Weyl O (2017) *Labeobarbus natalensis*. The IUCN Red List of Threatened Species 2017: e.T63294A100168851. Downloaded on **07 December 2017**. <<http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T63294A100168851.en>>
- Catchen JM, Amores A, Hohenlohe PA, Cresko WA, Postlethwait JH (2011) Stacks: Building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1**, 171-182.
- Catchen JM, Hohenlohe P, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.
- Chakona A, Malherbe WS, Gouws G, Swartz ER (2015) Deep genetic divergence between geographically isolated populations of the goldie barb (*Barbus pallidus*) in South Africa: potential taxonomic and conservation implications. *African Zoology* **50**, 5-10.
- Chakona A, Skelton PH (2017) A review of the *Pseudobarbus afer* (Peters, 1864) species complex (Teleostei, Cyprinidae) in the eastern Cape Fold Ecoregion of South Africa. *ZooKeys*, **657**, 109-140.
- Chakona A, Swartz ER, Gouws G (2013) Evolutionary drivers of diversification and distribution of a southern temperate stream fish assemblage: testing the role of historical isolation and spatial range expansion. *PLoS ONE* **8**, e70953.
- Chenuil A, Galtier N, Berrebi P (1999) A test of the hypothesis of an autopolyploid vs. allopolyploid origin for a tetraploid lineage: application to the genus *Barbus* (Cyprinidae). *Heredity* **82**, 373-380.

- Chojnowski JL, Kimball RT, Braun EL (2008) Introns outperform exons in analyses of basal avian phylogeny using clathrin heavy chain genes. *Gene* **410**, 89-96.
- Chow S, Hazama K (1998) Universal PCR primers for S7 ribosomal protein gene introns in fish. *Molecular Ecology* **7**, 1255-1256.
- Crass RS (1964) *Freshwater Fishes of Natal*. Shuter & Shooter, Pietermaritzburg.
- Creer S (2007) Choosing and using introns in molecular phylogenetics. *Evolutionary Bioinformatics* **3**, 99-108.
- Curto MA, Puppo P, Ferreira D, Nogueira M, Meimberg H (2012) Development of phylogenetic markers from single-copy nuclear genes for multi locus, species level analyses in the mint family (Lamiaceae). *Molecular Phylogenetics and Evolution* **63**, 758-767.
- DaCosta JM, Sorenson MD (2014) Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE* **9**, e106713.
- Davey JW, Cezard T, Fuentes-Utrilla P, *et al.* (2013) Special features of RAD sequencing data: implications for genotyping. *Molecular Ecology* **22**, 3151-3164.
- Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.
- de Graaf M, Nagelkerke LAJ, Palstra AP, Sibbing FA (2010) Experimental evidence for the biological species status in Lake Tana's *Labeobarbus* flock (Cyprinidae). *Animal Biology* **60**, 183-193.
- de Moor IJ, Bruton MN (1988) Atlas of alien and translocated indigenous animals in southern Africa. In: *South African National Scientific Programmes Report*, p. 317. National Scientific Programmes Unit: CSIR, Pretoria.
- Doadrio I (1994) Freshwater fish fauna of North Africa and its biogeography. *Annales-Musee Royal de l'Afrique Centrale. Sciences Zoologiques* **275**, 21-34.
- Doadrio I, Carmona JA, Machordom A (2002) Haplotype diversity and phylogenetic relationships among the Iberian barbels (*Barbus*, Cyprinidae) reveal two evolutionary lineages. *Journal of Heredity* **93**, 140-147.
- Dou J, Zhao X, Fu X, *et al.* (2012) Reference-free SNP calling: improved accuracy by preventing incorrect calls from repetitive genomic regions. *Biology Direct* **7**, 17.
- Dudgeon D, Arthington AH, Gessner MO, *et al.* (2006) Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews* **81**, 163-182.
- Durand JD, Tsigenopoulos CS, Ünlü E, Berrebi P (2002) Phylogeny and biogeography of the family Cyprinidae in the Middle East inferred from cytochrome *b* DNA - evolutionary significance of this region. *Molecular Phylogenetics and Evolution* **22**, 91-100.
- Eaton DA (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* **30**, 1844-1849.
- Eccles D (1986) Development of the gut in the South African cyprinid fish *Barbus aeneus* (Burchell). *African Zoology* **21**, 165-169.
- Eschmeyer WN, Fong JD (2017) Species by Family/Subfamily. Electronic version accessed **19 November 2017**
<<http://researcharchive.calacademy.org/research/ichthyology/catalog/SpeciesByFamily.asp>>
- Everett MV, Grau ED, Seeb JE (2011) Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources* **11**, 93-108.
- Ferchaud AL, Hansen MM (2016) The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: three-spine sticklebacks in divergent environments. *Molecular Ecology* **25**, 238-259.
- Fouché PSO (2009) Aspects of the ecology and biology of the Lowveld largescale yellowfish (*Labeobarbus marequensis*, Smith, 1843) in the Luvuvhu River, Limpopo River System, South Africa. *PhD thesis*, University of Limpopo (South Africa).

- Gagnaire P-A, Normandeau E, Pavey SA, Bernatchez L (2013) Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (*Coregonus clupeaformis*). *Molecular Ecology* **22**, 3036-3048.
- Gaigher I (1976) The reproduction of *Barbus cf. kimberleyensis* (Pisces, Cyprinidae) in the Hardap dam, South West Africa. *African Zoology* **11**, 97-110.
- Golubtsov AS, Krysanov EY (1993) Karyological study of some cyprinid species from Ethiopia. The ploidy differences between large and small *Barbus* of Africa. *Journal of Fish Biology* **42**, 445-455.
- Gonen S, Bishop SC, Houston RD (2015) Exploring the utility of cross-laboratory RAD-sequencing datasets for phylogenetic analysis. *BMC Research Notes* **8**, 299.
- Gotoh RO, Tamate S, Yokoyama J, Tamate HB, Hanzawa N (2013) Characterization of comparative genome-derived simple sequence repeats for acanthopterygian fishes. *Molecular Ecology Resources* **13**, 461-472.
- Groenewald AvJ (1958) A revision of the genera *Barbus* and *Varicorhinus* (Pisces: Cyprinidae) in Transvaal. *Annals of the Transvaal Museum* **23**, 263-330.
- Hand BK, Hether TD, Kovach RP, *et al.* (2015) Genomics and introgression: discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Current Zoology* **61**, 146-154.
- Hare MP (2001) Prospects for nuclear gene phylogeography. *Trends in Ecology & Evolution* **16**, 700-706.
- Harismendy O, Ng PC, Strausberg RL, *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10**, R32.
- Hassan M, Lemaire C, Fauvelot C, Bonhomme F (2002) Seventeen new exon-primed intron-crossing polymerase chain reaction amplifiable introns in fish. *Molecular Ecology Notes* **2**, 334-340.
- Hecht BC, Campbell NR, Holecek DE, Narum SR (2013) Genome-wide association reveals genetic basis for the propensity to migrate in wild populations of rainbow and steelhead trout. *Molecular Ecology* **22**, 3061-3076.
- Henning F, Lee HJ, Franchini P, Meyer A (2014) Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using dense linkage mapping in non-model organisms. *Molecular Ecology* **23**, 5224-5240.
- Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* **11**, 117-122.
- Hohenlohe PA, Day MD, Amish SJ, *et al.* (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology* **22**, 3002-3013.
- Houston RD, Davey JW, Bishop SC, *et al.* (2012) Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. *BMC Genomics* **13**, 244.
- Igea J, Juste J, Castresana J (2010) Novel intron markers to study the phylogeny of closely related mammalian species. *BMC Evolutionary Biology* **10**, 1-13.
- Impson D, Van der Walt R, Jordaan M (2017) *Labeobarbus seeberi*. The IUCN Red List of Threatened Species 2017: e.T63290A100163027. Downloaded on **07 December 2017**. <<http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T63290A100163027.en>>
- Jubb RA (1963) A revised list of the freshwater fishes of southern Africa. *Annals of the Cape Provincial Museums, Natural History* **3**, 5-39.
- Kai W, Nomura K, Fujiwara A, *et al.* (2014) A ddRAD-based genetic map and its integration with the genome assembly of Japanese eel (*Anguilla japonica*) provides insights into genome evolution after the teleost-specific genome duplication. *BMC Genomics* **15**, 233.

- Karssing RJ (2008) Status of the KwaZulu-Natal yellowfish *Labeobarbus natalensis* (Castelnau, 1861). In: *Technical Report on the State of Yellowfishes in South Africa 2007*. (eds. Impson ND, Bills IR, Wolhuter L), pp. 31-45. Water Research Commission, Gezina, Pretoria.
- Kruger E, Mulder P (1973) Gut length and food habits of fish - a note. *Newsletter of the Limnological Society South Africa* **20**, 1-7.
- Lamer JT, Sass GG, Boone JQ, *et al.* (2014) Restriction site-associated DNA sequencing generates high-quality single nucleotide polymorphisms for assessing hybridization between bighead and silver carp in the United States and China. *Molecular Ecology Resources* **14**, 79-86.
- Larson WA, McKinney GJ, Limborg MT, *et al.* (2016) Identification of multiple QTL hotspots in sockeye salmon (*Oncorhynchus nerka*) using genotyping-by-sequencing and a dense linkage map. *Journal of Heredity* **107**, 122-133.
- Larson WA, Seeb LW, Everett MV, *et al.* (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications* **7**, 355-369.
- Lepais O, Weir JT (2014) SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Molecular Ecology Resources* **14**, 1314-1321.
- Levin BA (2012) New data on morphology of the African scraping feeder *Varicorhinus beso* (Osteichthyes: Cyprinidae) with the special reference to specialized traits. *Journal of Ichthyology* **52**, 908-923.
- Levin BA, Golubtsov AS, Dgebuadze YY, Mugue NS (2013) New evidence of homoplasy within the African genus *Varicorhinus* (Cyprinidae): an independent origin of specialized scraping forms in the adjacent drainage systems of Ethiopia inferred from mtDNA analysis. *African Zoology* **48**, 400-406.
- Li C, Riethoven J-J, Ma L (2010) Exon-primed intron-crossing (EPIC) markers for non-model teleost fishes. *BMC Evolutionary Biology* **10**, 90.
- Lighten J, Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Molecular Ecology Resources* **14**, 753-767.
- Limborg MT, Waples RK, Seeb JE, Seeb LW (2014) Temporally isolated lineages of pink salmon reveal unique signatures of selection on distinct pools of standing genetic variation. *Journal of Heredity* **105**, 835-845.
- Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**, 981-994.
- Macher JN, Rozenberg A, Pauls SU, *et al.* (2015) Assessing the phylogeographic history of the montane caddisfly *Thremma gallicum* using mitochondrial and restriction-site-associated DNA (RAD) markers. *Ecology and Evolution* **5**, 648-662.
- Machordom A, Doadrio I (2001) Evolutionary history and speciation modes in the cyprinid genus *Barbus*. *Proceedings of the Royal Society London B* **268**, 1297-1306.
- Maddison WP (1997) Gene trees in species trees. *Systematic Biology* **46**, 523-536.
- Magnuson JJ (1962) An analysis of aggressive behavior, growth, and competition for food and space in medaka (*Oryzias latipes* (Pisces, Cyprinodontidae)). *Canadian Journal of Zoology* **40**, 313-363.
- Mazel AD (1989) People making history: The last ten thousand years of hunter-gatherer communities in the Thukela basin. *Natal Museum Journal of Humanities* **1**, 1-168.
- McKinney GJ, Waples RK, Seeb LW, Seeb JE (2017) Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources* **17**, 656-669.

- Metzker ML (2009) Sequencing technologies — the next generation. *Nature Reviews Genetics* **11**, 31-46.
- Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* **17**, 240-248.
- Mina M, Mironovsky A, Dgebuadze Y (1996a) Lake Tana large barbs: phenetics, growth and diversification. *Journal of Fish Biology* **48**, 383-404.
- Mina M, Mironovsky A, Dgebuadze Y (1996b) Morphometry of barbel of Lake Tana, Ethiopia: multivariate ontogenetic channels. *Folia Zoologica* **45**, 109-116.
- Mina MV, Mironovsky AN, Golani D (2001) Consequences and modes of morphological diversification of East African and Eurasian barbines (genera *Barbus*, *Varicorhinus* and *Capoeta*) with particular reference to *Barbus intermedius* complex. *Environmental Biology of Fishes* **61**, 241-252.
- Mina MV, Shkil FN, Dzerzhinskii KF, *et al.* (2012) Morphological diversity and age dependent transformations in progeny of the large barbs (*Barbus intermedius* complex *sensu* Banister) of several morphotypes from Lake Tana (Ethiopia). Results of a long-time experiment. *Journal of Ichthyology* **52**, 821-837.
- Moritz C (2002) Strategies to protect biological diversity and the evolutionary processes that sustain it. *Systematic Biology* **51**, 238-254.
- Moritz C, Dowling TE, Brown WM (1987) Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology and Systematics* **18**, 269-292.
- Moyle PB, Baltz DM (1985) Microhabitat use by an assemblage of California stream fishes: Developing criteria for instream flow determinations. *Transactions of the American Fisheries Society* **114**, 695-704.
- Mulder P (1973) Aspects on the ecology of *Barbus kimberleyensis* and *Barbus holubi* in the Vaal River. *African Zoology* **8**, 1-14.
- Mulder P, van Vuuren N, Ferreira J, van der Bank F (1990) A preliminary biochemical comparison of two species of the genus *Barbus* from the Vaal River system. *Water S.A.* **16**, 147-150.
- Myers G (1960) Preface to any future classification of the cyprinid fishes of the genus *Barbus*. *Stanford Ichthyological Bulletin* **7**, 212-215.
- Nagelkerke LA, Sibbing FA (2000) The large barbs (*Barbus* spp., Cyprinidae, Teleostei) of Lake Tana (Ethiopia), with a description of a new species, *Barbus osseensis*. *Netherlands Journal of Zoology* **50**, 179-214.
- Nakamura N, Kasahara S (1955) A study on the phenomenon of the Tobi-Koi or shoot carp. I. On the earliest stage at which the shoot carp appears. *Bulletin of the Japanese Society for the Science of Fish* **21**, 73-76.
- Nakamura N, Kasahara S (1956) A study on the phenomenon of the Tobi-Koi or shoot carp. II. On the effect of particle size and quantity of food. *Bulletin of the Japanese Society for the Science of Fish* **21**, 1022-1024.
- Nakamura N, Kasahara S (1961) A study on the phenomenon of the Tobi-Koi or shoot carp. IV. Effects of adding a small number of larger individuals to the experimental batches of carp fry and culture density upon the occurrence of shoot carp. *Bulletin of the Japanese Society for the Science of Fish* **27**, 958-962.
- Naran D, Skelton PH, Villet MH (2007a) Karyology of the redfin minnows, genus *Pseudobarbus* Smith, 1841 (Teleostei: Cyprinidae): one of the evolutionarily tetraploid lineages of South African barbines. *African Zoology* **41**, 178-182.
- Naran D, Skelton PH, Villet MH (2007b) Karyology of three evolutionarily hexaploid southern African species of yellowfish, *Labeobarbus* Rüppel, 1836 (Cyprinidae). *African Zoology* **42**, 254-260.

- Naylor GJP, Brown WM (1998) Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Systematic Biology* **47**, 61-76.
- Nel J, Murray K, Maherry A, *et al.* (2011) Technical Report for the National Freshwater Ecosystem Priority Areas project. WRC Report 1802/2/11.
- Nichols R (2001) Gene trees and species trees are not the same. *Trends in Ecology & Evolution* **16**, 358-364.
- O'Brien GC, Jacobs F, Cronje L, Wepener V, Smit NJ (2013) Habitat preferences and movement of adult yellowfishes in the Vaal River, South Africa. *South African Journal of Science* **109**, 1-8.
- Oellermann LK (1989) The karyology and taxonomy of the southern African yellowfish (Pisces, Cyprinidae). *MSc thesis*, Rhodes University, Grahamstown (South Africa).
- Oellermann LK, Skelton PH (1990) Hexaploidy in yellowfish species (*Barbus*, Pisces, Cyprinidae) from southern Africa. *Journal of Fish Biology* **37**, 105-115.
- Palumbi SR (1995) Nucleic acids II: the polymerase chain reaction. In: *Molecular Systematics*, 2nd edition (ed. Hillis D, Moritz C), pp. 205-247. Sinauer, Sunderland, MA.
- Palumbi SR (2003) Population genetics, demographic connectivity, and the design of marine reserves. *Ecological Applications* **13**, 146-158.
- Paris JR, Stevens JR, Catchen JM (2017) Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution* **8**, 1360-1373.
- Partridge T, Maud R (2000) Macro-scale geomorphic evolution of southern Africa. *Oxford Monographs on Geology and Geophysics* **40**, 3-18.
- Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7**, e37135.
- Pinho C, Rocha S, Carvalho B, *et al.* (2009) New primers for the amplification and sequencing of nuclear loci in a taxonomically wide set of reptiles and amphibians. *Conservation Genetics Resources* **2**, 181-185.
- Prychitko TM, Moore WS (1997) The utility of DNA sequences of an intron from the b-fibrinogen gene in phylogenetic analysis of woodpeckers (Aves: Picidae). *Molecular Phylogenetics and Evolution* **8**, 193-204.
- Prychitko TM, Moore WS (2000) Comparative evolution of the mitochondrial cytochrome *b* gene and nuclear β -fibrinogen intron 7 in woodpeckers. *Molecular Biology and Evolution* **17**, 1101-1111.
- Pujolar JM, Jacobsen M, Als TD, *et al.* (2014) Assessing patterns of hybridization between North Atlantic eels using diagnostic single-nucleotide polymorphisms. *Heredity* **112**, 627-637.
- Puritz JB, Hollenbeck CM, Gold JR (2015) Fishing for selection, but only catching bias: library effects in double-digest RAD data in a non-model marine species. Poster presented at: *Plant and Animal Genomes XXIII*, 10-15 January 2015, San Diego, CA.
- Quattro J, Jones W, Oswald K (2001) PCR primers for an aldolase-B intron in acanthopterygian fishes. *BMC Evolutionary Biology* **1**, 9.
- Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM (2013) Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology* **22**, 2953-2970.
- Rhymer JM, Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics* **27**, 83-109.
- Ricciardi A, Rasmussen JB (1999) Extinction rates of North American freshwater fauna. *Conservation Biology* **13**, 1220-1222.
- Ricker WE (1958) Handbook of computations for biological statistics of fish populations. *Journal of the Fisheries Research Board of Canada* **119**, 1-300.
- Rivers-Moore N, Goodman P, Nkosi M (2007) An assessment of the freshwater natural capital in KwaZulu-Natal for conservation planning. *Water SA* **33**, 665-674.

- Rochette NC, Catchen JM (2017) Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Protocols* **12**, 2640-2659.
- Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798-804.
- Roux F, Hoffman A, Engelbrecht J, R. B, Cambray J (2017) *Labeobarbus nelspruitensis*. The IUCN Red List of Threatened Species 2017: e.T63301A100174347. Downloaded on **07 December 2017**. <<http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T63301A100174347.en>>
- Sánchez CC, Smith TPL, Wiedmann RT, *et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics* **10**, 559.
- Sang T (2002) Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology* **37**, 121-147.
- Schierup MH, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879-891.
- Schliwen U, Rassmann K, Markmann M, *et al.* (2001) Genetic and ecological divergence of a monophyletic cichlid species pair under fully sympatric conditions in Lake Ejagham, Cameroon. *Molecular Ecology* **10**, 1471-1488.
- Schluter D (2001) Ecology and the origin of the species. *Trends in Ecology & Evolution* **16**, 372-380.
- Schuster SC (2007) Next-generation sequencing transforms today's biology. *Nature Methods* **5**, 16-18.
- Shirahata S (1964) On "cannibalism" in the rainbow trout fingerling, *Salmo gairdnerii irideus*. *Japanese Journal of Ecology* **14**, 25-32.
- Sivasundar A, Bermingham E, Ortí G (2001) Population structure and biogeography of migratory freshwater fishes (*Prochilodus*: Characiformes) in major South American rivers. *Molecular Ecology* **10**, 407-417.
- Skelton P, Bills R (2008) An introduction to African yellowfish and to this report. In: *Technical Report on the State of Yellowfishes in South Africa 2007* (eds. Impson ND, Bills IR, Wolhuter L), pp. 1-14. Water Research Commission, Pretoria.
- Skelton PH (1986) Fish of the Orange-Vaal system. In: *The Ecology of River Systems* (eds. Davies BR, Walker KF), pp. 143-162. Dr W. Junk Publishers, Dordrecht, The Netherlands.
- Skelton PH (2001) *A Complete Guide to the Freshwater Fishes of Southern Africa*. Struik, Cape Town.
- Skelton PH (2010) Flagships and fragments — perspectives on the conservation of freshwater fishes in southern Africa. *African Journal of Aquatic Science* **25**, 37-42.
- Smith SA, Bermingham E (2005) The biogeography of lower Mesoamerican freshwater fishes. *Journal of Biogeography* **32**, 1835-1854.
- Swartz E (2005) Phylogeography, phylogenetics and evolution of the redfins (Teleostei, Cyprinidae, *Pseudobarbus*) in southern Africa. *PhD thesis*, University of Pretoria (South Africa).
- Swartz ER, Chakona A, Skelton PH, Bloomer P (2014) The genetic legacy of lower sea levels: does the confluence of rivers during the last glacial maximum explain the contemporary distribution of a primary freshwater fish (*Pseudobarbus burchelli*, Cyprinidae) across isolated river systems? *Hydrobiologia* **726**, 109-121.
- Swartz ER, Mwale M, Hanner R (2008) A role for barcoding in the study of African fish diversity and conservation. *South African Journal of Science* **104**, 293-298.
- Swartz ER, Skelton PH, Bloomer P (2007) Sea-level changes, river capture and the evolution of populations of the Eastern Cape and fiery redfins (*Pseudobarbus afer* and *Pseudobarbus phlegethon*, Cyprinidae) across multiple river systems in South Africa. *Journal of Biogeography* **34**, 2086-2099.
- Swartz ER, Skelton PH, Bloomer P (2009) Phylogeny and biogeography of the genus *Pseudobarbus* (Cyprinidae): Shedding light on the drainage history of rivers associated with the Cape Floristic Region. *Molecular Phylogenetics and Evolution* **51**, 75-84.

- Templeton AR (2004) Statistical phylogeography: Methods of evaluating and minimizing inference errors. *Molecular Ecology* **13**, 789-809.
- Tsigenopoulos CS, Berrebi P (2000) Molecular phylogeny of North Mediterranean freshwater barbs (genus *Barbus*: Cyprinidae) inferred from cytochrome *b* sequences: Biogeographic and systematic implications. *Molecular Phylogenetics and Evolution* **14**, 165-179.
- Tsigenopoulos CS, Kasapidis P, Berrebi P (2010) Phylogenetic relationships of hexaploid large-sized barbs (genus *Labeobarbus*, Cyprinidae) based on mtDNA data. *Molecular Phylogenetics and Evolution* **56** 851-856.
- Unmack PJ (1999) Biogeography of Australian Freshwater Fishes. *PhD thesis*, Arizona State University (USA).
- van der Walt K-A, Swartz ER, Woodford D, Weyl O (2017) Using genetics to prioritise headwater stream fish populations of the Marico barb, *Enteromius motebensis* Steindachner 1894, for conservation action. *Koedoe* **59**, 1-7.
- van Vuuren N, Mulder P, Ferreira J, van der Bank FH (1989) The identification of hybrids of *Barbus aeneus* X *B. kimberleyensis* and *Labeo capensis* X *L. umbratus* in Hardap Dam, SWA/Namibia. *Madoqua* **16**, 27-34.
- Vilizzi L, Walker KF (1999) The onset of the juvenile period in carp, *Cyprinus carpio*: a literature survey. *Environmental Biology of Fishes* **56**, 93-102.
- von Bubnoff A (2008) Next-Generation Sequencing: The race is on. *Cell* **132**, 721-723.
- Vreven EJ, Musschoot T, Snoeks J, Schliewen UK (2016) The African hexaploid Torini (Cypriniformes: Cyprinidae): review of a tumultuous history. *Zoological Journal of the Linnean Society* **177**, 231-305.
- Wagner CE, Keller I, Wittwer S, *et al.* (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* **22**, 787-798.
- Waples R, Seeb L, Seeb J (2016) Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Molecular Ecology Resources* **16**, 17-28.
- Waters JM, Cambray JA (1997) Intraspecific phylogeography of the *Cape galaxias* from South Africa: evidence from mitochondrial DNA sequences. *Journal of Fish Biology* **50**, 1329-1338.
- Willing E-M, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics* **27**, 2187-2193.
- Willows-Munro S, Robinson TJ, Matthee CA (2005) Utility of nuclear DNA intron markers at lower taxonomic levels: Phylogenetic resolution among nine *Tragelaphus* spp. *Molecular Phylogenetics and Evolution* **35**, 624-636.
- Wilson AC, Cann RL, Carr SM, *et al.* (1985) Mitochondrial DNA and two perspectives on evolutionary genetics. *Biological Journal of the Linnean Society* **26**, 375-400.
- Wishart M, Hughes J, Stewart B, Impson D (2006) Extreme levels of intra-specific divergence among Cape Peninsula populations of the Cape galaxias, *Galaxias zebratus* Castelnau 1861, reveals a possible species complex. *African Journal of Aquatic Science* **31**, 99-106.
- Wohlfarth G (1977) Shoot carp. *Bamidgeh* **29**, 35-39.
- Xuzhen W, Shunping H, Yiyu C (2002) Sequence variations of the S7 ribosomal protein gene in primitive cyprinid fishes: Implication on phylogenetic analysis. *Chinese Science Bulletin* **47**, 1638-1643.
- Yang L, Sado T, Hirt MV, *et al.* (2015) Phylogeny and polyploidy: resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). *Molecular Phylogenetics and Evolution* **85**, 97-116.
- YWG (2013) Discussion at the Yellowfish Working Group Conference 2013, Vaal Streams Eco-Resort, Free State, 20 April 2013.

- Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Molecular Ecology* **11**, 1-16.
- Zardoya R, Doadrio I (1998) Phylogenetic relationships of Iberian cyprinids: systematic and biogeographical implications. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **265**, 1365-1372.
- Zardoya R, Doadrio I (1999) Molecular evidence on the evolutionary and biogeographical patterns of European cyprinids. *Journal of Molecular Evolution* **49**, 227-237.

Chapter 2: Exploring the phylogeography of a hexaploid freshwater fish by RAD sequencing

Abstract

The KwaZulu-Natal yellowfish (*Labeobarbus natalensis*) is an abundant cyprinid, endemic to KwaZulu-Natal Province, South Africa. In this study I developed a Single Nucleotide Polymorphism (SNP) dataset from double-digest Restriction-site Associated DNA (ddRAD) sequencing of samples across the distribution. I addressed several hidden challenges, primarily focussing on proper filtering of RAD data and selecting optimal parameters for data processing in polyploid lineages. I used the resulting high-quality SNP dataset to investigate the population genetic structure of *L. natalensis*. A small number of mitochondrial markers present in these data had disproportionate influence on the recovered genetic structure. The presence of singleton SNPs also confounded genetic structure. I found a well-supported division into northern and southern lineages, with further subdivision into five populations, one of which reflects north-south admixture. Approximate Bayesian Computation scenario testing supported a scenario where an ancestral population diverged into northern and southern lineages, which then diverged to yield the current five populations. All river systems showed similar levels of genetic diversity, which appears unrelated to drainage system size. Nucleotide diversity was highest in the smallest river system, the Mbokodweni, which, together with adjacent small coastal systems, should be considered as a key catchment for conservation.

2.1 Introduction

The Cyprinidae is the largest freshwater fish family, comprising approximately 80% of all freshwater fish species in temperate zones (Naran *et al.* 2007), and including over 2400 species (Swartz *et al.* 2008; de Graaf *et al.* 2010). Within Cyprinidae, the African genus *Labeobarbus* remains relatively understudied. This genus has recently been grouped into the tribe Torini (Yang *et al.* 2015). *Labeobarbus* is thought to have arisen from hybridisation between a tetraploid ancestor in Torini and a diploid ancestor of *Cyprinion* followed by autopolyploidisation, resulting in the current hexaploid lineage ($2N = \pm 150$ chromosomes) some time prior to their colonisation of Africa c. 13 mya (Oellermann & Skelton 1990; Tsigenopoulos *et al.* 2010; Yang *et al.* 2015). This lineage has subsequently speciated into 125 valid species (Vreven *et al.* 2016).

Seven species of *Labeobarbus* exist in southern Africa. Five of these (*L. aeneus*, *L. capensis*, *L. kimberleyensis*, *L. natalensis*, and *L. polylepis*) likely originated from a common ancestor invading the Orange River Basin c. 2-3 mya (Skelton 1986). Major geological events c. 5.1 mya resulted in deep riverine valleys separating current drainage systems across KwaZulu-Natal Province of South Africa (Partridge & Maud 2000; Rivers-Moore *et al.* 2007) which were later colonised by ancestors of the endemic KwaZulu-Natal yellowfish, *L. natalensis* (de Castelnau 1861). The prevalence of physical barriers such as waterfalls doubtless affected this process, restricting freshwater fish movement or leading to unidirectional movement. This, combined with the stenohaline nature of the fish, make it difficult to understand the dispersal pathways resulting in the now widespread occurrence of the species in the KwaZulu-Natal rivers.

Despite the current IUCN Red List assessment of *L. natalensis* as Least Concern (Cambray *et al.* 2017), the species may be declining (Karssing 2008). The genus is highly popular in South Africa both for subsistence and recreational anglers (Skelton & Bills 2008), and is also used as an indicator of river health – their presence showing low water pollution and few alien fish species (Skelton & Bills 2008). As such, conservation management is needed for this “flagship” species for freshwater

systems (Skelton & Bills 2008). This should include quantifying the genetic diversity of populations across the species' geographical range (Palumbi 2003; Smith & Bermingham 2005). Various phylogeographic studies have been conducted on cyprinids (Machordom & Doadrio 2001; Durand *et al.* 2002) although few of these have explored the South African branches of the family (but see Swartz *et al.* 2007; Swartz *et al.* 2009; Chakona *et al.* 2013; Swartz *et al.* 2014; Chakona *et al.* 2015; Chakona & Skelton 2017; van der Walt *et al.* 2017).

Previous analyses based on mitochondrial DNA (mtDNA) data showed substantial differences between populations of *L. natalensis* across its distribution (Bloomer *et al.* 2007; Bloomer *et al.* In prep). More variation was reported between these populations than between two other South African species, *L. aeneus* and *L. kimberleyensis* (Bloomer *et al.* 2007). Six primary mitochondrial haplogroups were identified for *L. natalensis*, matching major drainage systems – from north to south: the Umfolozi, Tugela, Umgeni, Mbokodweni, Mkomaas and Mzimkhulu systems (Bloomer *et al.* In prep). This suggests historical isolation among these drainage systems. The most notable divide was between the northern and southern drainage systems. This disjunction does not correspond closely with any known biogeographic transition. In general, KwaZulu-Natal has a rich and geographically varied freshwater fauna, but this diversity occurs as a complex regional mosaic, reflecting historical interchange among tropical and temperate faunal elements with substantial local endemism (Rivers-Moore *et al.* 2007; Perera *et al.* 2011). The initial *L. natalensis* phylogeographic study was based entirely on mitochondrial markers, and thus remains to be verified with genomic data. The processes that may have resulted in genetic structure also remain to be identified.

At present there is no close reference genome for *Labeobarbus*, in which ancestral hexaploidy has resulted in large and highly paralogous genomes. Consequently I decided to use a reduced representation approach, Restriction-site Associated DNA (RAD) sequencing (Miller *et al.* 2007; Baird *et al.* 2008), to understanding genomic diversity in *L. natalensis*. This method is popular and has

been used in many studies since its inception (Fig. 2.1). RAD sequencing has been used, particularly in fish, to identify population divergence (Larson *et al.* 2014; Boehm *et al.* 2015; Ferchaud & Hansen 2016), for SNP identification in polyploid fish (Hohenlohe *et al.* 2011; Ogden *et al.* 2013; Palti *et al.* 2014), in phylogeographic studies (Reitzel *et al.* 2013; Macher *et al.* 2015), for QTL analysis (Houston *et al.* 2012; Gagnaire *et al.* 2013; Yoshizawa *et al.* 2015), for linkage mapping (Brieuc *et al.* 2014; Henning *et al.* 2014), in hybridisation studies (Lamer *et al.* 2014; Pujolar *et al.* 2014; Hand *et al.* 2015), for exploration of genome architecture and evolution (Brawand *et al.* 2014; Kai *et al.* 2014; Waples *et al.* 2016), and in phylogenetic analyses (Wagner *et al.* 2013; Gonen *et al.* 2015). This methodology should be particularly suited to phylogeographic studies as the inference power from large numbers of markers may identify patterns that are not easily visible in traditional analyses based on relatively few loci (Davey *et al.* 2011). Quality control is critical for RAD sequencing analyses and is conducted at various stages via an analytical pipeline prior to interpreting results for meaningful biological relationships (Davey *et al.* 2013).

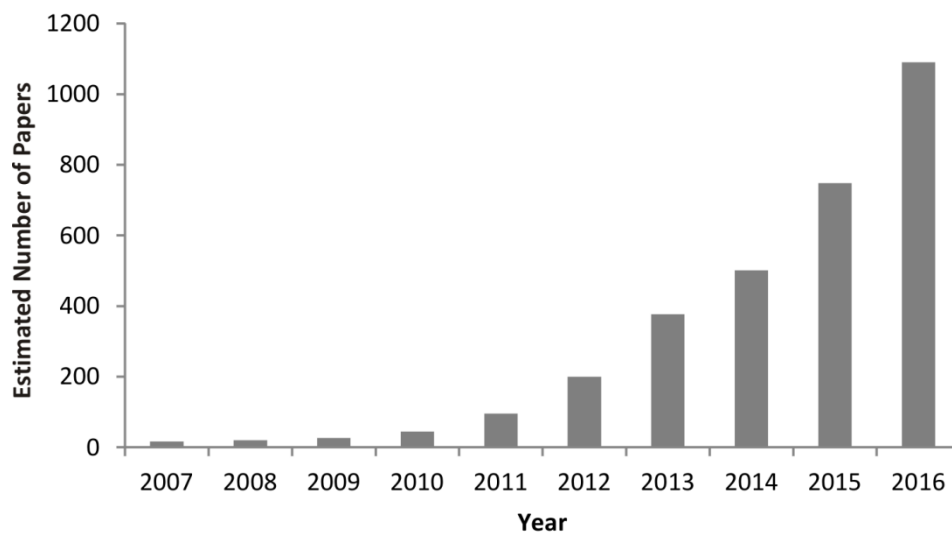


Fig. 2.1 Estimated number of research articles using RAD sequencing by year, indicating near-exponential growth. Number of research articles was estimated by the number of Google Scholar hits for “Restriction-site associated DNA”.

Double-digest RAD (ddRAD) sequencing (Peterson *et al.* 2012) addresses several coverage issues in the original RAD protocol by replacing random-shearing of fragments with a second restriction enzyme. Targeted fragments are defined on one end by a common restriction site as in standard RAD sequencing, but differ in being flanked by a less common restriction site at the other end (Peterson *et al.* 2012). This approach results in higher repeatability, better control over genome coverage, greater sharing of sequenced fragments, and similar sequence read proportions across individuals (Peterson *et al.* 2012). The additional restriction digestion may also introduce artefacts however, as mutations in restriction sites may result in underestimation of diversity due to allele dropout (Arnold *et al.* 2013). Indels, combined with stringent size selection, may also result in loci being dropped or included in particular individuals or populations during ddRAD sequencing (DaCosta & Sorenson 2014).

Polyploidy complicates most genetic analyses of *Labeobarbus*. Many studies of polyploids advocate analysis of non-nuclear markers or the transcriptome (Everett *et al.* 2011). However, a number of studies using RAD sequencing have recently tackled the challenge, particularly in tetraploid fish. Several strategies have emerged to circumvent the complicating issue of paralogy (reviewed in McKinney *et al.* 2017). These include removing diallelic markers yielding more than two alleles or haplotypes per individual and excluding loci where more than half the individuals genotyped appear heterozygous (Hohenlohe *et al.* 2011; Hohenlohe *et al.* 2013). Recently, McKinney *et al.* (2017) suggested the HDPlot approach, which compares heterozygosity at each diallelic locus across a population with read depth for each allele.

In this study I used ddRAD sequencing of samples from across the distribution of *L. natalensis* to identify phylogeographic patterns and processes affecting this species. To do this, I developed a pipeline to filter sequencing artefacts and paralogs from diallelic SNP data, resulting in a high-quality genomic resource for use in *Labeobarbus*.

2.2 Materials and methods

2.2.1 Sample collection and DNA extraction

Samples were collected from 13 localities across the KwaZulu-Natal Province, South Africa (Fig. 2.2; Supporting Information Table S2.1) between March 2003 and November 2007 by members of the Yellowfish Working Group. The localities selected represented most major drainage systems across the species distribution. Muscle and fin samples were obtained and stored in 96% ethanol at 4°C. DNA was extracted using the phenol-chloroform method (Sambrook *et al.* 1989).

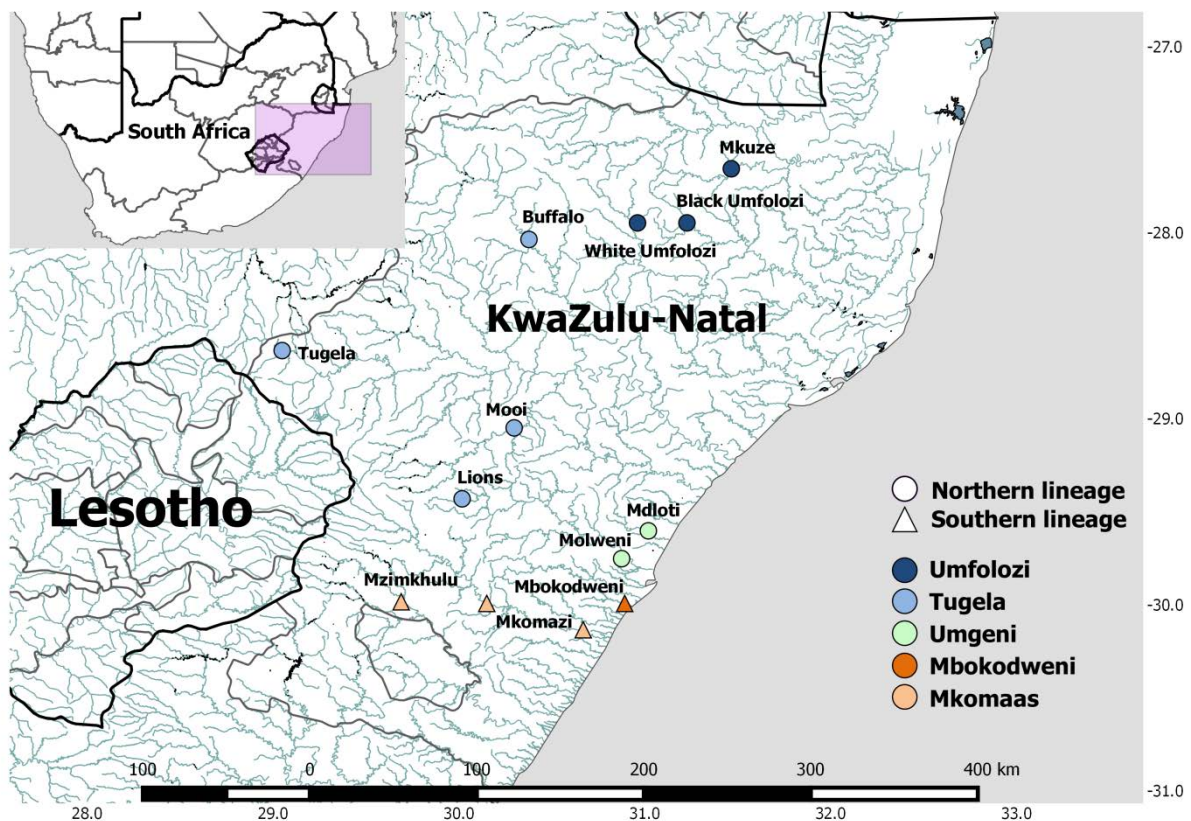


Fig. 2.2 Distribution of samples in this study across KwaZulu-Natal with reference to a map of South Africa (top left). River names are indicated at each sampling site. The colour of each sampling site corresponds to the putative population identified in this study. The shape of the symbol indicates an association to either the northern or the southern lineage. This map was produced using QGIS (QGIS Development Team, 2016. QGIS Geographic Information System. Open Source Geospatial Foundation Project. <http://www.qgis.org/>) and the National Freshwater Ecosystem Priority Areas (NFEPA) project (Nel *et al.* 2011).

A GeneQuant™ *pro* RNA/DNA calculator spectrophotometer (Amersham Biosciences, Freiberg, Germany), Qubit® 2.0 Fluorometer (Invitrogen, USA), and agarose gel electrophoresis were used to assess DNA quality and concentration. High quality samples were sent to Beijing Genomics Institute Hong Kong Co., Limited (BGI, Hong Kong) to undergo the ddRAD sequencing protocol as per Peterson *et al.* (2012). In total 23 high-quality samples were selected for analysis (Supporting Information Table S2.1).

2.2.2 Library preparation and sequencing

Samples were divided into two libraries, which were digested with the restriction enzymes *NlaIII* (CATG/) and *MluCI* (/AATT) following the double-digest paired-end protocol of Peterson *et al.* (2012). This combination of enzymes is anticipated to produce approximately 100,000 RAD-tags based on digestion of zebrafish, *Danio rerio* (Peterson *et al.* 2012). Each individual was tagged with a unique 4-8 base pair barcode, with two replicate individuals barcoded and sequenced in separate libraries, as controls. Each library was size selected for fragments of 200-400 bp. Final libraries were sequenced using 90 bp paired-end sequencing in a single lane of an Illumina HiSeq 2000 (Illumina Inc., USA). The resulting reads were screened for poor quality (reads with more than 50% low quality bases i.e. quality value ≤ 5 (E)), trimmed to remove adapters and barcodes, and demultiplexed prior to analysis.

2.2.3 Bioinformatics pipeline, quality control and read mapping

Data were cleaned both with custom bioinformatic expressions and using the program PROCESS_RADTAGS in STACKS 1.44 (Catchen *et al.* 2011; Catchen *et al.* 2013). Reads were first trimmed to a uniform read length of 80 bp to reduce the effect of sequencing error, after examination of SNP density spectra generated from the untrimmed data (Supporting Information Fig. S2.1). Quality of

the bases was assessed for each sample both before and after trimming using the program FASTQC (Andrews 2010) (Supporting Information Figs. S2-S3). After trimming, the parameters $-r$ (rescue RAD tags), $-c$ (clean data, remove reads with an uncalled base) and $-q$ (remove low quality reads) were specified in `PROCESS_RADTAGS`. Only reads that remained paired after processing were retained. Adapter-pollution, remnants of adapter sequences that were not removed by earlier trimming, was filtered using custom bioinformatic expressions. The degree of overlapping reads and incomplete restriction digestion in the dataset was estimated using custom bioinformatic expressions and CLC GENOMICS WORKBENCH 7.0.4 (CLC Inc., Aarhus, Denmark). The latter program was also used to evaluate read quality and to map filtered reads to the closest cyprinid reference genome, the common carp (*Cyprinus carpio*, GCF_000951615.1). Filtered reads were similarly mapped to the set of carp coding DNA sequences (CDS) from the same reference after removing mitochondrial genes. I used a mismatch cost of 2, insertion and deletion costs of 3, and length fractions and similarity fractions of 0.9 to retain only highly plausible mappings. Non-specific matches were ignored.

2.2.4 SNP discovery

Due to the lack of a close reference genome, the wrapper program `DENOVO_MAP.PL` was used to identify diallelic SNP loci by executing `USTACKS`, `CSTACKS` and `SSTACKS` (Catchen *et al.* 2011; Catchen *et al.* 2013). I chose to use only Read 1 files (*Nla*III cut-site) for the assembly to avoid duplication of SNPs due to overlapping reads from fragment ends and false SNPs caused by adapter-pollution. Optimal parameters were identified using the method outlined in Paris *et al.* (2017). Briefly, I constructed plots of the average read depth across samples while varying the minimum stack depth parameter $-m$ (Fig. 2.3). I then compared the number of SNPs, assembled loci and polymorphic loci for each sample and across samples using the 80% sample representation cut-off suggested by Paris *et al.* (2017) while varying the minimum stack depth ($-m$) and distance allowed between stacks ($-M$) from

defaults of $-m\ 5\ -M\ 3$ (Supporting Information Figs. S4-S5). Finally, the maximum distance required to merge catalog loci ($-n$) was assessed by evaluating the change in number of polymorphic loci for $n=M-1$, $n=M$ and $n=M+1$ (Supporting Information Table S2.2). The `--max_locus_stacks` default value was set to 7 to ensure adequate binning and avoid paralogs. The `-t` flag was specified while running `DENOVO_MAP.PL` to remove or split highly repetitive tags during `USTACKS`. From this procedure I identified optimal parameters for this dataset to be $-m\ 5\ -M\ 1\ -n\ 0$.

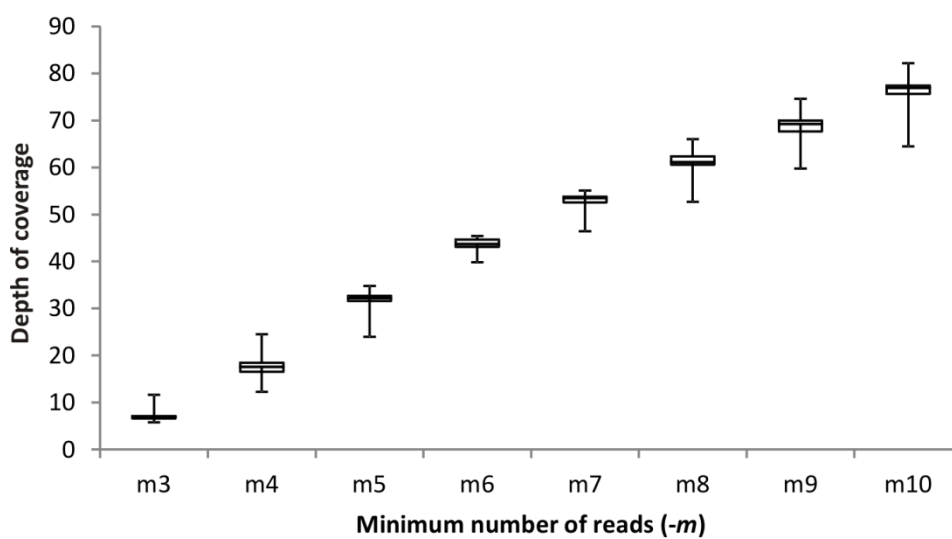


Fig. 2.3 Box-and-whisker plot showing the distribution of the mean depth of coverage across all 23 samples (plus two replicates) as the value for the minimum number of reads ($-m$) is varied from 3 to 10, as per Paris *et al.* (2017). Whiskers here indicate the maximum and minimum values across the samples. Paris *et al.* (2017) recommend a depth of coverage of $>25\times$, which would indicate that $-m = 5$ is most suitable for this dataset. The default parameter set was $-m\ 5\ -M\ 3\ -n\ 2$.

Mitochondrial reads were detected using nucleotide BLAST (Altschul *et al.* 1997) to match ($E \leq 1 \times 10^{-20}$) against the available mitogenomes for five related genera – *Labeobarbus* (*L. intermedius* NC_031531.1; *L. sp. Kongou* AP011324.1; *L. sp. Lucien* AP011323.1), *Varicorhinus* (NC_031528.1), *Tor* (NC_027617.1; KU870466.1; NC_021755.1; AP011326.1; NC_022702.1; JX444718.1; AP011372.1; KR868704.1), *Neolissochilus* (NC_026106.1; KU553349.1; AP011314.1; NC_031555.1) and *Hypselobarbus* (NC_031587.1). I removed potential paralogs that had been merged by identifying

loci that possessed more than two haplotypes within a single sample. I compared this haplotype approach with the HDPLLOT method of McKinney *et al.* (2017). A blacklist was constructed in POPULATIONS for loci identified as mitochondrial or paralogous.

I retained loci from POPULATIONS that were scored in at least 60% of all individuals. I used a minor allele frequency (MAF) filter of 0.04 to filter out singleton SNPs that may mask population structure (Roesti *et al.* 2012; Rodríguez-Ezpeleta *et al.* 2016), and a maximum observed heterozygosity filter of 0.99 to remove SNPs that were reported as heterozygotes in all samples the SNP was called in, which are potentially paralogous loci. Finally, because some analyses required a single SNP per locus, I filtered the dataset by selecting the most informative SNP per locus based on the number of minor alleles. Where multiple SNPs at a locus had the same number of minor alleles, I chose the first SNP with the best representation across samples. Loci that passed all filtering criteria were extracted using a whitelist and run through POPULATIONS again to produce the final dataset of 723 SNPs.

2.2.5 Population genetic parameters and structure

Output from POPULATIONS was exported in STRUCTURE and GENEPOP formats and converted to other formats, as needed, using PGD SPIDER (Lischer & Excoffier 2012). STRUCTURE 2.3.4 (Pritchard *et al.* 2000) was used to infer population structure with 100,000 chains as burn-in and 500,000 MCMC chains with 20 iterations for $K = 1-8$. I followed the same protocol for further hierarchical STRUCTURE runs for the two lineages identified from the primary run. The result files were run through STRUCTURE HARVESTER (Earl & VonHoldt 2011) and the optimal K -value was determined by the method of Evanno *et al.* (2005). CLUMPP 1.1.2 (Jakobsson & Rosenberg 2007) was used to visualise the data.

In addition to STRUCTURE, RADPAINTER and FINERADSTRUCTURE (Malinsky *et al.* 2016) were used as an independent assessment of population structure, as this package is designed to identify co-

ancestry from RAD data. Haplotypes were run through the FINERADSTRUCTURE pipeline using default parameters of 100,000 burn-in and 100,000 MCMC steps with sampling occurring every 1,000 MCMC steps. A tree was constructed with 10,000 hill-climbing iterations. Populations were defined as clusters within the FINERADSTRUCTURE tree and relatedness plot. The first three axes of variation were used in Principal Component Analysis (PCA) plots. Additionally, Factorial Correspondence Analysis (FCA) plots were generated in GENETIX 4.05.2 (Belkhir *et al.* 2004) from the same SNP dataset.

Pairwise population F_{ST} values were estimated among populations inferred from these analyses using ARLEQUIN 3.5.1.2 (Excoffier & Lischer 2010). Also, the contribution of highly differentiated markers on the observed population structure was assessed by identification of F_{ST} outlier loci using BAYESCAN 2.1 (Foll & Gaggiotti 2008) and comparing analyses of structure based on these loci, and excluding outlier loci. Finally, I identified a prevalent *HindIII* satellite using nucleotide BLAST ($E \leq 1 \times 10^{-10}$) against sequenced monomers available in cyprinids *Acrossocheilus paradoxus* (AJ241977.1) and *C. carpio* (M19418.1) and similarly assessed its contribution to population structure.

2.2.6 Population history

DIYABC 2.1.0 (Cornuet *et al.* 2014) was used to test a number of simple evolutionary scenarios. The dataset was reduced to 661 SNPs by excluding SNPs with missing data for entire populations. One million simulated datasets were generated per scenario at each stage of the scenario testing. I first tested six basic scenarios of a basal split separating one population from all others or a basal polytomy (Supporting Information Fig. S2.6A). I then tested 24 ladder-like scenarios of successive divergence (Supporting Information Fig. S2.6B). I also tested ten scenarios where the ancestral population split into two major lineages which then diverged into five regional populations (Supporting Information Fig. S2.6C). Finally, I compared the best supported scenarios from the above

tests as well as two variants that model the Umgeni population as a product of admixture (Supporting Information Fig. S2.6D). Scenarios were assessed using logistic regression analysis (1% of total datasets) comparing observed versus simulated values of standard summary statistics of genic diversity, population structure and Nei's distances with all other settings at the default values for SNP datasets.

The average nucleotide diversity across all loci ($\pi = 0.0035$) was used to determine the long-term effective population size (N_e) using the equation from Tajima (1983): $\pi = 4 * N_e * \mu$. The mutation rate per site per generation (μ) is calculated using a rate of between 1×10^{-8} and 1×10^{-9} per site per year in SNPs (Brumfield *et al.* 2003).

2.3 Results

2.3.1 Library preparation and sequencing

Illumina 90 bp paired-end sequencing produced over 137 million reads (Table 2.1) for the 23 individuals. This yielded over 12,027 million total base pairs prior to filtering and trimming, or an average of 523 million base pairs per individual with a range of 372-1,477 million. The average GC content of filtered reads per individual was between 38.5% and 40.8%. The Q20 scores of reads for each individual were within the range of 97.10% and 97.95%, while Q30 was 93.53% to 94.81%.

Table 2.1 Summary information from initial analysis of RAD sequencing data

Raw reads	Q20 %	Q30%	GC %
137,459,448	97.10 - 97.95	93.53 - 94.81	38.5 - 40.8

2.3.2 Bioinformatics, mapping and SNP discovery

Initial quality control of the data identified over 52 million high-quality paired reads without adapter-pollution. Mapping reads against the common carp reference genome resulted in an average of 830,241 reads mapped per individual (Table 2.2). The average coverage of mapped reads was 2.6% of the *C. carpio* genome. In contrast, mapping against *C. carpio* nuclear coding sequences yielded an average number of 55,844 mapped reads per individual with an average coverage of 3.2% of the *C. carpio* CDS reference.

Table 2.2 Mapping of RAD sequencing reads to the *C. carpio* nuclear (GCF_000951615.1) genome

Mapping	Average reads mapped per individual	Average coverage per individual (%)	Total coverage (%)
<i>Cyprinus carpio</i> nuclear genome	830,241	2.6	10
<i>Cyprinus carpio</i> entire CDS	55,844	3.2	14

Selecting only the first read from each pair gave the current list of 50,740 loci and 17,256 SNPs identified through the STACKS pipeline (Table 2.3). The two methods for identification of potential paralogs, excess haplotypes per individual and HDPLLOT, produced very different results (see discussion below). These methods identified 2,435 and 463 loci respectively (Supporting Information Fig. S2.7). I opted to follow the excess haplotype approach as it listed more putative paralogs. Filtering in POPULATIONS of mitochondrial loci (121), potentially paralogous loci (2,435), SNPs present in less than 60% of samples (536), SNPs with a MAF of less than 0.04 (1,142) and SNPs recorded as heterozygotes across all samples (125) resulted in a dataset of 826 high-quality SNPs. This dataset was further reduced to a single SNP per locus based on the number of minor alleles present to accommodate some downstream analysis programs, resulting in the final dataset of 723 SNPs. SNPs found from the initial STACKS pipeline and in the final dataset show transition/transversion ratios of

1.58 and 2.04, respectively. The SNP identification process is summarised in Table 2.3 and distribution of SNPs across individuals in Supporting Information Fig. S2.8.

Table 2.3 Summary of the SNP identification process and filtering steps to the final SNP dataset

Total raw reads	137,459,448
Processed single reads	52,951,112
Loci assembled	50,740
SNPs identified	17,256
Less filters:	
Mitochondrial loci	121
Potentially paralogous loci	2,435
SNPs represented across <60% individuals	536
SNPs with MAF < 0.04	1,142
SNPs with $H_o > 0.99$	125
SNPs passing all filters	826
One SNP per locus based on number of minor alleles	723

MAF = minor allele frequency; H_o = Observed heterozygosity

2.3.3 Population structure and variation

Analysis of 723 loci with STRUCTURE revealed a well-supported split at $K = 2$ between the northern and southern populations (Fig. 2.4). Individuals from the lower Umgeni (geographically between the Mbokodweni and Tugela drainage systems) appeared as potentially admixed individuals between these northern and southern lineages. Some evidence of further structure at $K = 3-5$ is present but this is not as well supported. However, hierarchical STRUCTURE (Fig. 2.5) showed further subdivision into five populations: Umfolozi, Tugela, Umgeni, Mbokodweni and Mkomaas, from north to south. The two samples from the Mzimkhulu system, near the southern distribution limit, were not distinguished from those in the adjacent Mkomaas, and I treat these here as a single population. Similarly, the single sample from Lions River, a tributary of the upper Umgeni, clustered closely with those from the Tugela system, rather than the lower Umgeni, and was considered part of the Tugela population in subsequent analyses (see discussion below).

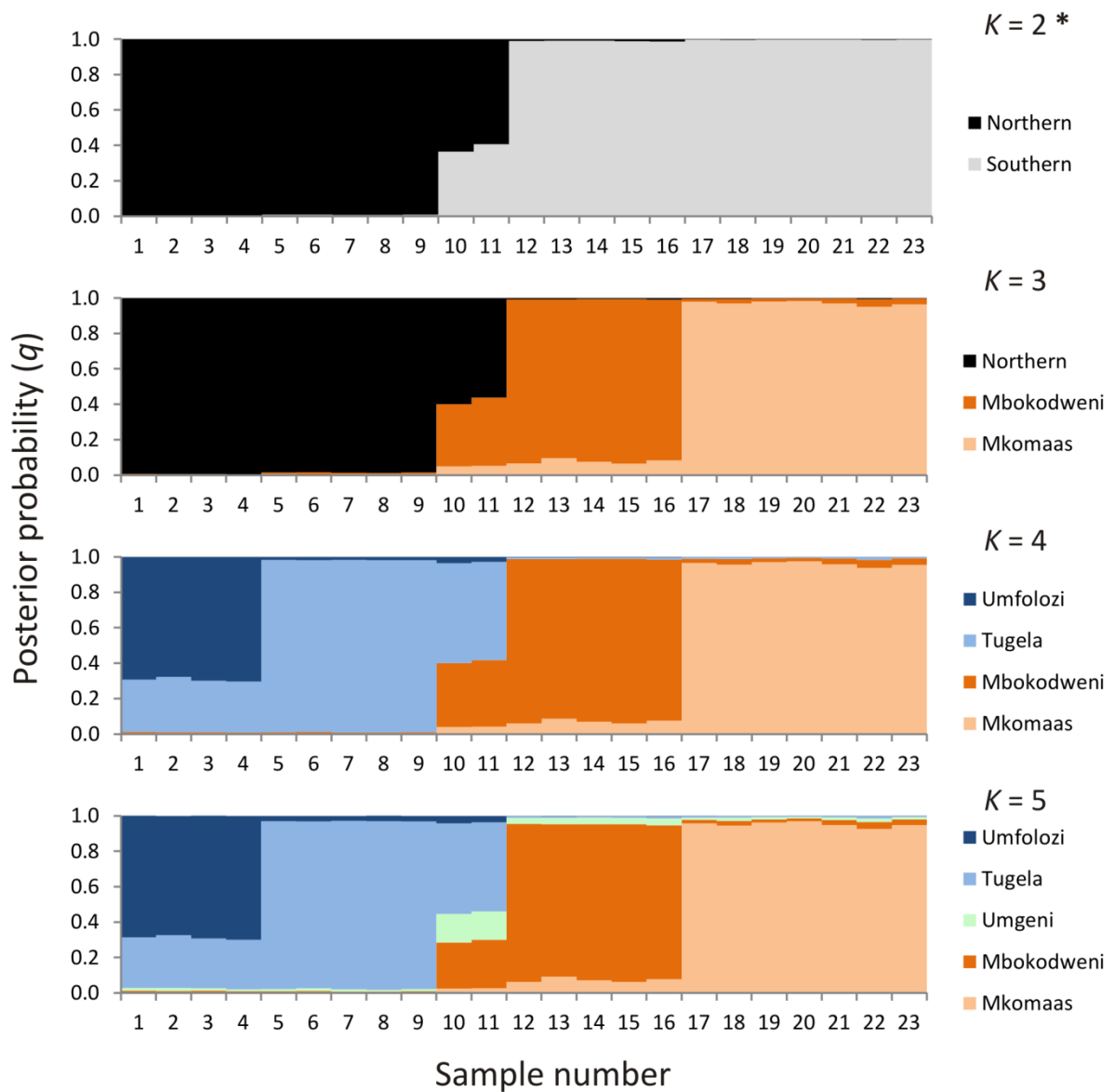


Fig. 2.4 STRUCTURE analysis with $K = 2-5$ using 723 high-quality filtered loci. Each individual (indicated as columns along the X-axis) is probabilistically assigned (probability of assignment q on the Y-axis) to one of the inferred genetic clusters. $K = 2$ was recovered as having the most support with the Evanno method (Evanno *et al.* 2005). CLUMPP was used to produce this representation from 20 replicates.

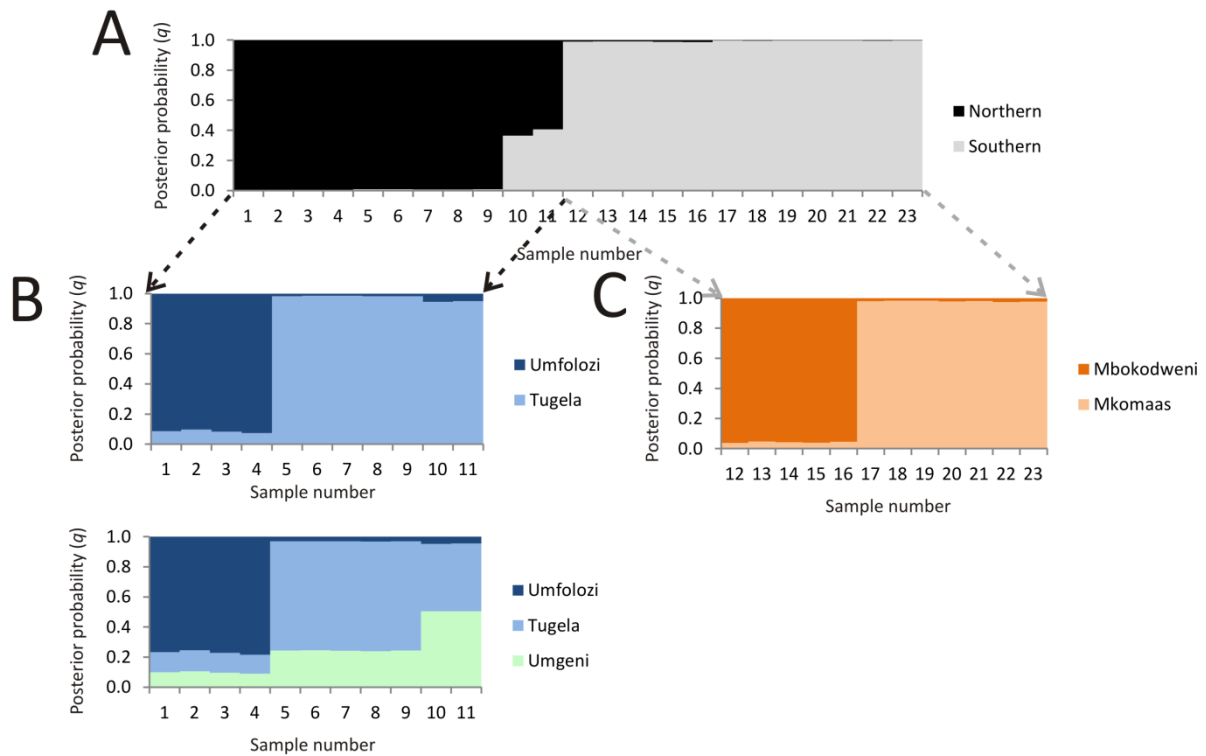


Fig. 2.5 Hierarchical STRUCTURE results for the northern and southern lineages. Each individual (indicated as columns along the X-axis) is probabilistically assigned (probability of assignment q on the Y-axis) to one of the inferred genetic clusters. A – primary run of STRUCTURE for $K=2$ on which the hierarchical STRUCTURE was based. B – northern lineage hierarchical STRUCTURE for $K=2$ and $K=3$ showing the results for minor clusters (4/20 replicates and 6/20 replicates respectively) only when location is specified as prior. C – southern lineage hierarchical STRUCTURE for $K=2$ showing the major cluster (14/20 replicates) with location as prior.

FINERADSTRUCTURE was used to generate a co-ancestry matrix and tree (Fig. 2.6) showing five populations – the Umfolozi, Tugela, Mkomaas, Mbokodweni, and Umgeni. Further support for these populations was shown in PCA and FCA plots from FINERADSTRUCTURE (Fig. 2.7) and GENETIX (Fig. 2.8). The plots also display variance within populations. The Umgeni was shown to be similar to both neighbouring populations (Tugela and Mbokodweni) and was plotted between these two groups.

Positive selection may potentially affect phylogeographic analyses. Consequently, loci in this dataset were filtered for F_{ST} outliers using BAYESCAN 2.1 across the five populations using a False Discovery Rate (FDR) of 0.05 and default parameters (20 pilot runs of 5,000 iterations with a burn-in and final run of 50,000 iterations each). I identified 24 loci as potential F_{ST} outliers. However I did not

find any difference in genetic signal when comparing these to other loci, and so all loci were combined for downstream analyses. Removing *HindIII* satellite loci was found to produce similar results, leading to their retention in the final dataset.

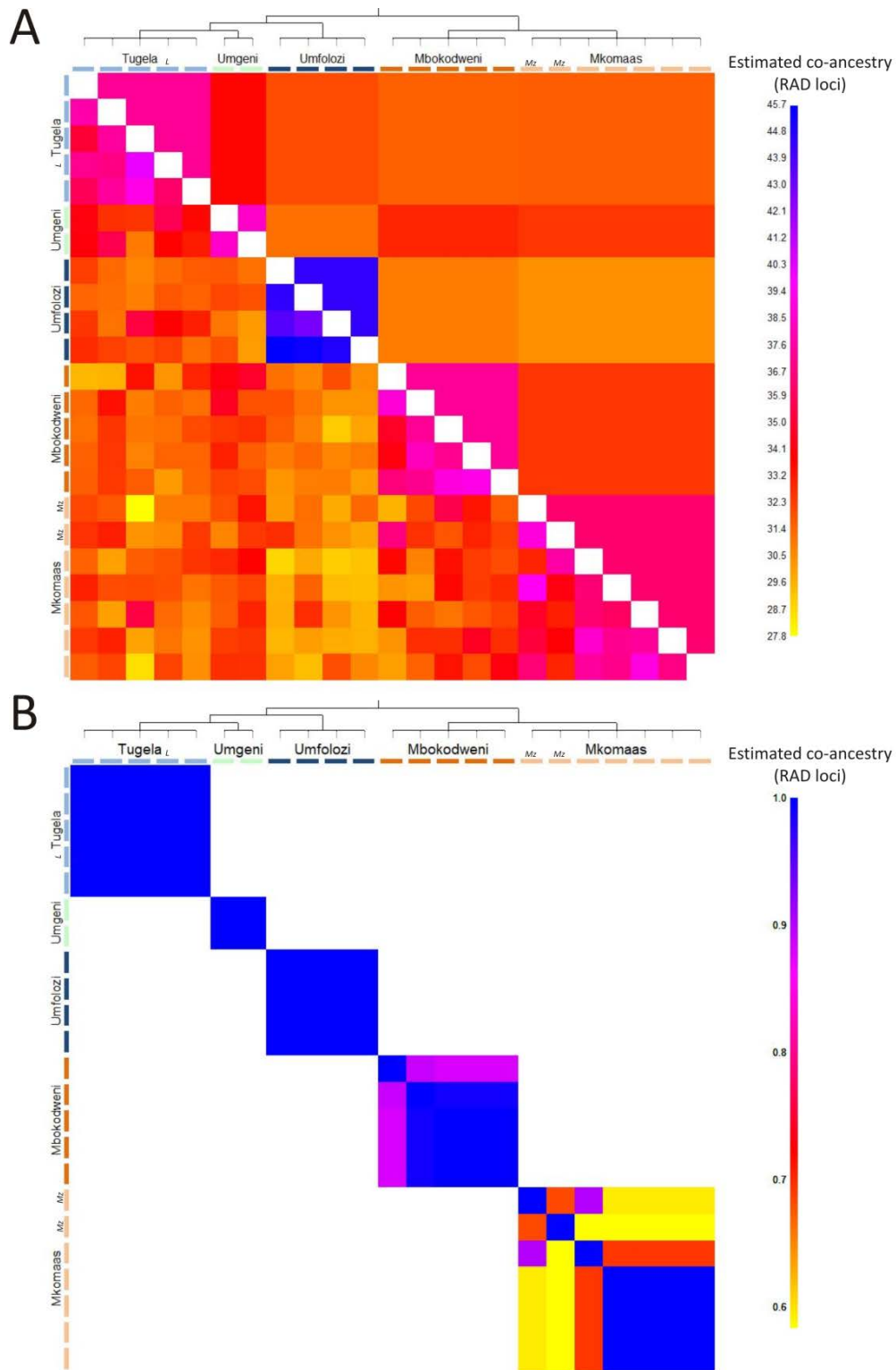


Fig. 2.6 A – FINERADSTRUCTURE co-ancestry matrix, indicating pairwise co-ancestry between individuals. The lower diagonal indicates raw copy numbers whereas the upper diagonal shows aggregated copy numbers. Individuals clustering into populations are indicated by clustering in the accompanying tree and along the diagonal of the plot. The sample labelled ‘L’ is the individual from the Lions River, whereas the samples marked ‘Mz’ are the two individuals from the Mzimkhulu. B –

MCMC pairwise comparison heat plot. Pairwise co-ancestry less than 0.6 are not shown (white blocks). Populations are broadly indicated by squares clustering together along the diagonal. Intra-population divergence is observed within the Mkomaas and Mzimkhulu lineages, as shown by pairwise comparisons not in blue (1.0 estimated co-ancestry). This is particularly obvious between the Mkomaas and Mzimkhulu Rivers (within the Mkomaas population).

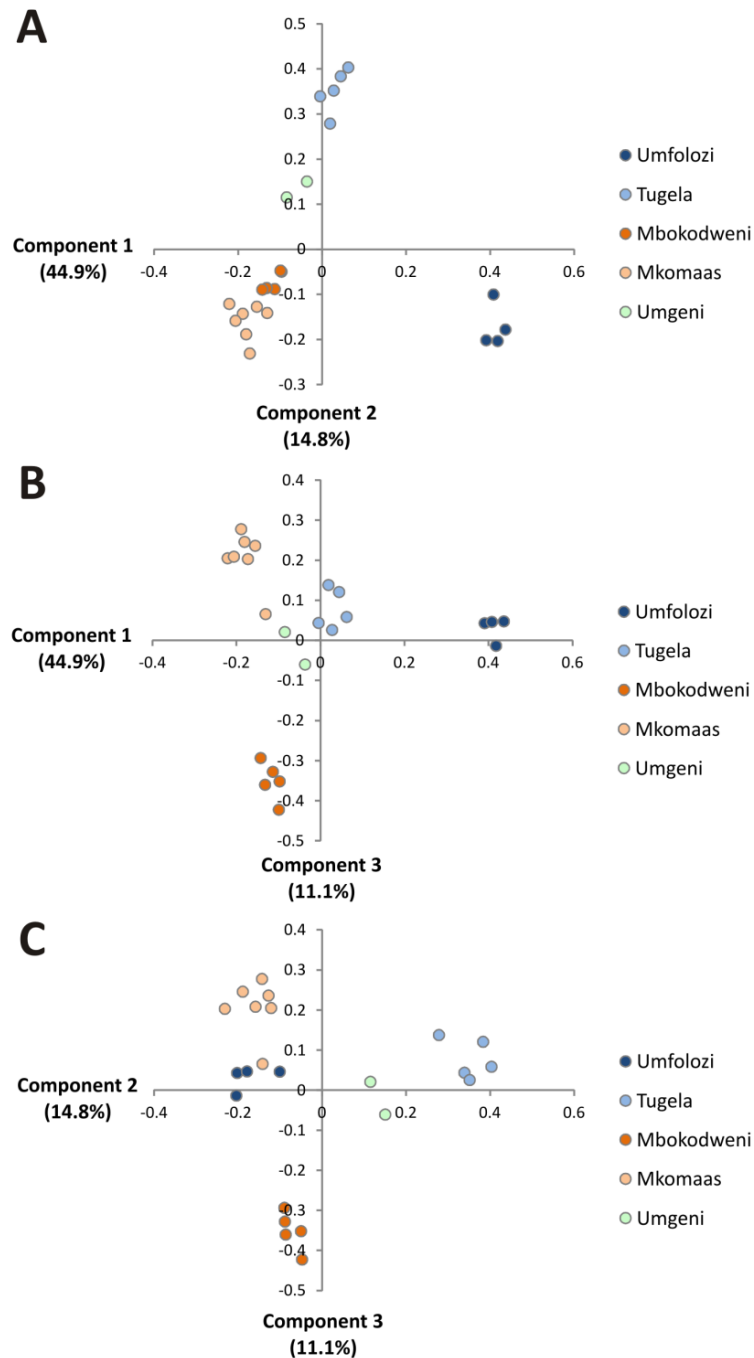


Fig. 2.7 Principal Components Analysis plots indicating the distribution of individuals into populations according to the first three principal components identified in FINERADSTRUCTURE, accounting for 70.8% of the eigenvalues. Component 1 splits Umfolozi from the rest. Component 2 splits Tugela (and by extension Umgeni) from all others. Component 3 isolates Mbokodweni from the other populations. A – The first two components which split samples into all five observed populations. The southern lineage (Mkomaas and Mbokodweni) clusters closely together. B – Components 1 and 3 split the samples into five populations. Subdivision within the five major populations is most apparent. C – Components 2 and 3 split the samples into five populations, although one sample of the Mkomaas population (KNU018) clusters closely with the Umfolozi population.

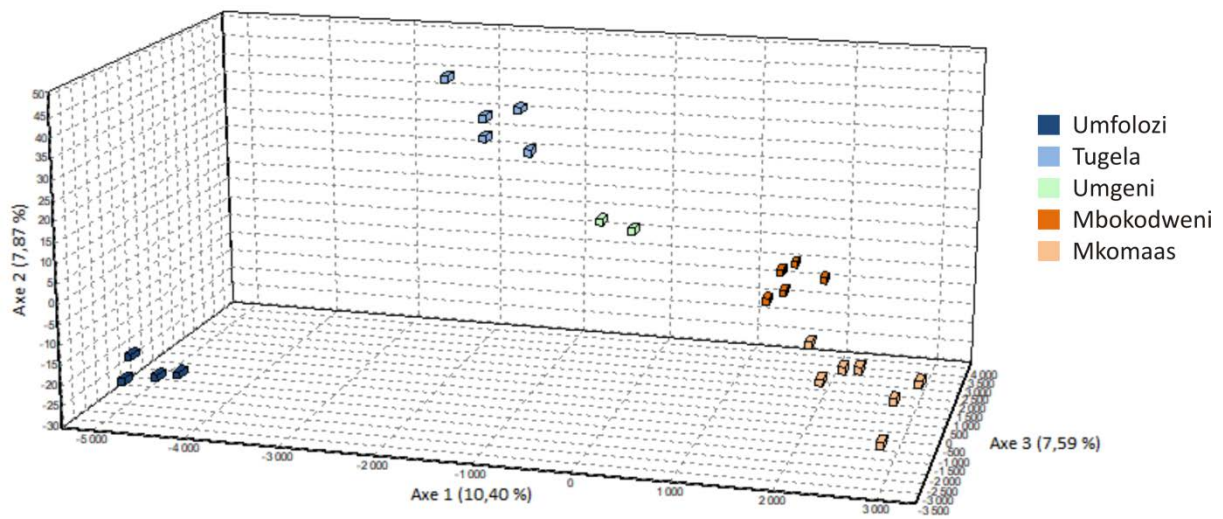


Fig. 2.8 Factorial Correspondence Analysis plot showing the clustering of individuals into the five populations according to the first three components, accounting for 25.86% of the variance in the data.

The number of private alleles varied from 35 in Umfolozi to 57 in the Mkomaas population, except Umgeni which had only 6 private alleles (Table 2.4). There was a marked increase of private alleles when the northern (excluding the potentially admixed Umgeni lineage) and southern lineages were specified (107 and 170 respectively). These results show that the northern lineages share 34 alleles which are not found in the southern lineages, and the southern lineages share 77 alleles which are not found in the north. Extending this approach I removed the potentially admixed Umgeni lineage and categorised each allele as private within a single population, shared between two or more populations only, or shared between all populations (Fig. 2.9).

Table 2.4 Summary data produced by the POPULATIONS program in STACKS for variant positions (top) and all positions (bottom). Populations are: Umf = Umfolozi; Tug = Tugela; Mko = Mkomaas; Mbo = Mbokodweni; Umg = Umgeni. Summary data were also calculated for all individuals as a single population (Total) and for the northern (North) and southern (South) lineages identified in this study, excluding the potentially admixed Umgeni lineage.

	Pop	<i>N</i>	Pvt	Sites	% Poly Sites	<i>P</i>	<i>H_O</i>	<i>H_E</i>	<i>F_{IS}</i>	π
Variant sites	Umf	3.58	35	533	59.1	0.814	0.370	0.231	-0.179	0.270
	Tug	4.21	38	639	62.4	0.817	0.364	0.232	-0.180	0.266
	Mko	6.17	57	607	71.0	0.806	0.387	0.249	-0.227	0.271
	Mbo	4.22	36	643	65.8	0.800	0.399	0.248	-0.211	0.284
	Umg	2.00	6	492	50.6	0.810	0.374	0.222	-0.118	0.295
	North	7.07	107	719	76.4	0.818	0.357	0.246	-0.186	0.267
	South	9.76	170	720	85.1	0.799	0.392	0.266	-0.234	0.282
	Total	18.5		723	100	0.807	0.373	0.270	-0.218	0.278
All sites	Umf	3.80	35	51,579	0.61	0.9981	0.0038	0.0024	-0.002	0.0028
	Tug	4.65	38	55,557	0.72	0.9979	0.0042	0.0027	-0.002	0.0031
	Mko	6.63	57	55,596	0.78	0.9979	0.0042	0.0027	-0.003	0.0030
	Mbo	4.65	36	54,846	0.77	0.9977	0.0047	0.0029	-0.003	0.0033
	Umg	2.00	6	51,058	0.49	0.9982	0.0036	0.0021	-0.001	0.0028
	North	8.09	107	57,520	0.95	0.9977	0.0045	0.0031	-0.002	0.0033
	South	11.0	170	57,600	1.06	0.9975	0.0049	0.0033	-0.003	0.0035
	Total	21.0		57,840	1.25	0.9976	0.0047	0.0034	-0.003	0.0035

N = average number of individuals genotyped at each locus; Pop = populations; Pvt = number of private alleles; Sites = number of variant sites (top) and total sites (bottom); % Poly Loci = percentage of sites found to be polymorphic; *P* = average frequency of major allele; *H_O* = average observed heterozygosity; *H_E* = average expected heterozygosity; *F_{IS}* = average Wright's inbreeding coefficient; π = mean nucleotide diversity

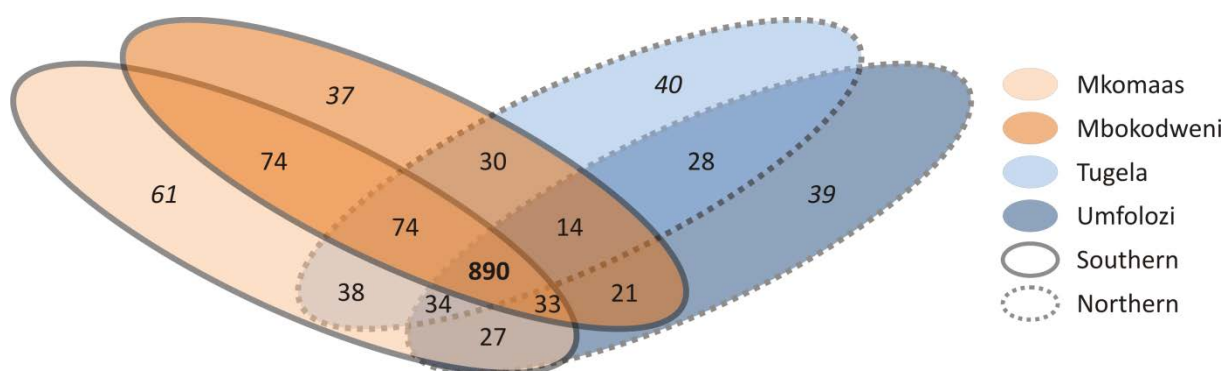


Fig. 2.9 Venn diagram illustrating the association of alleles between four populations (Mkomaas, Mbokodweni, Tugela and Umfolozi) after samples from the potentially admixed Umgeni population are removed, resulting in 720 diallelic loci. Values in italics indicate private alleles for each of the four populations, whereas the value in bold indicates alleles found in all populations.

Although the dataset comprises variable loci, only a small proportion of all nucleotide sites (1.25%) were polymorphic. Despite exclusion of singleton SNP loci the major allele frequency was relatively high across variable sites, suggesting that most loci comprise a common allele and a rare variant. Observed heterozygosity was similar across populations and higher than expected heterozygosity, resulting in negative F_{IS} values (see discussion below). Nucleotide diversity was similar across populations, varying from 0.0028 in Umgeni to 0.0033 in Mbokodweni and 0.0035 overall.

Pairwise F_{ST} values between four of the five populations (Table 2.5) were significant at $p = 0.01$. All pairwise comparisons with Umgeni yielded negative non-significant values, probably because this sample comprised two individuals with indications of admixture. The highest pairwise F_{ST} values were recorded for the Umfolozi drainage system compared to Mkomaas ($F_{ST} = 0.057$), followed by Mkomaas versus Tugela system ($F_{ST} = 0.039$) and Mbokodweni versus Tugela ($F_{ST} = 0.039$). The lowest positive pairwise F_{ST} values were recorded between the two northern populations (Umfolozi and Tugela, $F_{ST} = 0.002$) and the two southern populations (Mkomaas and Mbokodweni, $F_{ST} = 0.007$).

Table 2.5 Pairwise F_{ST} values between the five populations identified in this study

	Umfolozi	Tugela	Mkomaas	Mbokodweni
Umfolozi				
Tugela	0.002*			
Mkomaas	0.057*	0.039*		
Mbokodweni	0.032*	0.039*	0.007*	
Umgeni	-0.075	-0.096	-0.043	-0.073

* $p < 0.01$

2.3.4 Population history

The best supported scenario from DIYABC analysis involved a split into northern and southern lineages followed by subdivision into three northern and two southern populations (Fig. 2.10). Other scenarios received similar high support, such as a latitudinal series of splits, either from north to

south or vice versa, or a split into two northern and two southern populations with admixture in the Umgeni system (Supporting Information Fig. S2.6D). The effective population size is estimated to lie between 87,500 – 875,000 individuals.

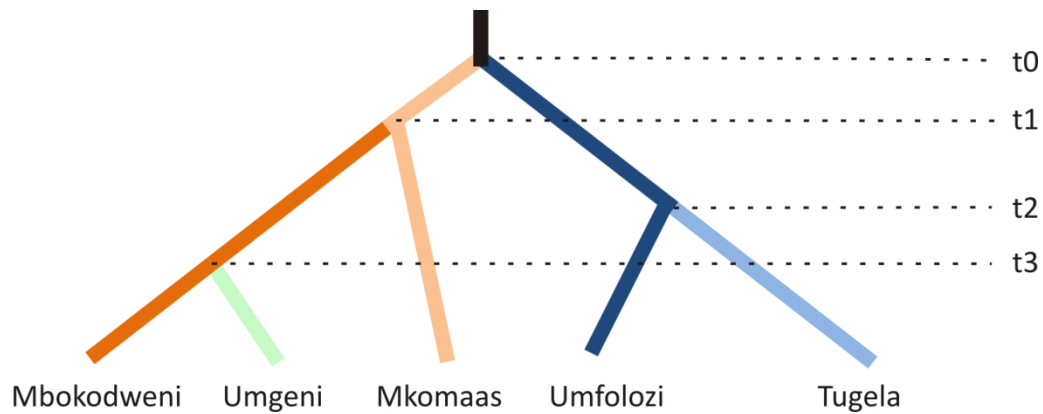


Fig. 2.10 Scenario determined as most likely by DIYABC scenario testing using the logistic regression approach. Divergence times are indicated on the right. Divergence points are not drawn to scale. The node *t0* was fixed in time as the oldest point, whereas all other nodes were allowed to vary in relation to one another.

2.4 Discussion

2.4.1 Sequencing and mapping

The average GC content of the reads obtained through RAD sequencing, between 38.5% and 40.8%, was similar to that reported for the *D. rerio* genome (38.6%) (Zhou *et al.* 2004) giving confidence that these data were not particularly biased by the choice of AT-rich restriction sites (contrary to DaCosta & Sorenson 2014; Campagna *et al.* 2015). The relatively high transition/transversion ratio of 2.04 after filtering for paralogs and mitochondrial loci may indicate a bias towards genic regions, as SNPs occur more frequently as transitions in exons than in introns (Park *et al.* 2010). This value may also reflect effective filters to reduce sequencing error in the final SNP dataset (Pujolar *et al.* 2013; Rašić *et al.* 2014; Zhang *et al.* 2015). Initial examination of the SNP data, prior to filtering, showed many SNPs in the last few base pairs of reads (Supporting Information Fig. S2.1) that may reflect

sequencing errors rather than true variants (Pujolar *et al.* 2013). Trimming removed most of these errors, as is apparent from the SNP density spectrum (Supporting Information Fig. S2.1), with further errors removed by filtering for singleton alleles and excess heterozygosity.

A general reason for the unexpectedly high genome coverage observed is the paralogous origins of most loci in *L. natalensis* when mapped against lower-ploidy species such as *C. carpio*. Since the hexaploid lineage *L. natalensis* has 150 chromosomes (Oellermann & Skelton 1990) versus 100 in the tetraploid lineage *C. carpio* (Ráb *et al.* 1989), a larger amount of the genomic information in the former species originates from shared ancestral sequence duplications. Therefore the high percentage of mapped reads seen here undoubtedly includes some error from merging of paralogs. Marginally greater coverage was observed in the *C. carpio* nuclear CDS than in the entire nuclear genome, which suggests a bias in the presence of cut sites towards coding sequences, in agreement with the observed transition/transversion ratio. However, this may also reflect a greater likelihood of mapping reads in coding regions, which are more conserved across these distantly-related taxa.

The protocol and enzymes chosen for the ddRAD method were estimated to obtain 1.60% of the *D. rerio* genome (Peterson *et al.* 2012). When the reads obtained for each *L. natalensis* individual were mapped against the *C. carpio* genome, an average of 2.6% coverage per individual was observed. In total, across individuals, reads mapped to about 10% of the *C. carpio* genome. This is despite the distant relation between taxa and the stringent mapping parameters employed to reduce spurious matches. Incomplete enzyme digestion yields loci not accounted for in the estimate of coverage from the *D. rerio* genome. However, less than 9% of sequence reads include additional target restriction sites suggesting that this is only a contributing factor. Although the same fragment size range was targeted as in Peterson *et al.* (2012), the high proportion of overlap between paired reads indicated imperfect size selection. This may be a common issue with genomic methods involving size selection, as low concentrations of non-target fragments may be favoured by biased

amplification and sequencing of short fragments. One consequence of poor enforcement of the size threshold is that a considerably larger portion of the genome was sampled.

2.4.2 Bioinformatics and SNP discovery

A paired-end ddRAD sequencing approach was utilised in this study to improve the fragment read depth by avoiding the random shearing step in single-digest RAD. This method offers the added benefits of requiring less genomic DNA, high repeatability within and among individuals, reduced library construction costs and allowing highly multiplexed libraries (Peterson *et al.* 2012). A number of studies have also shown the method to require small sample sizes to produce highly informative population-level results (Willing *et al.* 2012; Boehm *et al.* 2015; Macher *et al.* 2015).

Some predicted drawbacks of this method are the inability of the process to combine stacks of reads as longer contigs, with consequent reduced down-stream applications, and potential for bias in estimates of population parameters (Arnold *et al.* 2013; DaCosta & Sorenson 2014). In practice I did not observe these limitations, my mapping showed more extensive coverage than anticipated and this was reflected in the number of stacks retrieved.

Over 66% of the paired reads were overlapping, indicating inefficient size selection. Additionally, some of the consensus fragments were so short (< 90 bp) that sequencing had extended into the adapters and barcode at the 5' end. Both of these observations may indicate a need for the selection of a larger fragment size or a more stringent size selection step in the ddRAD protocol. Reads contaminated by adapter-pollution were filtered with custom scripts using regular expressions based on restriction enzyme recognition sequences. I both used only first read sequences and chose a single SNP per locus to avoid duplication due to overlap between paired reads. I observed a high degree of incomplete enzyme digestion, with 9% of reads containing more than one of the same restriction site.

Overall, the method was found to produce relatively low read depths at lower $-m$ thresholds (minimum read depth per stack) across the large number of sites identified, probably due to the high frequency of restriction sites for both enzymes, combined with incomplete digestion and size selection. These enzymes were selected to access many loci across the genome. Although efforts were made to restrict fragment size and to achieve complete restriction enzyme digestion, the retention of low levels of non-target fragments are common to these methods and may have disproportionate influence on the sequence data. An unanticipated by-product of these technical imperfections, combined with polyploidy of this lineage, is high genome coverage at low read depth. As is typical of many genomic studies, this resulted in a large loss of data through stringent filtering for quality control. In future ddRAD studies this could be remedied by using one rare-cutting enzyme to reduce the number of fragments sequenced, or to increase overall sequencing coverage to improve read depth per site. However, by specifying a minimum depth of five reads per stack, I was able to obtain loci of sufficient coverage ($>25x$) for population genomic analyses (Paris *et al.* 2017).

Because coverage is so important for identifying errors from sequencing variation, the substantial variance in loci observed when varying $-m$ may indicate that high read-depth cut-off should be favoured in analyses (Mastretta-Yanes *et al.* 2015). However, setting the cut-off value too high would result in allele dropout, leading to further errors (Mastretta-Yanes *et al.* 2015). Here I used the approach of Paris *et al.* (2017) to determine the optimal STACKS parameters by comparing the number of assembled loci, polymorphic loci and SNPs obtained. This allowed me to generate intuitive graphical representations of how parameters in STACKS influenced the dataset (Fig. 2.3; Supporting Information Figs. S2.4-S2.6; Supporting Information Table S2.2). From these plots I chose an optimal set of parameter values for the current dataset.

A large number of paralogous sequences were expected due to the ancestral polyploidy of *L. natalensis*. Beyond the default STACKS parameters involved in identifying and filtering potentially paralogous loci, I employed two further approaches: identifying loci with more than two haplotypes

in a sample and the HDPLLOT technique of McKinney *et al.* (2017). The first approach identified a large proportion of loci as potential paralogs, but this likely also included many loci affected by sequencing error or adapter-pollution, which would be removed in any event. In contrast, the HDPLLOT method was difficult to interpret due to the sparsity of samples (Supporting Information Fig. S2.7), but identified 463 SNPs as potential paralogs. Of these, 42 were not identified by the excess haplotype method or other filters and were retained in the final 723 loci. Although the authors of HDPLLOT compare an excess-haplotype approach to their own method (McKinney *et al.* 2017), they assessed the haplotypes at the population level, without considering individual diploid genotypes, and therefore did not make full use of these data. Here I demonstrate that removing loci with excess haplotypes yields more putative paralogous loci and may be more useful for studies with few samples of relatively low coverage. The HDPLLOT approach is likely still useful as it would identify diverged paralogs, which are fixed for alternate alleles, and should be excluded but which would not yield more than two haplotypes within a single individual (McKinney *et al.* 2017). Despite my best efforts at isolating paralogous loci, the F_{IS} output by STACKS reported negative values which are indicative of paralogous loci retained in the dataset (discussed below). This shows that paralogs are resilient to the numerous filtering approaches used here.

Criteria for SNP representation across individuals have a strong effect on the data available for analysis. Recently there has been a trend advocating the use of datasets with considerable missing data (Rubin *et al.* 2012; Buerkle & Gompert 2013; Wagner *et al.* 2013; Chattopadhyay *et al.* 2014; DaCosta & Sorenson 2014; Huang & Knowles 2014), but this has been argued against in other studies (Henning *et al.* 2014). Initially I opted for a conservative approach to minimise missing data and thereby reduce uncertainty in population analyses. However, I found that selecting low levels of missing data, at a cost of a smaller dataset, resulted in a loss of power to detect phylogeographic structure (results not shown). As the level of missing data is allowed to increase, so does the signal of phylogeographic structure. This phenomenon has been briefly described in the literature (Wagner

et al. 2013; Huang & Knowles 2014; Takahashi *et al.* 2014; Campagna *et al.* 2015) and should be viewed as a motivation to include more missing data to reduce potential biases from only examining highly conserved regions. However, my final datasets retained relatively low levels of missing data – 19.7% within the 723 SNP dataset and 0.63% within the DIYABC reduced SNP dataset of 661 markers. The difference in missing data observed between these datasets must be due to removing SNPs which contain missing data for entire populations in the latter dataset. This suggests that most missing data observed is due to mutations within one of the restriction sites leading to locus dropout at the population level (Arnold *et al.* 2013). These missing data are therefore not due to technical errors or low coverage, but real biological signal from private population mutations.

I found that some signal of finer-scale structure was being driven by mitochondrial SNPs. STRUCTURE plots generated prior to mitochondrial SNP filtering showed support for higher levels of *K*, depending on the level of missing data allowed (results not shown). This was surprising given that there were only 121 mtDNA markers out of 50,740 loci prior to other filters. This highlights the need for effective mitochondrial-marker filtering of RAD datasets, as these few SNPs influenced signal from all other SNPs. Additionally, I found that the presence of large numbers of singleton SNPs drowned out the signal of genetic differentiation, as observed previously by Rodríguez-Ezpeleta *et al.* (2016). This resulted in STRUCTURE plots with no differentiation between populations (data not shown). However, this was resolved by filtering for a minimal minor allele frequency (MAF) set to remove singleton alleles.

Cyprinid genomes include extensive repetitive regions (Henkel *et al.* 2012) such as the *HindIII* satellite (Datta *et al.* 1998; Tseng *et al.* 2008), which is prevalent in the dataset at a frequency of 4.26% prior to filtering, with 7.47% of the loci in the final dataset potentially affiliated with this satellite. The prevalence of this satellite in the data exceeds that of all satellites across the *C. carpio* genome (2.46%) (Xu *et al.* 2011). A *HindIII* satellite has been shown to exhibit intraspecific concerted evolution in *Cyprinodon variegatus*, whereby it shows low levels of variability within populations and

individuals but is distinct between local populations (Elder & Turner 1994). This is thought to occur either by genetic isolation between populations or by the propagation of new mutational variants across neighbouring populations through molecular biological processes such as biased gene conversion (Elder & Turner 1994). Similar results were recorded in *Acrossocheilus paradoxus* (Tseng *et al.* 2008) for the satellite sequence to which the *Labeobarbus* satellite matches. In agreement with this, I analysed a set of *HindIII* satellite SNPs independently to the neutral SNPs and found similar results of genetic structure between the populations identified in this study (data not shown). I similarly tested whether the 24 loci potentially under selection identified through BAYESCAN influenced the STRUCTURE results, but found that I obtained the same results whether I excluded these loci or not. As a result, both satellite loci and loci potentially under selection were retained in the final dataset.

2.4.3 Population genetic parameters and structure

Differentiation between groups using STRUCTURE at $K = 2$ identified a divide between the northern and southern populations. The split between lineages appears to have occurred around Durban. Although this does not coincide with any well recognised biogeographic or climatic boundary it is consistent with a general transition from a speciose tropical fauna, to a highly endemic warm temperate fauna in aquatic organisms (Seymour *et al.* 2001; Alexander *et al.* 2004; Perera *et al.* 2011). Further subdivision into five populations broadly follows the division of KwaZulu-Natal into the aquatic biogeographic regions of Zululand (Umfolozi), Tugela, Umgeni and Mzimkhulu (including Mkomaas) (Rivers-Moore *et al.* 2007), with the fifth population (Mbokodweni) being more unexpected. Similar divisions within lineages to the biogeographic regions have been observed in freshwater crabs (Gouws *et al.* 2015) and in vertebrate fauna overall (Perera *et al.* 2011).

The division into two broad lineages dominated my primary analysis in STRUCTURE and subdivision into the five populations is not as well supported. However, further investigation using hierarchical STRUCTURE and FINERADSTRUCTURE revealed additional structure. Despite variation among individuals, PCA and independently-generated FCA plots also clearly grouped these into five populations, mainly delimited by river systems.

The two individuals from the lower Umgeni drainage system appeared as potentially admixed samples, which may indicate ancestral contact in this system between northern and southern lineages. All F_{ST} values for comparisons with Umgeni were negative, suggesting that variance within Umgeni was greater than that between Umgeni and other populations. Unfortunately only two samples were available from the lower Umgeni, which would affect permutation tests of this result. The Umgeni samples grouped between the neighbouring Tugela and Mbokodweni populations in every spatial analysis performed (Fig. 2.7-2.8) which again suggests admixture.

The single sample from the upper Umgeni system (Lions River) consistently grouped closely with those from the neighbouring Tugela system (Fig. 2.4-2.8). This is likely a translocated individual from the Mooi River, a tributary of the Tugela, via the Mooi-Mgeni Transfer Scheme (MMTS), an interbasin connection in continuous operation since 2003. I therefore included this upper Umgeni sample in the Tugela population. Similarly, two samples from the Mzimkhulu River near the southernmost point of the distribution were found to consistently cluster with the neighbouring Mkomaas River samples as a single lineage, despite grouping as a distinct mtDNA lineage (Bloomer *et al.* In prep). These samples do not form a distinct group but are responsible for much of the variation within the Mkomaas population, as demonstrated in the FINERADSTRUCTURE (Fig. 2.6-2.7) results.

I was not able to include samples from the southernmost limit of distribution, the Mtamvuna River. Similarly, the Mbokodweni River has two neighbouring river systems of similar size, the Umlazi and Illovo Rivers which were unsampled but may represent additional range for the unique lineage

identified currently solely from the Mbokodweni. Further sampling throughout the Umgeni system should be a priority to determine the effect of the ongoing transfer scheme on the upper and lower systems. Other unsampled river systems, including the Umvoti, Amatiqulu and Mhlathuze, are lower priorities for analysis as these medium sized systems are flanked by the larger Tugela and Umfolozi.

The private alleles identified here are a useful resource to distinguish populations. In addition, many alleles shared among the northern populations (excluding Umgeni) were not found in the southern lineages (34 alleles) and vice versa (77 alleles). Further investigation of private alleles after removing the potentially admixed Umgeni lineage revealed a split between the northern and southern lineages and the four populations. The Umfolozi lineage is clearly the most divergent according to the association of alleles between populations as it shares the least out of all populations. F_{ST} values also support the deeper split between the northern and southern populations, as pairwise comparisons between groups were consistently higher than those within these groups. Although all efforts were made to remove paralogs from analyses, negative F_{IS} values within all populations suggest that some remained in the final dataset. Closely similar paralogs would be combined as allelic variants in STACKS, resulting in excess heterozygosity. Where nucleotide substitutions result in consistent differences among paralogs these combined loci would be excluded by the heterozygosity filter. Difficulties distinguishing interlocus from allelic variation are expected in polyploid species, resulting in excess heterozygosity (Soltis & Soltis 2000) due to the additional paralogs present.

The identification of five different populations in this study contrasts against the six haplogroups identified using mitochondrial data (Bloomer *et al.* In prep). This could be due to retention of ancestral polymorphisms at nuclear loci whereas the lower effective population size of the mitochondrial genome would allow population variation fixation at a more rapid rate. Alternatively, this level of fine-scale differentiation may be beyond the current approach where major historical events could be masking signal from more recent or smaller-scale events, such as in the case of the

primary STRUCTURE analysis where $K = 2$ was determined to be most likely. However, the most likely explanation for this incongruence is that there is gene flow occurring between these putatively isolated populations such that the finer-scale nuclear structure within the northern and southern lineages is not substantial, but the mitochondrial locus is rapidly fixed within local populations and reflects a larger degree of difference between these populations. This may suggest a flaw with popular methods such as DNA barcoding, where the sequencing of mitochondrial genes is solely used to delimit populations or even species. Further investigation is necessary to determine whether this is the case here.

2.4.4 Population history

Similar levels of support were received for three models in scenario testing with DIYABC. The model with the most support matches the results previously observed in other analyses, where an ancestral population diverged into the northern and southern lineages, which then underwent further subdivision into the current five populations. The ancestral lineage was likely located in Zululand (Umfolozi drainage system), although I cannot rule out the possibility of an Mkomaas ancestral group (Supporting Information Fig. S2.6).

The N_e produced here is an estimate assuming random mating and constant population size, and is associated with the long-term population size, not the contemporary census size. Parameters of population history were not estimated in DIYABC due to uncertainty of the priors. Because *L. natalensis* is distributed throughout most rivers of KwaZulu-Natal, it was assumed that the effective population size and hence nucleotide diversity would be correlated with the geographic size of the drainage systems or with the number of associated rivers in a system. This was reflected in the nucleotide diversities for the Umfolozi, Tugela and Mkomaas systems, however, the Mbokodweni population had nucleotide diversity even greater than the larger systems of Tugela and Mkomaas

despite its restricted range of a single known river. Because this is a reflection of the long-term effective population size, this may indicate that this population historically occupied a more widespread range, or could suggest that I have not yet found the full extent of the distribution of this lineage across the coastal rivers in this area. This highlights the need for more comprehensive sampling across this area to define the current range of this population. The current restricted distribution of this population as shown in this study places it as a priority for conservation purposes.

2.5 Conclusion

In this study I used the ddRAD sequencing approach to reduce the genome complexity of a South African endemic hexaploid fish. My SNP identification protocol was optimised using the new approach of Paris *et al.* (2017) and extended by comparing two different approaches for paralog identification and removal, filtering of mitochondrial loci which influenced STRUCTURE results, filtering of singleton alleles which masked genetic structure, and retaining satellite loci and loci potentially under selection which both showed similar results to putatively neutral loci. Although a moderate level of missing data was observed, I show that this was likely to be due to locus dropout caused by lineage-specific mutations in one of the two restriction sites. I used the final dataset of 723 SNPs to characterise that two major lineages – northern and southern – diverge into five regional populations – Umfolozi, Tugela, Umgeni, Mbokodweni and Mkomaas – across the distribution. I found some evidence for north-south admixture within the Umgeni and translocation of a Tugela sample into the Umgeni. Private alleles were identified which support the proposed relationship between the populations. Disparity between previous mitochondrial results and the results presented in this study is likely explained by gene flow between populations. Approximate Bayesian Computation testing suggested a scenario of divergence within the northern and southern lineages into the five current populations. Finally, a number of population genetic parameters are

provided in this study including the first estimate of long-term effective population size and genetic diversity. These indices indicate that the Mbokodweni population may be a key target for conservation efforts. The approaches I utilised together with the resources established in this study will aid in combating the dearth of genetic data available for *Labeobarbus* and other cyprinids.

2.6 Acknowledgements

This study is dedicated to the memory of our colleague, the late Rob Karssing. I would like to thank the following colleagues for contributing to the sampling in this study: M Nkosi, N Rivers-Moore, J Craigie, H Plank, HE Filter, T Wilkinson, R Arderne, J Wakelin and A Howell. I would also like to thank Vincent Savolainen, Bengt Hansson, Ulrich Schliewen, Emmanuel Vreven, Alexander Papadopoulos, Ilkser Erdem Kiper and their research groups for their contributions to this project. Thank you also to two anonymous reviewers who provided invaluable suggestions and guidance. This project was funded by the South African National Research Foundation (NRF; Innovation doctoral scholarship awarded to CSS and Incentive funding to PB, grant number 77240) and the University of Pretoria's Genomics Research Institute (GRI). Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.

2.7 References

- Alexander G, Harrison J, Fairbanks D, Navarro R (2004) Biogeography of the frogs of South Africa, Lesotho and Swaziland. In: *Atlas and Red Data Book of the Frogs of South Africa, Lesotho and Swaziland. SI/MAB Series #9* (eds. Minter LR, Burger M, Harrison J, Braack HH, Bishop PJ), pp. 31-47. Smithsonian Institution, Washington, DC.
- Altschul SF, Madden TL, Schaffer AA, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402.
- Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: <<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>>

- Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology* **22**, 3179-3190.
- Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, 1-7.
- Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2004) GENETIX 4.05, logiciel sous Windows pour la génétique des populations. *Laboratoire génome, populations, interactions, CNRS UMR 5171, Université Montpellier II, Montpellier*, 1996-2004.
- Bloomer P, Bills IR, van der Bank FH, *et al.* (2007) Multidisciplinary investigation of differences and potential hybridization between two yellowfish species *Labeobarbus kimberleyensis* and *L. aeneus* from the Orange-Vaal system. Follow-up study 2004-2007. Yellowfish Working Group.
- Bloomer P, Stobie CS, Reid K, Oosthuizen CJ (In prep) Divergent mitochondrial lineages in the KwaZulu-Natal yellowfish (*Labeobarbus natalensis*) across ancient river systems. Target journal: *BMC Evolutionary Biology*.
- Boehm J, Waldman J, Robinson JD, Hickerson MJ (2015) Population genomics reveals seahorses (*Hippocampus erectus*) of the western mid-Atlantic Coast to be residents rather than vagrants. *PLoS ONE* **10**, e0116219.
- Brawand D, Wagner CE, Li YI, *et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513**, 375-381.
- Brieuc MSO, Waters CD, Seeb JE, Naish KA (2014) A dense linkage map for Chinook salmon (*Oncorhynchus tshawytscha*) reveals variable chromosomal divergence after an ancestral whole genome duplication event. *G3: Genes, Genomes, Genetics* **4**, 447-460.
- Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution* **18**, 249-256.
- Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology* **22**, 3028-3035.
- Cambray J, Bills R, Chakona A, Coetzer W, Weyl O (2017) *Labeobarbus natalensis*. The IUCN Red List of Threatened Species 2017: e.T63294A100168851. Downloaded on **14 December 2017**. <<http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T63294A100168851.en>>
- Campagna L, Gronau I, Silveira LF, Siepel A, Lovette IJ (2015) Distinguishing noise from signal in patterns of genomic divergence in a highly polymorphic avian radiation. *Molecular Ecology* **24**, 4238-4251.
- Catchen JM, Amores A, Hohenlohe PA, Cresko WA, Postlethwait JH (2011) Stacks: Building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1**, 171-182.
- Catchen JM, Hohenlohe P, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.
- Chakona A, Malherbe WS, Gouws G, Swartz ER (2015) Deep genetic divergence between geographically isolated populations of the goldie barb (*Barbus pallidus*) in South Africa: potential taxonomic and conservation implications. *African Zoology* **50**, 5-10.
- Chakona A, Skelton PH (2017) A review of the *Pseudobarbus afer* (Peters, 1864) species complex (Teleostei, Cyprinidae) in the eastern Cape Fold Ecoregion of South Africa. *ZooKeys* **657**, 109-140.
- Chakona A, Swartz ER, Gouws G (2013) Evolutionary drivers of diversification and distribution of a southern temperate stream fish assemblage: testing the role of historical isolation and spatial range expansion. *PLoS ONE* **8**, e70953.
- Chattopadhyay B, Garg KM, Ramakrishnan U (2014) Effect of diversity and missing data on genetic assignment with RAD-Seq markers. *BMC Research Notes* **7**, 841.

- Cornuet J-M, Pudlo P, Veyssier J, *et al.* (2014) DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics* **30**, 1187-1189.
- DaCosta JM, Sorenson MD (2014) Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE* **9**, e106713.
- Datta U, Dutta P, Mandal RK (1988) Cloning and characterization of a highly repetitive fish nucleotide sequence. *Gene* **62**, 331-336.
- Davey JW, Cezard T, Fuentes-Utrilla P, *et al.* (2013) Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology* **22**, 3151-3164.
- Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.
- de Graaf M, Nagelkerke LAJ, Palstra AP, Sibbing FA (2010) Experimental evidence for the biological species status in Lake Tana's *Labeobarbus* flock (Cyprinidae). *Animal Biology* **60**, 183-193.
- Durand JD, Tsigenopoulos CS, Ünlü E, Berrebi P (2002) Phylogeny and biogeography of the family Cyprinidae in the Middle East inferred from cytochrome *b* DNA— evolutionary significance of this region. *Molecular Phylogenetics and Evolution* **22**, 91-100.
- Earl DA, VonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetic Resources* **4**, 359-361.
- Elder JF, Turner BJ (1994) Concerted evolution at the population level: pupfish *HindIII* satellite DNA sequences. *Proceedings of the National Academy of Sciences* **91**, 994-998.
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.
- Everett MV, Grau ED, Seeb JE (2011) Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources* **11**, 93-108.
- Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.
- Ferchaud AL, Hansen MM (2016) The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: three-spine sticklebacks in divergent environments. *Molecular Ecology* **25**, 238-259.
- Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics* **180**, 977-993.
- Gagnaire P-A, Normandeau E, Pavay SA, Bernatchez L (2013) Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (*Coregonus clupeaformis*). *Molecular Ecology* **22**, 3036-3048.
- Gonen S, Bishop SC, Houston RD (2015) Exploring the utility of cross-laboratory RAD-sequencing datasets for phylogenetic analysis. *BMC Research Notes* **8**, 299.
- Gouws G, Peer N, Perissinotto R (2015) MtDNA lineage diversity of a potamonautid freshwater crab in KwaZulu-Natal, South Africa. *Koedoe: African Protected Area Conservation and Science* **57**, 1-12.
- Hand BK, Hether TD, Kovach RP, *et al.* (2015) Genomics and introgression: discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Current Zoology* **61**, 146-154.
- Henkel CV, Dirks RP, Jansen HJ, *et al.* (2012) Comparison of the exomes of common carp (*Cyprinus carpio*) and zebrafish (*Danio rerio*). *Zebrafish* **9**, 59-67.
- Henning F, Lee HJ, Franchini P, Meyer A (2014) Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using dense linkage mapping in non-model organisms. *Molecular Ecology* **23**, 5224-5240.

- Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* **11**, 117-122.
- Hohenlohe PA, Day MD, Amish SJ, *et al.* (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology* **22**, 3002-3013.
- Houston RD, Davey JW, Bishop SC, *et al.* (2012) Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. *BMC Genomics* **13**, 244.
- Huang H, Knowles LL (2014) Unforeseen consequences of excluding missing data from Next-Generation Sequences: simulation study of RAD sequences. *Systematic Biology* **65**, 357-365.
- Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.
- Kai W, Nomura K, Fujiwara A, *et al.* (2014) A ddRAD-based genetic map and its integration with the genome assembly of Japanese eel (*Anguilla japonica*) provides insights into genome evolution after the teleost-specific genome duplication. *BMC Genomics* **15**, 233.
- Karssing RJ (2008) Status of the KwaZulu-Natal yellowfish *Labeobarbus natalensis* (Castelnau, 1861). In: *Technical Report on the State of Yellowfishes in South Africa 2007* (eds. Impson ND, Bills IR, Wolhuter L), pp. 31-45. Water Research Commission: Gezina, Pretoria.
- Lamer JT, Sass GG, Boone JQ, *et al.* (2014) Restriction site-associated DNA sequencing generates high-quality single nucleotide polymorphisms for assessing hybridization between bighead and silver carp in the United States and China. *Molecular Ecology Resources* **14**, 79-86.
- Larson WA, Seeb LW, Everett MV, *et al.* (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications* **7**, 355-369.
- Lischer H, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298-299.
- Macher JN, Rozenberg A, Pauls SU, *et al.* (2015) Assessing the phylogeographic history of the montane caddisfly *Thremma gallicum* using mitochondrial and restriction-site-associated DNA (RAD) markers. *Ecology and Evolution* **5**, 648-662.
- Machordom A, Doadrio I (2001) Evidence of a Cenozoic Betic-Kabilian connection based on freshwater fish phylogeography (*Luciobarbus*, Cyprinidae). *Molecular Phylogenetics and Evolution* **18**, 252-263.
- Malinsky M, Trucchi E, Lawson D, Falush D (2016) RADpainter and fineRADstructure: population inference from RADseq data. *bioRxiv*, 057711.
- Mastretta-Yanes A, Arrigo N, Alvarez N, *et al.* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology Resources* **15**, 28-41.
- McKinney GJ, Waples RK, Seeb LW, Seeb JE (2017) Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources* **17**, 656-669.
- Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* **17**, 240-248.
- Naran D, Skelton PH, Villet MH (2007) Karyology of the redfin minnows, genus *Pseudobarbus* Smith, 1841 (Teleostei: Cyprinidae): one of the evolutionarily tetraploid lineages of South African barbines. *African Zoology* **41**, 178-182.

- Nel J, Murray K, Maherry A, *et al.* (2011) Technical Report for the National Freshwater Ecosystem Priority Areas project. WRC Report 1802/2/11.
- Oellermann LK, Skelton PH (1990) Hexaploidy in yellowfish species (*Barbus*, Pisces, Cyprinidae) from southern Africa. *Journal of Fish Biology* **37**, 105-115.
- Ogden R, Gharbi K, Mugue N, *et al.* (2013) Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Molecular Ecology* **22**, 3112-3123.
- Palti Y, Gao G, Miller MR, *et al.* (2014) A resource of single-nucleotide polymorphisms for rainbow trout generated by restriction-site associated DNA sequencing of doubled haploids. *Molecular Ecology Resources* **14**, 588-596.
- Palumbi SR (2003) Population genetics, demographic connectivity, and the design of marine reserves. *Ecological Applications* **13**, 146-158.
- Paris JR, Stevens JR, Catchen JM (2017) Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution* **8**, 1360-1373.
- Park S, Yu H-J, Mun J-H, Lee S-C (2010) Genome-wide discovery of DNA polymorphism in *Brassica rapa*. *Molecular Genetics and Genomics* **283**, 135-145.
- Partridge T, Maud R (2000) Macro-scale geomorphic evolution of southern Africa. *Oxford Monographs on Geology and Geophysics* **40**, 3-18.
- Perera SJ, Ratnayake-Perera D, Proches S (2011) Vertebrate distributions indicate a greater Maputaland-Pondoland-Albany region of endemism. *South African Journal of Science* **107**, 1-15.
- Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7**, e37135.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.
- Pujolar JM, Jacobsen M, Als TD, *et al.* (2014) Assessing patterns of hybridization between North Atlantic eels using diagnostic single-nucleotide polymorphisms. *Heredity* **112**, 627-637.
- Pujolar JM, Jacobsen MW, Frydenberg J, *et al.* (2013) A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel. *Molecular Ecology Resources* **13**, 706-714.
- Ráb P, Pokorný J, Roth P (1989) Chromosome studies of the common carp, *Cyprinus carpio*. I. Karyotype of Amurian carp, *C. carpio haematopterus*. *Caryologia* **42**, 27-36.
- Rašić G, Filipović I, Weeks AR, Hoffmann AA (2014) Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, *Aedes aegypti*. *BMC Genomics* **15**, 275.
- Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM (2013) Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology* **22**, 2953-2970.
- Rivers-Moore N, Goodman P, Nkosi M (2007) An assessment of the freshwater natural capital in KwaZulu-Natal for conservation planning. *Water SA* **33**, 665-674.
- Rodríguez-Ezpeleta N, Bradbury IR, Mendibil I, *et al.* (2016) Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection. *Molecular Ecology Resources* **16**, 991-1001.
- Roesti M, Salzburger W, Berner D (2012) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology* **12**, 94.
- Rubin BER, Ree RH, Moreau CS (2012) Inferring phylogenies from RAD sequence data. *PLoS ONE* **7**, e33394.
- Sambrook J, Fritsch E, Maniatis T (1989) Extraction with phenol: chloroform. *Molecular cloning: A laboratory manual* **3**.

- Seymour CL, De Klerk HM, Channing A, Crowe TM (2001) The biogeography of the Anura of sub-equatorial Africa and the prioritisation of areas for their conservation. *Biodiversity and Conservation* **10**, 2045-2076.
- Skelton PH (1986) Fish of the Orange-Vaal system. In: *The ecology of river systems* (eds. Davies BR, Walker KF), pp. 143-162. Springer: Netherlands.
- Skelton P, Bills R (2008) An introduction to African yellowfish and to this report. In: *Technical Report on the State of Yellowfishes in South Africa 2007* (eds. Impson ND, Bills IR, Wolhuter L), pp. 1-14. Water Research Commission: Pretoria.
- Smith SA, Bermingham E (2005) The biogeography of lower Mesoamerican freshwater fishes. *Journal of Biogeography* **32**, 1835-1854.
- Soltis PS, Soltis DE (2000) The role of genetic and genomic attributes in the success of polyploids. *Proceedings of the National Academy of Science of the United States of America* **97**, 7051-7057.
- Swartz ER, Chakona A, Skelton PH, Bloomer P (2014) The genetic legacy of lower sea levels: does the confluence of rivers during the last glacial maximum explain the contemporary distribution of a primary freshwater fish (*Pseudobarbus burchelli*, Cyprinidae) across isolated river systems? *Hydrobiologia* **726**, 109-121.
- Swartz ER, Mwale M, Hanner R (2008) A role for barcoding in the study of African fish diversity and conservation. *South African Journal of Science* **104**, 293-298.
- Swartz ER, Skelton P, Bloomer P (2007) Sea-level changes, river capture and the evolution of populations of the Eastern Cape and fiery redfins (*Pseudobarbus afer* and *Pseudobarbus phlegethon*, Cyprinidae) across multiple river systems in South Africa. *Journal of Biogeography* **34**, 2086-2099.
- Swartz ER, Skelton PH, Bloomer P (2009) Phylogeny and biogeography of the genus *Pseudobarbus* (Cyprinidae): Shedding light on the drainage history of rivers associated with the Cape Floristic Region. *Molecular Phylogenetics and Evolution* **51** 75-84.
- Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437-460.
- Takahashi T, Nagata N, Sota T (2014) Application of RAD-based phylogenetics to complex relationships among variously related taxa in a species flock. *Molecular Phylogenetics and Evolution* **80**, 137-144.
- Tseng M-C, Chiang T-Y, Wang J-P (2008) Characterization and genetic variations of satellite DNAs in *Acrossocheilus paradoxus* (Günther, 1868) (Cyprinidae) indicate population expansion. *Journal of Fish Biology* **72**, 1138-1153.
- Tsigenopoulos CS, Kasapidis P, Berrebi P (2010) Phylogenetic relationships of hexaploid large-sized barbs (genus *Labeobarbus*, Cyprinidae) based on mtDNA data. *Molecular Phylogenetics and Evolution* **56** 851-856.
- van der Walt K-A, Swartz ER, Woodford D, Weyl O (2017) Using genetics to prioritise headwater stream fish populations of the Marico barb, *Enteromius motebensis* Steindachner 1894, for conservation action. *Koedoe* **59**, 1-7.
- Vreven EJ, Musschoot T, Snoeks J, Schliewen UK (2016) The African hexaploid Torini (Cypriniformes: Cyprinidae): review of a tumultuous history. *Zoological Journal of the Linnean Society* **177**, 231-305.
- Wagner CE, Keller I, Wittwer S, et al. (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* **22**, 787-798.
- Waples R, Seeb L, Seeb J (2016) Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Molecular Ecology Resources* **16**, 17-28.

- Willing E-M, Dreyer C, Van Oosterhout C (2012) Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. *PLoS ONE* **7**, e42649.
- Xu P, Li J, Li Y, et al. (2011) Genomic insight into the common carp (*Cyprinus carpio*) genome by sequencing analysis of BAC-end sequences. *BMC Genomics* **12**, 188.
- Yang L, Sado T, Hirt MV, et al. (2015) Phylogeny and polyploidy: resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). *Molecular Phylogenetics and Evolution* **85**, 97-116.
- Yoshizawa M, Robinson BG, Duboué ER, et al. (2015) Distinct genetic architecture underlies the emergence of sleep loss and prey-seeking behavior in the Mexican cavefish. *BMC biology* **13**, 1.
- Zhang B-D, Xue D-X, Wang J, et al. (2015) Development and preliminary evaluation of a genomewide single nucleotide polymorphisms resource generated by RAD-seq for the small yellow croaker (*Larimichthys polyactis*). *Molecular Ecology Resources* **16**, 755-768.
- Zhou Y, Bizzaro JW, Marx KA (2004) Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+C)% composition. *BMC Genomics* **145**, 95.

2.8 Supporting Information

Table S2.1 Sampling locations of *L. natalensis* used in this study. Each of the samples listed yielded DNA of a quality satisfactory for RAD sequencing. Sample numbers match those indicated in the STRUCTURE results.

Geographic location	Drainage system	Coordinates	Year	Samples	Sample no.
Black Umfolozi River	Umfolozi	-27.93852 S 31.20945 E	2007	BLA004; BLU003	1, 2
Mkuze River, Farm Verdrukt (444)	Mkuze / Umfolozi	-27.68 S 31.44 E	2003	KZF223	3
White Umfolozi River, Lenjanedrif, Vryheid	Umfolozi	-27.93503 S 30.95519 E	2007	WUR008	4
Buffalo River, Tayside	Buffalo / Tugela	-28.05957 S 30.37355 E	2007	BUF002; BUF008	5, 6
Mooi River, Winterhoek Weir	Tugela	-29.04956 S 30.30382 E	2006	MOO007	7
Tugela River, Hlalanathi, below DWAF gauging weir	Tugela	-28.65275 S 29.04352 E	2006	TUG002	8
Lions River	Tugela	-29.42 S 30.01 E	2003	KNT212	9
Mdloti River, North Coast, Durban	Umgeni	-29.60266 S 31.00874 E	2006	MDL003	10
Molweni River, Durban Metro	Umgeni	-29.75176 S 30.88918 E	2006	MOL004	11
Mbokodweni River, Durban Metro	Mbokodweni	-30.00925 S 30.89019 E	2006	MBO007; MBO008; MBO009; MBO010; MBO011	12, 13, 14, 15, 16
Mkomaas, below Josephine's Bridge	Mkomaas	-30.01 S 30.15 E	2003	KNK315	17
Mkomazi River	Mkomaas	-30.13815 S 30.67455 E	2006	MKO001; MKO002; MKO003; MKO004	18, 19, 20, 21
Mzimkhulu, 2 km above Polela junction	Mzimkhulu / Mkomaas	-29.96 S 29.69 E	2003	KNU018; KNB017	22, 23

Table S2.2 Statistics generated as $-n$ is varied from $-n = -M - 1$ to $-n = -M + 1$ using the script COUNT_FIXED_SNPS.PY and the populations module of STACKS with a SNP sample representation cut-off of 80% ($-r = 0.8$), as per Paris *et al.* (2017). I followed Paris *et al.*'s (2017) recommendation in selecting the parameter set which yielded the greatest number of polymorphic loci ($-n = 0$).

count_fixed_snps.py		$m5M1n0$ ($-n = -M - 1$)	$m5M1n1$ ($-n = -M$)	$m5M1n2$ ($-n = -M + 1$)
Heterozygous SNPs per sample	BLA004	2059	2053	2033
	BUF008	1853	1863	1857
	MOL004	1001	1029	1069
	MBO009	673	700	725
	KNK315	711	695	688
	MBO008	849	853	870
	MOO007	665	682	699
	TUG002	564	578	598
	MBO007	285	289	299
	KNB017	2637	2568	2544
	BUF002	665	634	605
	KNT212	324	318	311
	KNA017	103	94	93
	WUR008	407	407	411
	MDL003	465	460	467
	KZF223	390	386	388
	MKO001	817	812	802
	BLU003	383	371	366
	MKO003	416	419	417
	BLB004	418	403	387
	KNU018	382	379	382
	MBO010	331	337	330
	MBO011	497	500	494
	MKO002	322	323	320
	MKO004	462	468	462
	Variable sites across population	17679	17621	17617
	Heterozygous SNPs in the catalog	17679	18980	21471
	SNPs across all samples and in catalog	17679	18634	20617
	Fixed SNPs only in the catalog	0	1013	3000
Populations ($-r = 0.8$)	Number of loci	2006	1982	1759
	Number of polymorphic loci	1198	1185	1070
	Number of SNPs	1923	1910	1645

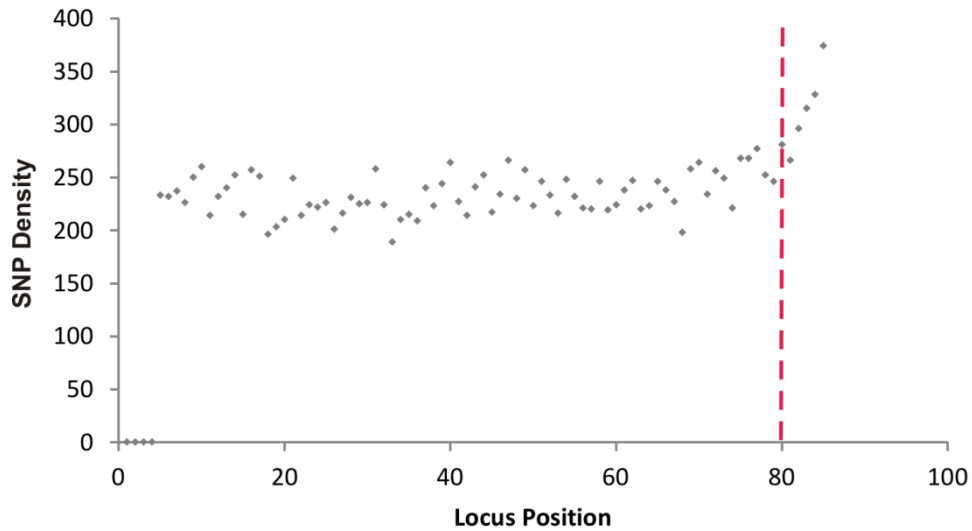


Fig. S2.1 SNP density plot for 85-bp Read 1 fragments after initial processing in STACKS using the parameter set $-m\ 5\ -M\ 2\ -n\ 1$. The initial uniform distribution suggests low levels of sequencing error, whereas the slight increase in SNPs towards the end of the loci is indicative of sequencing error. Trimming to a length of 80-bp (indicated by the dashed line) removes most of these errors.

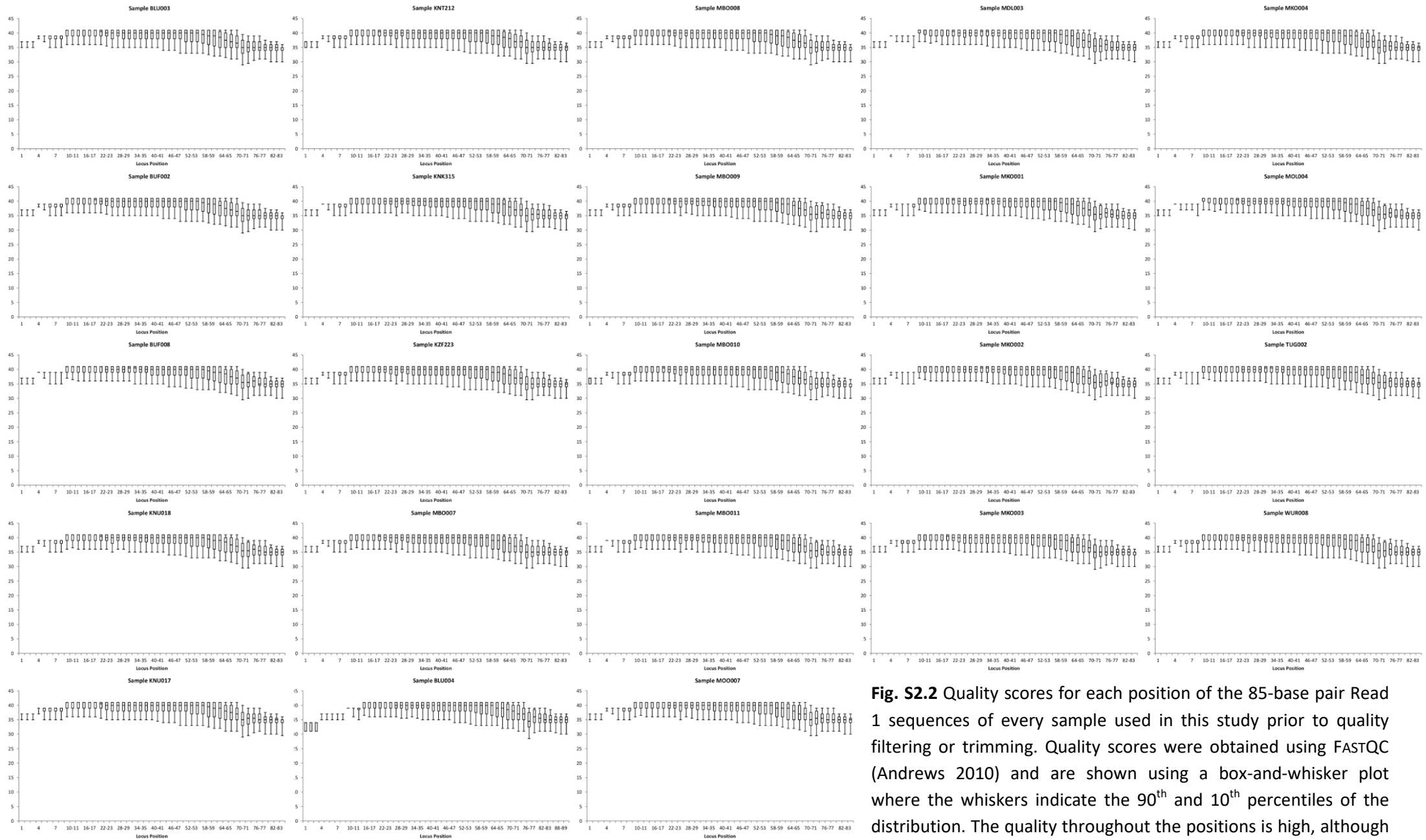


Fig. S2.2 Quality scores for each position of the 85-base pair Read 1 sequences of every sample used in this study prior to quality filtering or trimming. Quality scores were obtained using FASTQC (Andrews 2010) and are shown using a box-and-whisker plot where the whiskers indicate the 90th and 10th percentiles of the distribution. The quality throughout the positions is high, although this deteriorates towards the end of the reads.

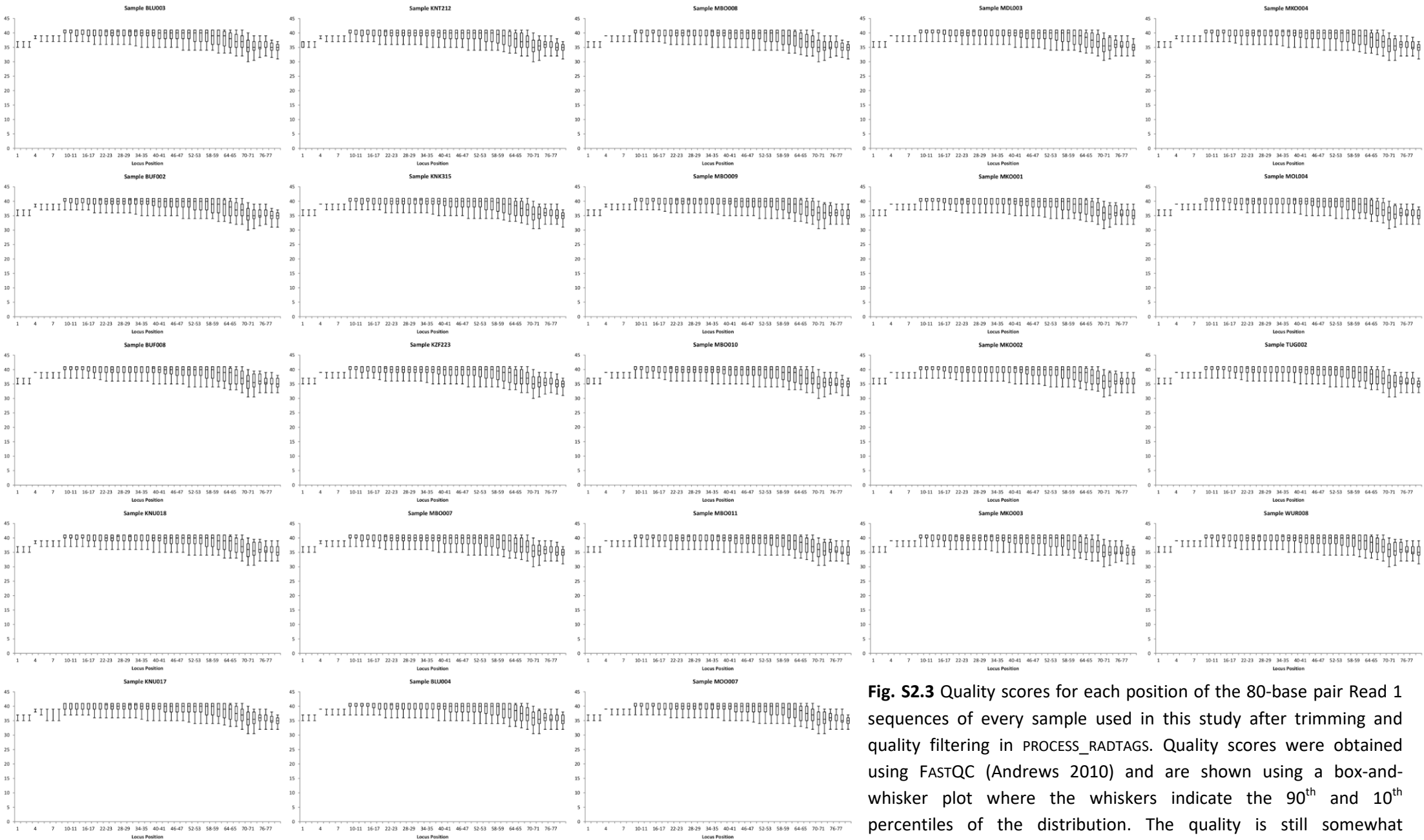


Fig. S2.3 Quality scores for each position of the 80-base pair Read 1 sequences of every sample used in this study after trimming and quality filtering in PROCESS_RADTAGS. Quality scores were obtained using FASTQC (Andrews 2010) and are shown using a box-and-whisker plot where the whiskers indicate the 90th and 10th percentiles of the distribution. The quality is still somewhat diminished towards the end of the reads, but is improved compared to the results prior to trimming.

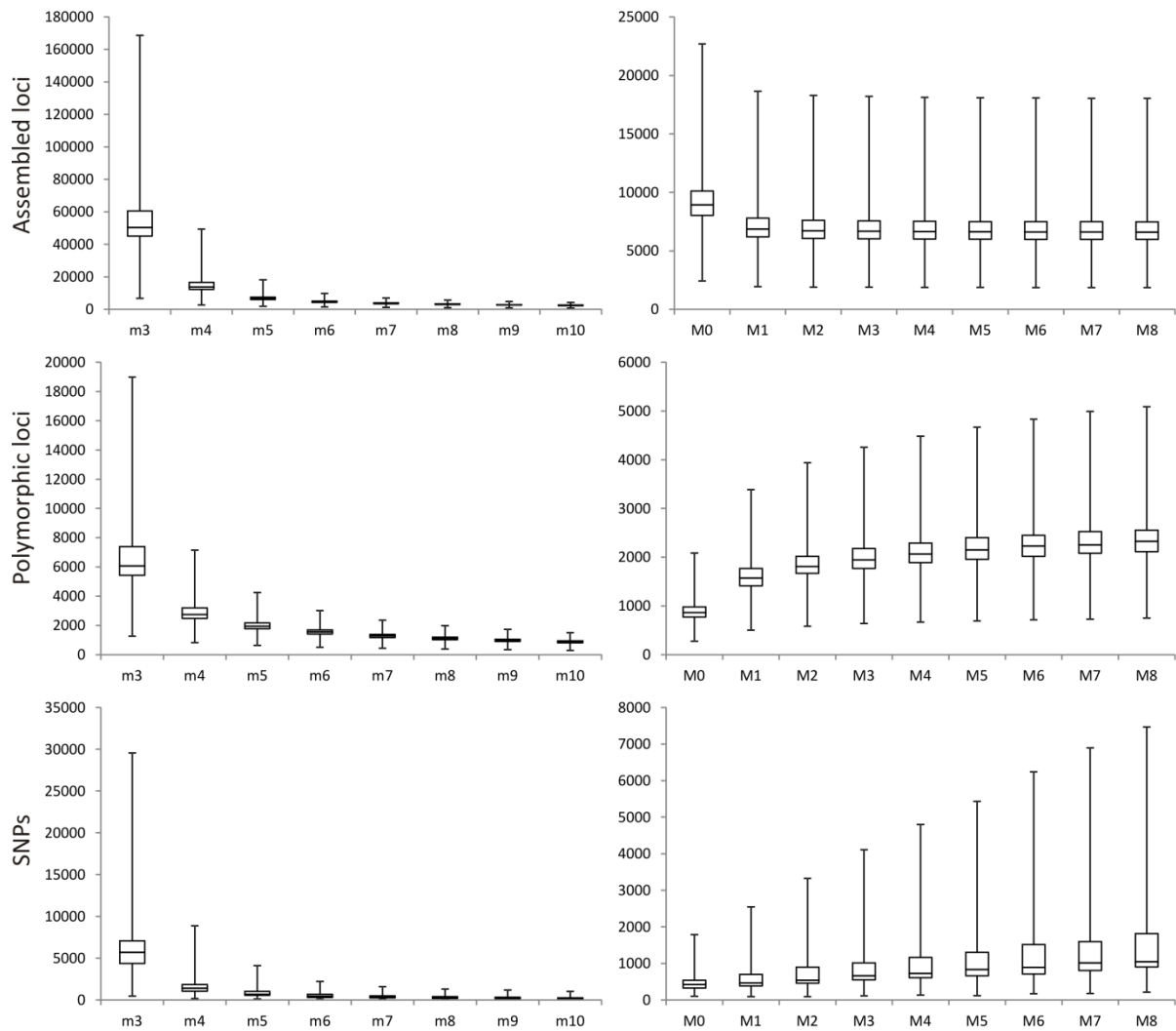


Fig. S2.4 Parameter optimisation of the STACKS parameters $-m$ and $-M$ measured by assessing the total number of assembled loci, polymorphic loci and SNPs identified as the default parameter set ($m5M3n2$) is changed by the variable on the X-axis as per Paris *et al.* (2017). The distribution of results across the 23 samples (plus two replicates) is given by a box-and-whisker distribution where the whiskers indicate maximum and minimum values. The trend for $-M$ approaches a plateau at $-M = 1$.

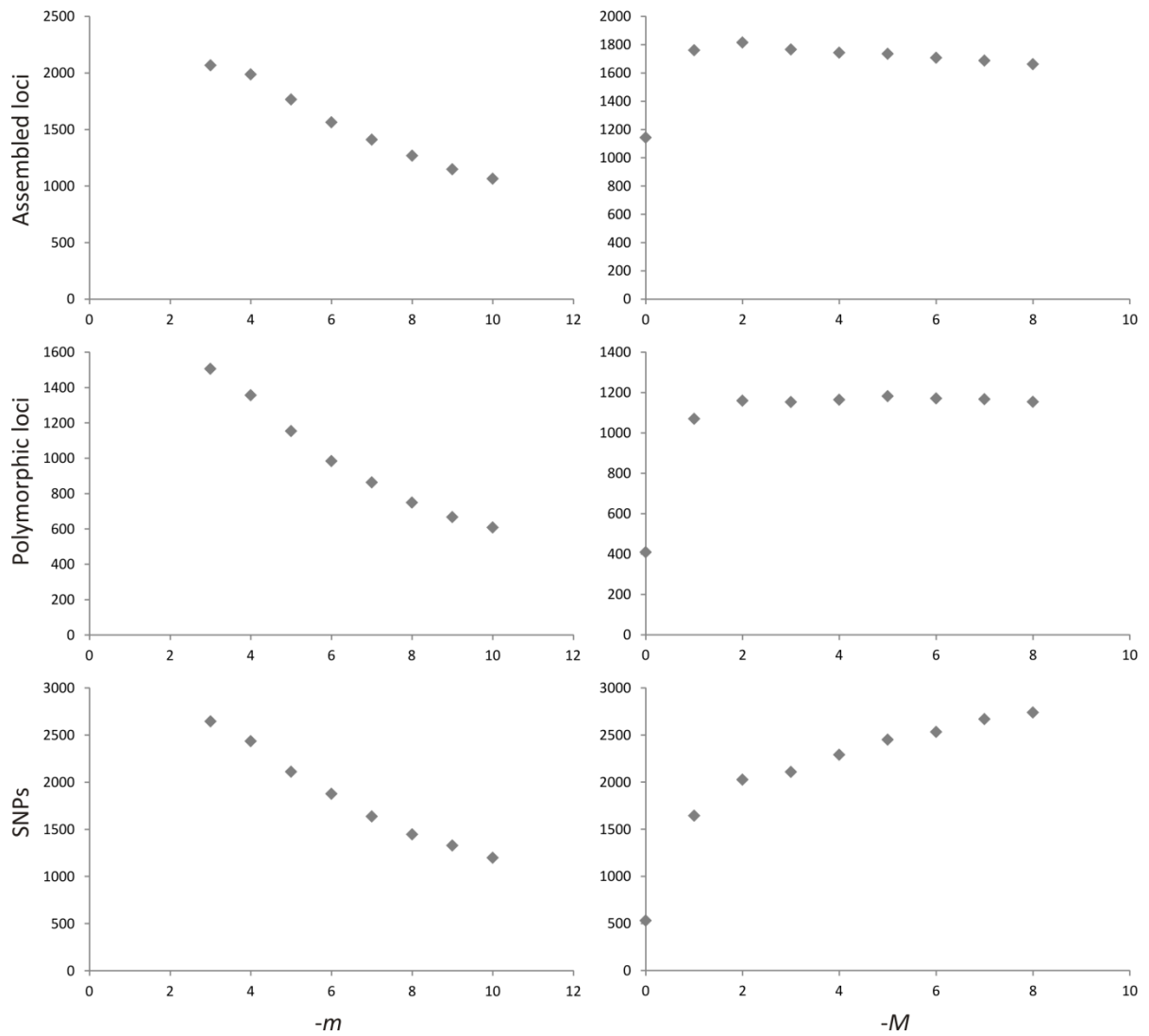


Fig. S2.5 Parameter optimisation of the STACKS parameters $-m$ and $-M$ measured by assessing the total number of assembled loci, polymorphic loci and SNPs identified as the default parameter set ($m5M3n2$) is changed by the variable on the x-axis as per Paris *et al.* (2017). The total values for the entire population (requiring a SNP to be present in at least 80% of samples as per Paris *et al.* (2017)) are indicated. The trend for $-M$ again approaches a plateau at $-M = 1$.



Fig. S2.6 All evolutionary pathways tested in DIYABC leading to the current five observed populations. Populations are abbreviated: Mbo = Mbokodweni; Mko = Mkomaa; Umf = Umfolozi; Umg = Umgeni; Tug = Tugela. The scenarios were tested in batches A, B, C and D. Scenarios with the most support using the logistic regression approach for each batch are indicated in bold. A: Basic scenarios of a polytomy were tested. B: Simple evolutionary ladder-like splits resulting in the five populations were tested, keeping the first diverged population as one of the two outgroups identified in A. C: A split into two lineages which later diverged into the current five populations was tested. Note that node ages were allowed to vary, resulting in only the ten scenarios tested. D: All of the most supported scenarios from the previous batches were tested against each other. The fifth scenario was most supported, although scenarios 3 and 7 received very similar levels of support.

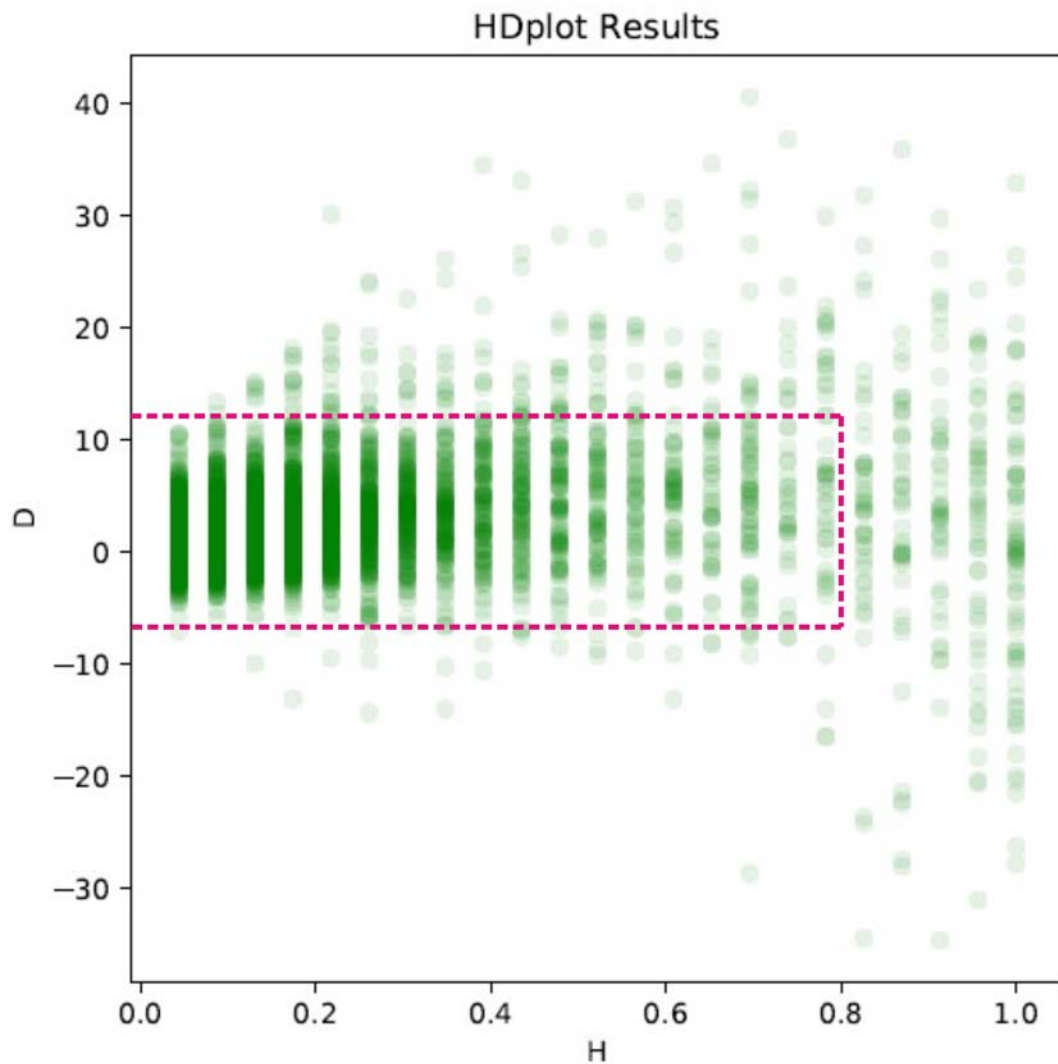


Fig. S2.7 Results of running the HDPLOT script (McKinney *et al.* 2017) on the current dataset with no threshold for the number of samples a locus must appear in to be called ($-r = 0$). Due to the sparsity of samples, only a certain number of values are valid for heterozygosity (H). The y-axis indicates deviation of both alleles of a diallelic locus from 1:1. The primary cloud (indicated by the rectangle) indicates non-paralogous loci, with duplicated loci separating above this cloud and diverged duplicates clustering to the right of the plot. Setting thresholds of $H < 0.8$ and $-7 < D < 12$ yielded 463 of 16,893 SNPs which were identified as potentially paralogous loci.

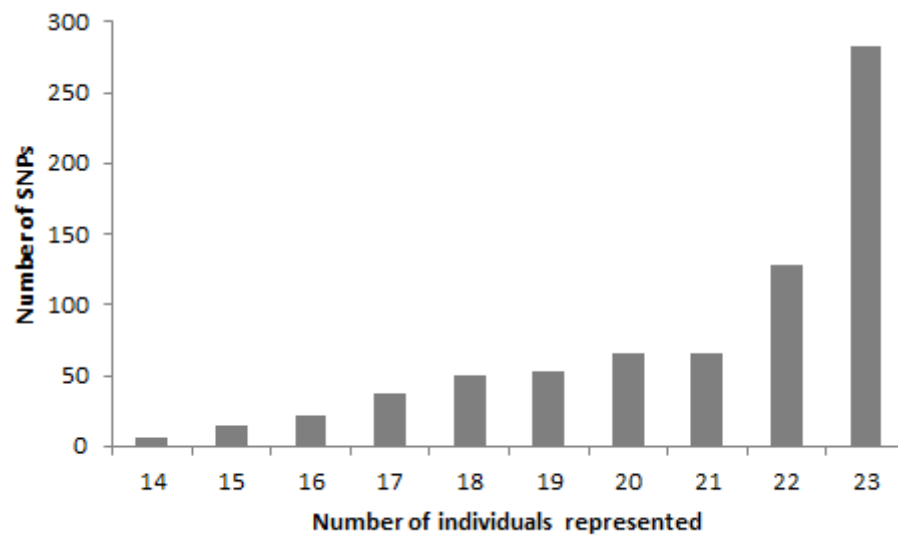


Fig. S2.8 Distribution of the final set of SNPs across all individuals used in this study due to the requirement for SNPs to be present in at least 60% (=14) of all samples.

Chapter 3: Finding stories in noise: Mitochondrial portraits from RAD data

Abstract

Restriction-site Associated DNA (RAD) Sequencing has become a highly popular Next-Generation Sequencing (NGS) technique utilised in a wide range of studies. However, research groups wanting to use this technique may experience a deluge of information without clear advice on how to proceed once they have obtained their raw data. In a previous paper I highlighted the importance of parameter estimation and optimisation. One proposed method for determining the reliability of RAD data has been to independently sequence one or more mitochondrial genes and compare the results of the nuclear genome RAD dataset to the mitochondrial gene dataset. However, a number of studies using this approach appear to be unaware that a large proportion of RAD SNPs that make it through the processing pipeline are themselves mitochondrial, and may bias analyses. Here I present a novel approach for verifying nuclear genomic signal through comparisons with mitochondrial markers present in the RAD dataset. Additionally, I mine this dataset to map up to 98% of the *Labeobarbus* mitogenome. I validate these mitogenome reconstructions by comparisons with cytochrome *b* gene sequences obtained through traditional Sanger sequencing. Finally, I investigate mitochondrial relationships between three species of South African *Labeobarbus* and identify potential instances of contemporary and more ancient genetic exchange between populations within *L. natalensis* as well as interspecific hybridisation between *L. aeneus* and *L. kimberleyensis*.

3.1 Introduction

Labeobarbus is a genus of hexaploid (± 150 chromosomes) (Oellermann & Skelton 1990) freshwater fish distributed throughout Africa. Its position in the family Cyprinidae has historically been a matter of much debate, although a recent phylogeny has placed the genus within the new tribe Torini (Yang *et al.* 2015). The genus is thought to have arisen from hybridisation of a tetraploid ancestral Torini with a diploid *Cyprinion* (Yang *et al.* 2015), which then underwent spontaneous chromosome duplication or autopolyploidy (Oellermann & Skelton 1990). The genome of *Labeobarbus* is still largely polyploid, which must be taken into account during genetic analyses. Some authors (*cf.* Levin 1983; Zhan *et al.* 2014; Van de Peer *et al.* 2017) 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, following placement of *Varicorhinus* in the synonymy of *Labeobarbus* (Tsigenopoulos *et al.* 2010; Yang *et al.* 2015; Vreven *et al.* 2016). 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 Orange-Vaal smallmouth yellowfish *L. aeneus* and the KwaZulu-Natal yellowfish *L. natalensis*. *Labeobarbus natalensis* is endemic to KwaZulu-Natal Province, whereas *L. aeneus* is distributed widely across the country, primarily within the Orange-Vaal river system. Mitochondrial phylogeographic studies have shown that *L. natalensis* consists of a number of lineages associated with separate drainage systems (Bloomer *et al.* 2007). Initial whole genome data are congruent with this, and separated the species into five major lineages (Chapter 2): the Umfolozi, Tugela, Mbokodweni, Mkomaas and Umgeni populations. *Labeobarbus aeneus* is thought to hybridise with Orange-Vaal largemouth yellowfish

L. kimberleyensis (Gaigher 1976; Eccles 1986; van Vuuren *et al.* 1989; Mulder *et al.* 1990). However, morphometrics, allozyme and mitochondrial DNA studies of introgression between these forms have been inconclusive (Bloomer *et al.* 2007). My ongoing research, using genomic data, is aimed at resolving this.

In order to investigate the phylogeography of these hexaploid species effectively I have opted to use a Next-Generation Sequencing (NGS) approach to analyse genomewide single nucleotide polymorphisms (SNPs) which can add to the previous mitochondrial results. Restriction-site Associated DNA (RAD) sequencing (Miller *et al.* 2007; Baird *et al.* 2008) has become a popular NGS 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 (Davey *et al.* 2011; Davey *et al.* 2013; Andrews *et al.* 2016).

The power, utility and popularity of this technique have resulted in a number of modified versions of the protocol, such as double-digest RAD (ddRAD) (Peterson *et al.* 2012). This approach utilises whole genome digestion with both a rare-cutting and a common-cutting restriction enzyme, removing the need for fragmentation by sonication (Peterson *et al.* 2012). Fragments within a given size range possessing both enzyme cut sites are then sequenced (Peterson *et al.* 2012). This method was designed to have several advantages over traditional RAD sequencing, namely: 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 new pitfalls, such as compounding allele frequency estimation bias from restriction site mutations (Arnold *et al.* 2013; Eaton *et al.* 2017), indels eliminating alleles from size-selection (DaCosta & Sorenson 2014), and overrepresentation of particular genomic regions due to enzyme choice (DaCosta & Sorenson 2014). I also found that incomplete restriction digestion and size selection may both have marked

effects on the results obtained using ddRAD sequencing, complicating downstream analyses (Chapter 2).

A popular verification method for RAD data has been to sequence other markers for independent analysis, such as one or more mitochondrial DNA fragments (Streicher *et al.* 2014; Herrera *et al.* 2015; Jezkova *et al.* 2015; Macher *et al.* 2015; Pante *et al.* 2015). This allows one to show whether observed trends are corroborated by other markers, although in some cases a different pattern will result from this mitochondrial sequencing (e.g. Cruaud *et al.* 2014; Jezkova *et al.* 2015; Macher *et al.* 2015; Puckett *et al.* 2015). Nuclear-mitochondrial discordance here may be indicative of a problem during the RAD data analysis or an evolutionary process responsible for the discrepancy such as introgression (reviewed in Ballard & Whitlock 2004). To my knowledge, few studies have attempted to mine the mitochondrial data present in RAD datasets for independent analysis (but see Truong *et al.* 2012; Pujolar *et al.* 2014). Here I present RAD data analyses following my previous study (Chapter 2) where I validate my results using mined mitochondrial markers. I also contrast this to previously conducted mitochondrial phylogeographic analysis from one of the species (Bloomer *et al.* 2007). I construct partial mitogenomes for each of the samples to investigate the evolutionary relationships between the three species. And finally, I validate the mitochondrial mappings by comparing them to Sanger sequences of a mitochondrial gene.

3.2 Materials and Methods

3.2.1 Sample collection and DNA extraction

I used the 23 samples previously described in Chapter 2 for the KwaZulu-Natal yellowfish (*L. natalensis*), and an additional 39 samples of Orange-Vaal yellowfishes (24 *L. aeneus* and 15 *L. kimberleyensis*). The Orange-Vaal samples were collected between January 2004 and October 2006 from ten localities in South Africa by Roger Bills (Fig. 3.1; Supporting Information Table 3.1). 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.

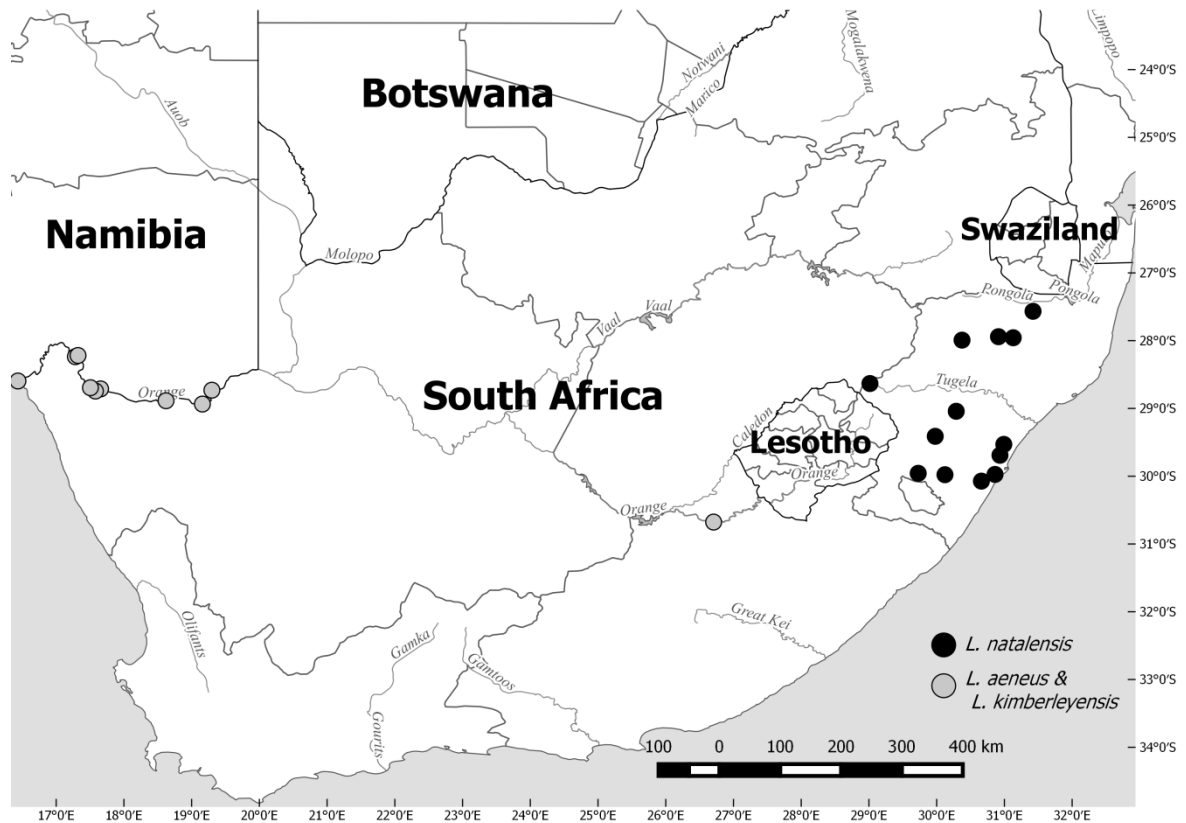


Fig. 3.1 Sampling locations across South Africa showing major river systems. Sampling localities in black correspond to *L. natalensis*, localities in grey are *L. aeneus* and *L. kimberleyensis*. This map was produced using QGIS (QGIS Development Team, 2016. QGIS Geographic Information System. Open Source Geospatial Foundation Project. <http://www.qgis.org/>).

A GeneQuant™ *pro* RNA/DNA calculator spectrophotometer (Amersham Biosciences, Freiburg, 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 libraries, with a separate library for the KwaZulu-Natal yellowfish (Supporting Information Table S3.1).

3.2.2 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 commonly throughout the entire 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. An additional size-selection step of 300-500 bp was implemented. Library sequencing (90 bp paired-end) occurred in a single lane of an Illumina HiSeq 2000 (Illumina Inc., USA) for the first Orange-Vaal library, 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 me for further analyses. The reads were then trimmed to a consistent length of 80 bp after removing adapter and barcode sequences.

3.2.3 Bioinformatics and mitochondrial SNP discovery

Raw RAD data files were put through the pipeline described in Chapter 2, using STACKS 1.44 (Catchen *et al.* 2011; Catchen *et al.* 2013) for stack-building and SNP identification. CLC GENOMICS WORKBENCH 7.0.4 (CLC Inc., Aarhus, Denmark), FASTQC (Andrews 2010) and custom scripts were used to assess quality, GC-content and levels of error in the libraries. Paired reads were physically joined for SNP identification using custom scripts (AS Papadopoulos personal communication), resulting in a single file per individual containing reads of 160 bp. Following the STACKS parameter-testing method of Paris *et al.* (2017) I identified an optimal parameter set of (*-m* 5 *-M* 2 *-n* 2) for

SNP identification in paired reads from the *L. natalensis* library (Supporting Information Figs. S3.1-S3.3, Table S3.3). For conformity, I adopted the same parameters for Orange-Vaal yellowfishes.

Once stacks had been built and SNPs identified, reads that had passed through the entire pipeline were then run through CLC MAIN WORKBENCH 6.9 (CLC Inc., Aarhus, Denmark) local BLASTN (Altschul *et al.* 1997) against the following 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 have 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 $1E-20$ 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. All SNPs with more than a single allele per individual were removed from the analysis as potential nuclear mitochondrial DNA segments (NUMTs) or containing excessive sequencing error.

A whitelist was constructed for all of these mitochondrial SNPs, which had been recovered by STACKS and identified using BLAST. The POPULATIONS module of STACKS was run again with this whitelist and 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 NUMT) and 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. The SNPs were converted to NEXUS format using PGD SPIDER 2.0.7.2 (Lischer & Excoffier 2012) and imported into TCS 1.21 (Clement *et al.* 2000) to produce haplotype networks.

3.2.4 Mitogenome assembly

Individual read files, which had been filtered for adaptor pollution, were mapped in CLC GENOMICS WORKBENCH 7.0.4 (CLC Inc., Aarhus, Denmark) against the mitogenome of the East African *L. intermedius*, as the most closely related available reference mitogenome. Overlapping paired reads were merged prior to mapping. I included seven replicated samples (the same sample sequenced twice in separate runs with different barcodes) as controls to determine whether sequencing the same individual twice yields the same results in this form of analysis. This should give an indication of how robust these methods are. I performed an *in silico* digestion of the *L. intermedius* mitogenome to produce an estimate of the maximum proportion of the *Labeobarbus* mitochondrial genome expected assuming incomplete restriction digestion is prevalent (Supporting Information Table S3.2). Mappings used default parameters except for fraction length and similarity which were both set to 0.9. Non-specific mappings were ignored. Consensus read mappings were generated by requiring a minimum depth of 3 reads to minimise possible errors 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 *et al.* 2013) for manual checking of the alignment. Extraneous gaps were removed and the alignment was exported in NEXUS format. The NEXUS file was used in TCS to

produce haplotype networks with 95% connection limits. More connecting steps were allowed where possible. Model checking was performed in MEGA to identify the optimal nucleotide substitution model for the dataset. The alignment was then exported into BEAUTI 1.7.4 (Drummond *et al.* 2012) to produce the XML file for use in BEAST 1.7.4 (Drummond *et al.* 2012). The substitution model was changed to HKY+I+G set as the site heterogeneity model (Hasegawa *et al.* 1985). I used a strict clock with a cyprinid rate for cytochrome *b* (cyt *b*) of 0.53% per lineage per million years (Dowling *et al.* 2002) which has previously been used for *Labeobarbus* (Tsigenopoulos *et al.* 2010) and other cyprinids (Jeffries *et al.* 2016). I specified a constant size coalescent tree prior with a chain of 50×10^6 MCMC steps, sampling every 5,000 steps. All other parameters were left at their default values. The tree produced from this analysis was annotated using TREEANNOTATOR 1.7.4 (<http://beast.bio.ed.ac.uk/TreeAnnotator>) and viewed in FIGTREE 1.3.1 (<http://tree.bio.ed.ac.uk/software/figtree/>). I produced a phylogram in MRBAYES (Ronquist & Huelsenbeck 2003) from the same input file, with the same substitution model as in the BEAST analysis, and 30×10^6 MCMC steps. All other parameters were retained at default values. Output was viewed in FIGTREE and the outgroup was trimmed.

3.2.5 Comparison 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 *et al.* 1996) for all samples of *L. natalensis* used in this study. I 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 1x 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. The 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 NaAc. Precipitated products were sequenced in both directions using the same primers as in the PCR amplification.

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 mitogenome mapping and Sanger sequence for each sample. Replicates were included in this analysis. The total number of pairwise differences was used to calculate an error rate by presenting it as a proportion of the total number of base pairs recovered in both approaches, i.e. removing bases that were only recovered from a single approach.

3.3 Results

3.3.1 Library preparation and sequencing

Illumina 90PE sequencing of the first Orange-Vaal library resulted in less than 83.5 million reads (Table 3.1) over all 11 individuals. The average number of raw reads per individual was 7,587,820. In contrast, the second Orange-Vaal library consisted of almost 336 million reads across 28 individuals (four of which were replicated). The average number of reads per sample in this library was 10,494,963 reads. Filtering of each library for adaptor pollution and quality resulted in a significant drop in read count (Table 3.1). Both the Q20 percentage and GC-content of the Orange-Vaal libraries were comparable to that observed for *L. natalensis* (Chapter 2), although the quality of bases was generally superior in the former.

Table 3.1 Summary statistics from initial analysis of RAD sequencing data

Library	Raw reads	Filtered reads	Q20 (%)	GC (%)
<i>L. natalensis</i>	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		

3.3.2 Bioinformatics and mitochondrial SNP discovery

I found a total of 102 polymorphic putative mitochondrial SNPs in the *L. natalensis* dataset and 98 polymorphic putative mitochondrial SNPs in the Orange-Vaal yellowfish libraries. All of these loci were verified using BLASTN. A number of these SNPs (12 and 13 respectively) had to be removed for occurring within a locus that had a heterozygous SNP, which is potentially indicative of sequencing error, erroneous merging of loci, contamination or the presence of NUMTs. Duplicated SNPs originating from overlapping sequences between paired reads were also

identified and removed (14 and 19 respectively). After filtering I retained 77 polymorphic mitochondrial SNPs for *L. natalensis* and 66 polymorphic mitochondrial SNPs for the Orange-Vaal yellowfishes.

Haplotype networks obtained using only mitochondrial SNPs identified from the STACKS pipeline (Figs. 3.2-3.3) showed a general division into species and catchment associated populations. I recovered the previously identified populations of *L. natalensis* (Chapter 2), as well as a novel split within the southernmost population with separate haplogroups in the Mkomaas and Mzimkhulu River catchments. The Mkomaas haplogroup was the most divergent from other lineages, while the Mzimkhulu lineage had the highest divergence among haplotypes within a haplogroup. Orange-Vaal yellowfishes broadly split into two major haplogroups – one corresponding to each species (Fig. 3.3), but with a few misplaced individuals, where assignment to an mtDNA lineage conflicted with phenotypic identification of the fish.

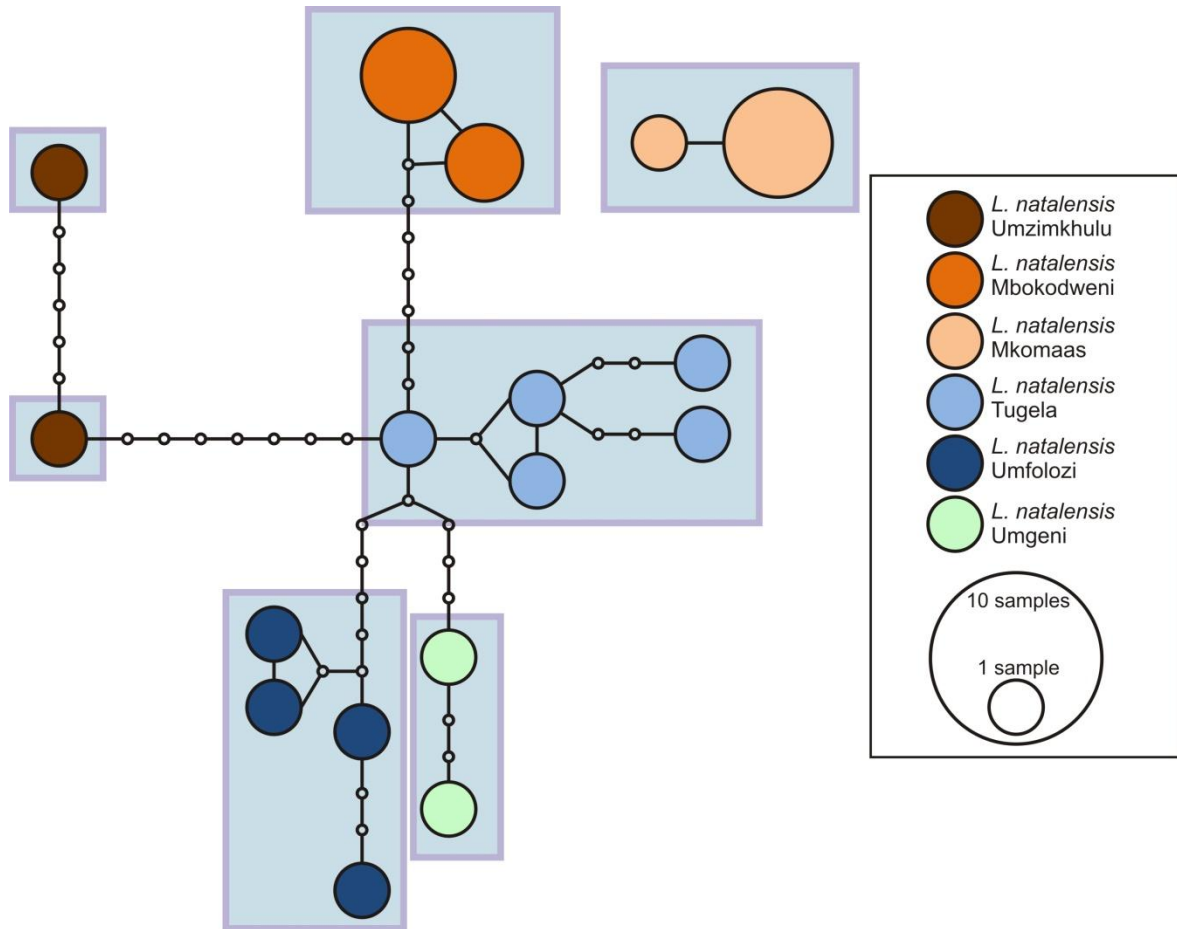


Fig. 3.2 Haplotype network produced from 77 polymorphic mitochondrial SNPs identified using the STACKS pipeline in *L. natalensis*. Boxes indicate 95% connection limits (three steps). A maximum of nine connecting steps were allowed.

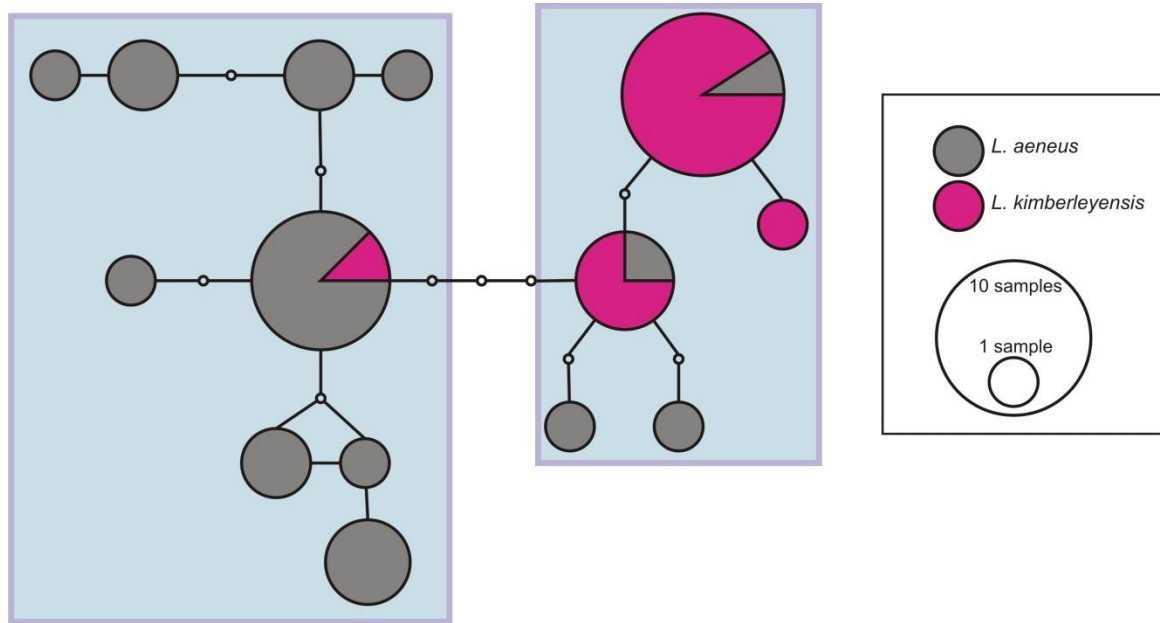


Fig. 3.3 Haplotype network from 66 polymorphic mitochondrial SNPs identified using the STACKS pipeline in Orange-Vaal yellowfish. Boxes indicate 95% connection limits (three steps). A maximum of four connecting steps were allowed.

3.3.3 Mitogenome assembly

In total, 467,848 reads were mapped to the *L. intermedius* mitogenome across all individuals (~0.1% of all reads). This proportion of mitochondrial to nuclear reads is comparable to studies in birds (Oswald *et al.* 2017) and far exceeds the proportion of mitochondrial DNA within the total genome. *In silico* restriction digestion of the *L. intermedius* mitogenome gave an expected 84% coverage if 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%), with chimeric coverage, combining all mapped reads across individuals, amounting to 98% of the mitogenome. Imposing a minimum depth of three reads at a site reduced the average coverage to 54% (range = 15-76%).

Haplotype networks produced from the consensus mitogenomes for each library (Figs. 3.4-3.5) were found to be consistent with previous studies (Chapter 2; Bloomer *et al.* 2007) and with the STACKS-based approach of haplotype assignment from mitochondrial SNPs. However, the haplotype mapping approach yielded finer resolution of divergences and greater separation of populations, due to the greater quantity of data generated by mapping directly to a reference genome. This also suggests that the mapping approach retained many SNPs which were dropped or excluded by the STACKS SNP discovery pipeline. For *L. natalensis*, haplogroups were generally separated by a large number of mutations with relatively few mutations separating haplotypes within populations. The situation for the Orange-Vaal yellowfishes is more complex, with two to three major clusters observed, broadly split into the two different species. However, some haplotypes from *L. aeneus* were found to cluster within the predominantly *L. kimberleyensis* haplogroup. Three of these samples matched those from the STACKS-based approach. Replicate samples were expected to sometimes not cluster together as although the biological samples are identical, the sequences used to generate these results are not and differ by coverage and depth of each replicate.

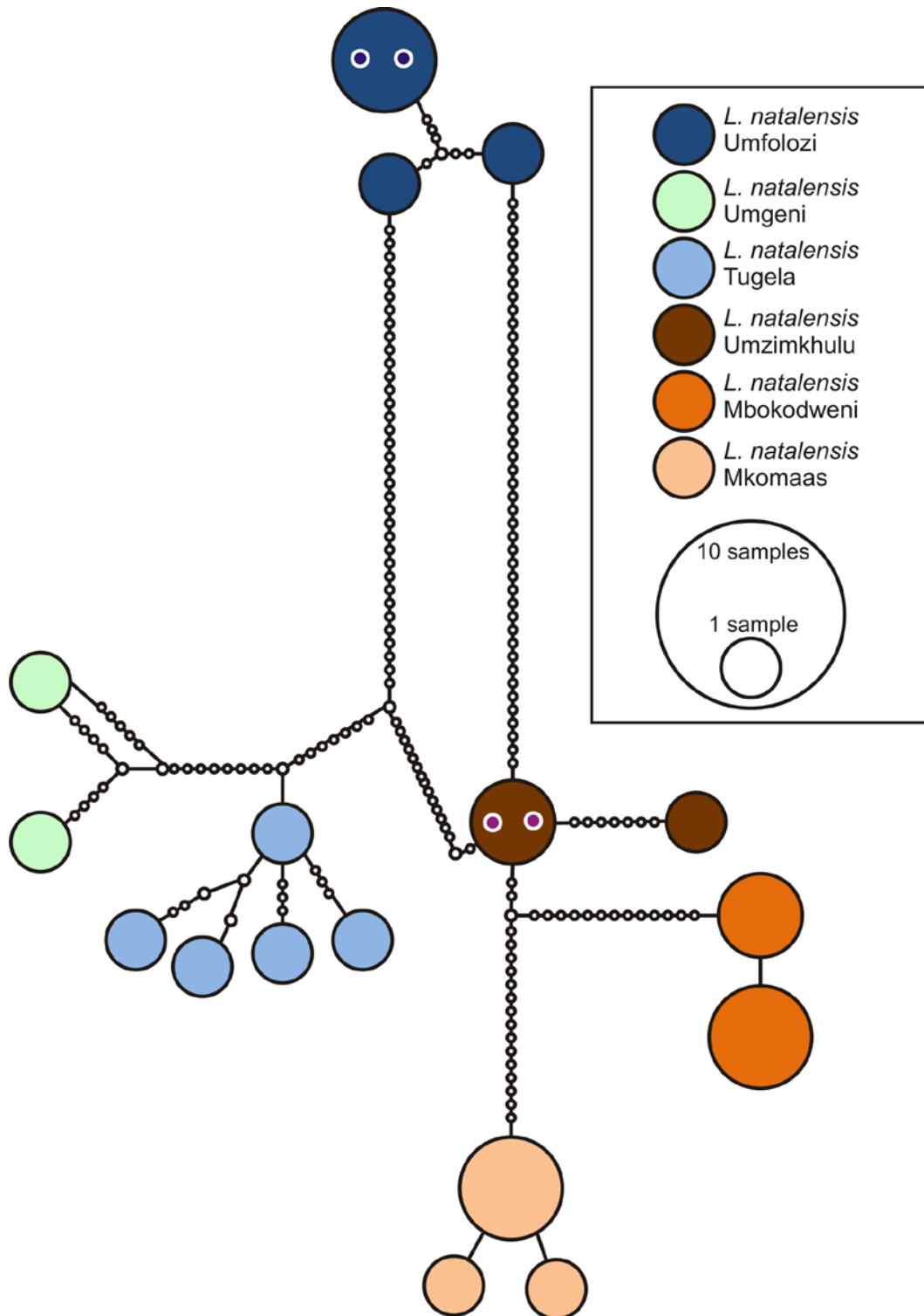


Fig. 3.4 Haplotype network produced from consensus mitochondrial genome sequences found by mapping read data of 25 *L. natalensis* samples against the *L. intermedius* reference mitogenome. Two replicate pairs are included and indicated by white-bordered circles with shades of purple indicating pairs.

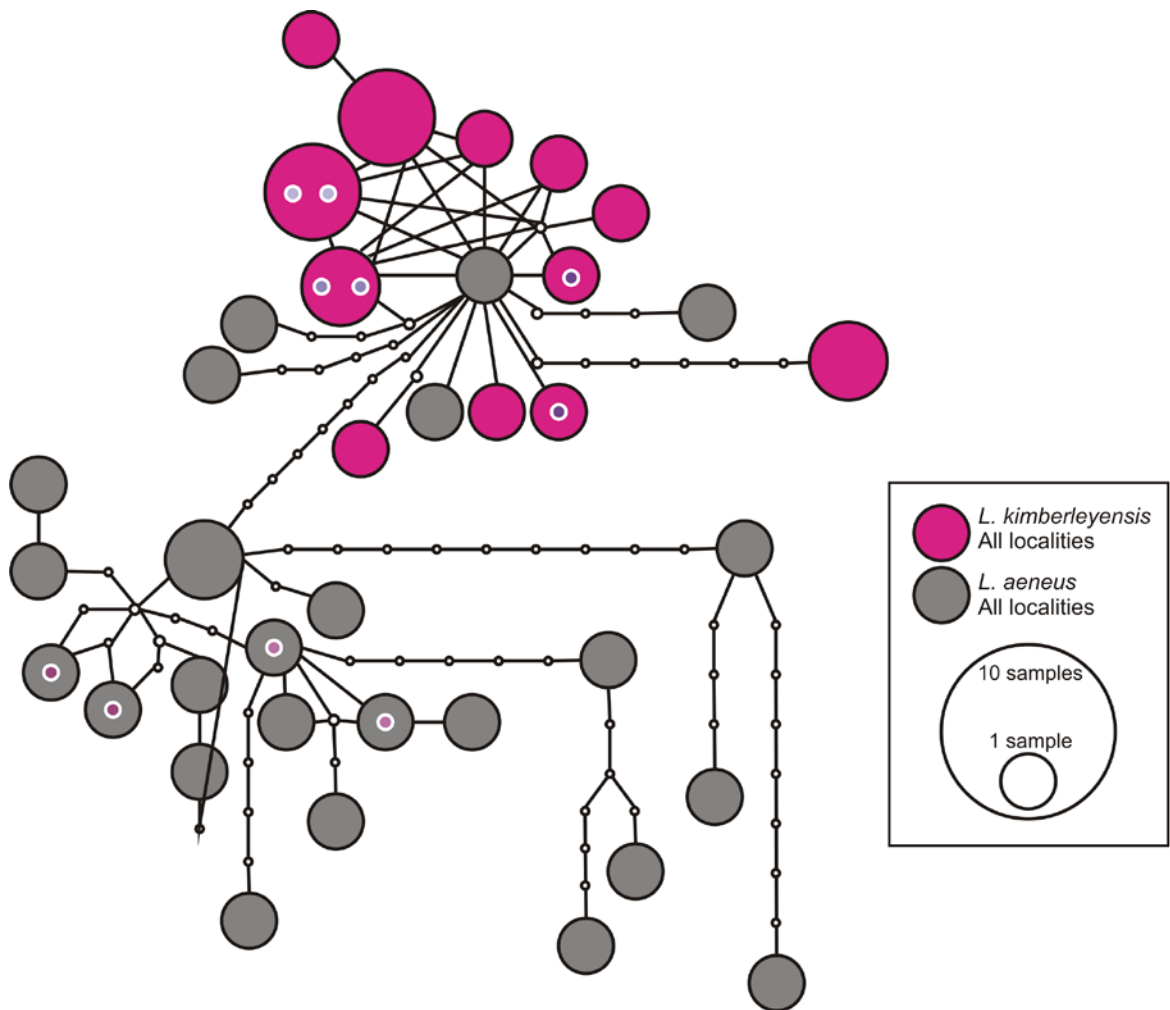


Fig. 3.5 Haplotype network for the 44 Orange-Vaal yellowfish produced using consensus mitochondrial genome sequences found by mapping read data against *L. intermedius*. Five replicate pairs are included and indicated by white-bordered circles with shades of purple indicating the pairs.

The Bayesian Constant Coalescent chronogram (Fig. 3.6), produced from the consensus mitogenome sequences generated with the minimum depth requirement of three reads, clearly splits the South African yellowfish into the *L. natalensis* and Orange-Vaal lineages first, with further subdivision into the six populations previously identified for *L. natalensis* and into the two Orange-Vaal species, *L. aeneus* and *L. kimberleyensis*. I observed no further clustering by locality

for the Orange-Vaal yellowfishes, which was unexpected given the geographic distance between the upper Orange sampling site near Aliwal North and the remaining sampling sites in the lower Orange river, over 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 me to tentatively date divergence between the species used in this analysis. I estimated the divergence between the ancestors of *L. intermedius* and the South African lineage to have occurred 7.82 mya, with *L. natalensis* splitting from the other South African yellowfish 1.74 mya and subsequent division between the Orange-Vaal species occurring around 1.45 mya. The major populations within *L. natalensis* diverged between 0.35-1.06 mya. Coalescence within the Orange-Vaal yellowfish lineages occurred more recently – from 0.43 mya within *L. kimberleyensis* and from 0.25 mya within *L. aeneus*.

The Bayesian phylogram produced in MRBAYES from these data (Fig. 3.7) supported previous findings with each species forming a monophyletic clade, and clear substructure within *L. natalensis*. Most internal nodes on the tree were strongly supported, with the exception of the basal node in this unrooted tree. All replicates either clustered together or clustered within groups with low posterior probability support for separation, indicating that this method is robust.

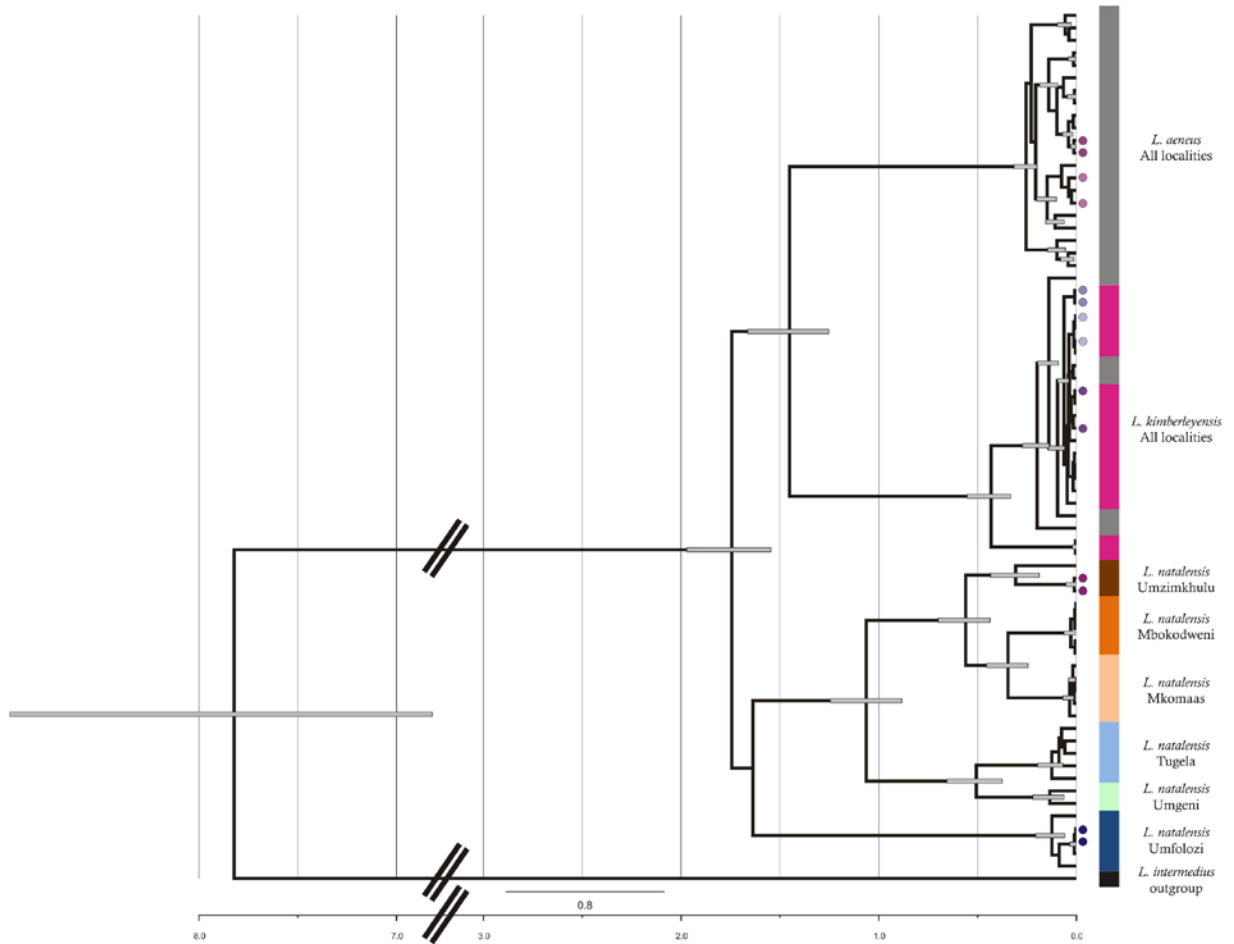


Fig. 3.6 Bayesian constant coalescent chronogram based on partial mitogenome reconstructions for 69 *Labeobarbus* samples (including seven replicate samples indicated by paired purple circles). The reference mitogenome of the East African *L. intermedius* was used as outgroup. Populations and species identified in this analysis are indicated by the coloured bar on the side. Colours of the *L. natalensis* populations match those described in Chapter 2. 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.

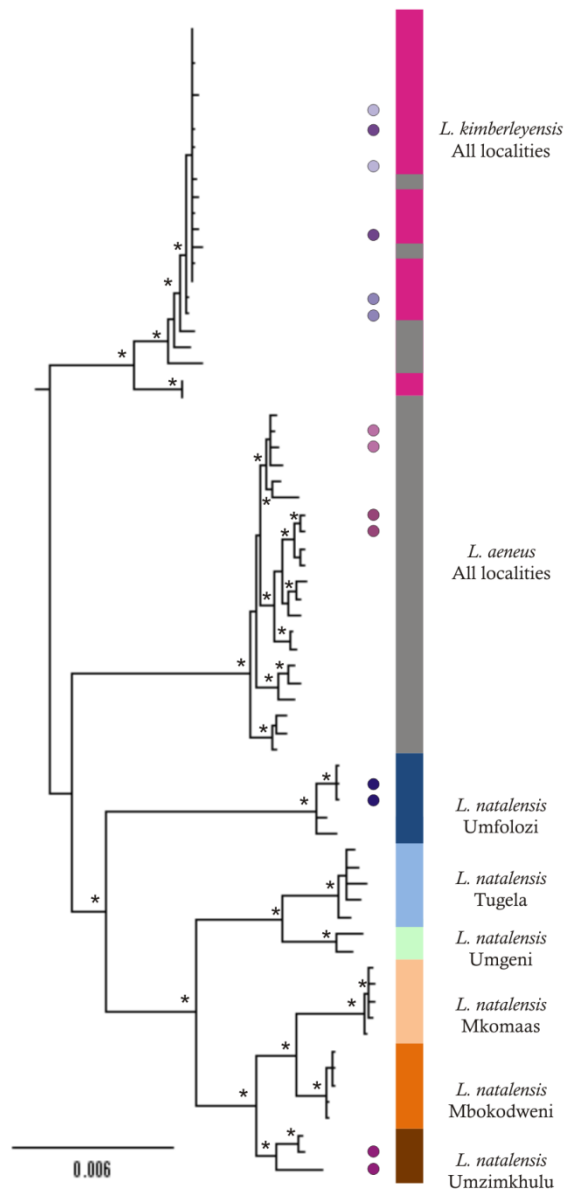


Fig. 3.7 Phylogram produced in MRBAYES from partial mitogenome reconstructions. Different yellowfish populations identified in this study are indicated by the coloured bar. Replicate pairs are represented by paired purple circles. Nodal probability support values equal to or exceeding 0.95 are indicated by a *. The outgroup was trimmed from the tree following rooting.

3.3.4 Comparison with Sanger sequencing

Sequences were obtained for all specimens of *L. natalensis* (23 samples) for 928 bp of cyt *b*. This allowed me to compare 4,919 base pairs that were obtained in both the Sanger sequencing and mitogenome mapping approaches in 27 pairwise comparisons due to inclusion of replicated samples. Of these 4,919 bases, I found only two errors, which occurred in the same individual from the same read mapping, giving an estimated error rate of 0.04%.

3.4 Discussion

3.4.1 Sequencing and mapping

The RAD sequencing of two additional species of *Labeobarbus* in this study greatly improves the resolution for the genus, particularly as the second Orange-Vaal yellowfish library was at a higher depth and produced a greater quantity of data than the other two libraries combined. This may influence the comparison of the nuclear SNPs of these libraries, but due to the abundance of the comparatively small mitochondrial genome, it appeared to have minimal impact here. This is shown by the highly comparable coverage across samples in different species and libraries, although the second Orange-Vaal library does have higher average coverage. Sites that are not represented due to uneven sequencing of samples are unlikely to drastically influence biological signal due to the stochasticity from which uneven sequencing arose (Eaton *et al.* 2017). In addition, this library also has the highest GC-content and superior base quality scores than the other two. The high GC-content is likely due to the repeats which are overrepresented in this particular library, and may not represent any real difference between species.

Previously I reported a higher than expected coverage of the nuclear genome from mapping RAD sequencing reads to the *Cyprinus carpio* nuclear genome (Chapter 2). In this study, I find a degree of coverage of the mitochondrial genome only slightly greater than would be expected from *in silico* digestion using the *L. intermedius* mitogenome. This near-complete representation of the *Labeobarbus* mitogenome is due to the *in silico* digestion being based on a single sample, whereas my results use all of the samples. This increases the coverage to include regions found in only single samples due to point mutations between samples resulting in the recognition site of the restriction enzymes being introduced in new genomic regions and lost in others. This allows me to reconstruct a greater proportion of the mitogenome than would otherwise be expected. The average mitogenome mapping coverage is lower than the estimate produced by *in silico* digestion as this estimate assumes incomplete enzyme digestion such that some enzyme recognition sites will be recovered undigested within some reads.

3.4.2 STACKS-based mitochondrial SNP identification

I identified a relatively large proportion of mitochondrial SNPs through the STACKS pipeline using my approach. The size of the mitochondrial genome led me to expect fewer mitochondrial SNPs to be present, as did the requirement of mitochondrial SNPs to be called either due to alternate homozygotes with an absence of heterozygotes as a result of haploidy in a diploid model, or as a result of sequencing error. However, the comparative abundance of mitogenome reads appears to have resulted in a high degree of coverage across samples which otherwise may not share nuclear SNPs due to sequencing depth limitations. These markers are therefore prime candidates to successfully pass through a traditional STACKS pipeline, although I have previously seen that many of these markers are filtered as F_{ST} outliers (Chapter 2; Gleason & Burton 2016).

The mitochondrial SNPs passing through the STACKS pipeline provided variation exceeding that expected by resequencing a single mitochondrial gene, demonstrating the usefulness of this approach. The haplotype networks produced from these mitochondrial SNPs split the populations for each species as expected, although the Orange-Vaal haplogroups showed a few individuals clustering in a different group to that which they had been morphologically assigned. This may be a result of incorrect morphological assignment, mitochondrial introgression, or ancestral polymorphism. Because these species have been hypothesised to hybridise (Gaigher 1976; Eccles 1986; van Vuuren *et al.* 1989; Mulder *et al.* 1990), I believe that mitochondrial introgression is the most likely explanation.

3.4.3 Mitogenome mapping SNP identification

The mitogenome mapping haplotype networks, Bayesian constant coalescent chronogram and phylogram revealed five samples of Orange-Vaal yellowfish clustering into different haplogroups than would be expected by morphological assignment. This is concordant with the STACKS-based haplotype network and reaffirms that this is unlikely to be due to an error during analysis. However, only three of these “misassigned” samples coincided between approaches. I suggest that differences in samples identified as belonging to alternate haplogroups is likely due to differences in coverage and depth between samples, as each of the samples which were only identified in a single approach were from the low-coverage library and possessed relatively few reads. This, combined with the relative paucity of markers used in the STACKS-based approach, would indicate that the five misassigned samples identified from mitogenome mapping are likely the most relevant. All of these samples were morphologically classified as *L. aeneus* but in this analysis were found to cluster with *L. kimberleyensis*. This suggests that male *L. aeneus* are fertilising female *L. kimberleyensis* eggs despite

temporal separation of their breeding seasons (Mulder 1973; Tómasson *et al.* 1984). Asymmetrical hybridisation has been observed between Carpathian and common barbels (*Barbus carpathicus* and *B. barbus*), where it is thought to occur due to size differences in the two parental species or from unequal effects of selection on viability or fertility of female hybrids (Lajbner *et al.* 2009). In this case the asymmetry may merely be explained by the relative rarity of *L. kimberleyensis* across the distribution in combination with the increased time to reproductive age in this species (Mulder 1973). Additionally, it is thought that *L. aeneus* spawns earlier than *L. kimberleyensis*, at a time before gonad development is complete in the latter species (Mulder 1973; Tómasson *et al.* 1984). However, *L. aeneus* has an extended spawning season (Mulder 1973) and may remain active for a period that overlaps with *L. kimberleyensis* spawning. Nuclear SNPs should be analysed before conclusions can be drawn on potential hybridisation to remove the possibility of erroneous morphological assignment.

Very little other genetic structure was observed within the *L. aeneus* and *L. kimberleyensis* clades, whereas *L. natalensis* exhibited a high degree of structuring. This is perhaps not so surprising when one considers that the former species occur in a single large river system, whereas *L. natalensis* is found in isolated drainage systems, and all species are migratory. Additionally, the dating of nodes on the constant size coalescent tree indicate that diversification within both Orange-Vaal lineages has occurred recently (within the last 0.43 million years). However, I observe that all three of the samples of *L. aeneus* from the upper Orange River were among the five samples clustering in a different species group (*L. kimberleyensis*). This may indicate that previous study findings of genetic and morphological differences between *L. aeneus* of the upper and lower Orange River (Bloomer & Naran 2006; Bloomer *et al.* 2007) could be due to a greater degree of hybridisation occurring in the upper Orange River, with Augrabies Falls potentially acting as a barrier between the two regions. The

falls have been identified as a biogeographic barrier separating historically distinct fish faunas (Skelton 1986). Further studies using nuclear markers are necessary to evaluate this.

The haplotype networks produced from the mapping approach include more variability than the STACKS-based approach, as exemplified by the separation between the *L. natalensis* haplogroups/populations and between the major haplogroups of the Orange-Vaal yellowfish. I observed relatively few mutations separating samples within populations of *L. natalensis*. The overall trend from both mapping haplotype networks matches closely to the trends seen in the haplotype networks from the STACKS-based approach, but with a finer degree of resolution. For example, *L. natalensis* from the Mzimkhulu system carry a distinct mitochondrial lineage, as was previously suggested from a study analysing mitochondrial control region sequences (Bloomer *et al.* 2007), but these individuals were not distinguished from the Mkomaas population using nuclear RAD data (Chapter 2). This may be indicative of recent divergence events that are not yet evident from the nuclear genome, or irregular gene flow between populations which has resulted in the more rapid fixation of mitochondrial haplotypes while the nuclear genomes remain relatively similar.

Additionally, the Umgeni population was also identified as a unique lineage, which is contrary to initial nuclear data suggesting that these individuals are potential hybrids between the northern and southern populations (Chapter 2). This discrepancy may be indicative of relatively ancient admixture events between the northern and southern lineages within this putatively northern lineage which is closely related to, but distinct from the Tugela population. Additionally, the sample from the upper Umgeni system (Lions River) clustered with the Tugela system in both the mitochondrial and nuclear RAD analyses. This is suggestive of 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 fish sampled from the Lions River must be an early translocated individual from the Mooi River (Tugela population) which explains its affiliation to that population in both the

nuclear and mitochondrial analyses. Viewing the mitochondrial and nuclear analyses as a whole, this suggests that the Umgeni system has its own native population of yellowfish which may potentially be undergoing ongoing admixture with the Tugela population due to the MMTS, and which may have historically introgressed with the southern lineage (Mkomaas or Mbokodweni). Although the lower Umgeni samples may not show evidence of the contemporary admixture event, the MMTS has been operational for some time (since 2003, with some earlier infrequent transfers) (Hunter 2009) and because the sampling of this region occurred during 2003-2006, I do not know what the state of the Umgeni population currently is.

The species that was selected for mapping, *L. intermedius*, is fairly distantly related to South African yellowfish, as expected given geographic separation and the history of southward radiation of this genus (Tsigenopoulos *et al.* 2010). As a result of this distance I observe a long branch connecting it to the three South African *Labeobarbus* in the Bayesian constant coalescent chronogram. I tentatively dated the node of the common ancestors between *L. intermedius* and the South African lineage as occurring 7.82 mya, which fits with the proposed theory of invasion of Africa by early *Labeobarbus* in the Late Miocene (Tsigenopoulos *et al.* 2010). However the same node based on *cyt b* divergence alone was previously estimated to date to 4.23 mya (Tsigenopoulos *et al.* 2010). This may be due to the use of a pairwise rate in the previous paper as opposed to the per-lineage rate here, or by the difference between using partial mitogenome reconstructions as opposed to a single gene. Divergence within the South African lineage occurs from around 1.74 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).

3.4.4 Sanger sequencing validation

The validation of the mitogenome mapping approach in comparison with Sanger sequences was successful – the only two discrepancies yielded a putative error rate of 0.04%. This shows that some mistakes are incorporated in the mapping approach, but their occurrence is fairly rare. The success of this approach will also be heavily influenced by divergence between the source organism and the organism mapped against – I found a much greater degree of error when I used *N. hexagonalepis* as the reference mitogenome for mapping (data not shown). However, the use of strict mapping parameters coupled with visual inspection should remove most problematic areas from a mitogenome mapping, allowing one to extract useful mitogenomic data from an approach that has traditionally only been used for nuclear data.

3.5 Conclusion

In this chapter I have presented a novel approach of mining RAD data for mitochondrial SNPs which can be isolated and analysed independently, potentially validating nuclear RAD SNP results or identifying genomic processes responsible for discordance such as introgression. I identified these markers using two different methods: processing all reads through the SNP identification pipeline and then filtering mitochondrial SNPs using BLAST and preprocessing RAD data files to minimise errors followed by mapping these files against the mitogenome of a species from the same genus. The first method found 77-66 of 5,991 (*L. natalensis*) to 117,183 (*L. aeneus* and *L. kimberleyensis*) total SNPs which passed through the entire SNP identification pipeline to be mitochondrial (0.06-1.29%). This reflects a higher proportion of mitochondrial SNPs in the *L. natalensis* final dataset than is expected (0.1% of all reads mapped as mitochondrial), whereas the other two species had a lower proportion than expected. These SNPs were used to identify the mitochondrial genetic structure

independently of the nuclear RAD data. The results were similar – mitochondrial SNPs split *L. natalensis* into more groups (six total groups), than is evident from nuclear RAD data. The results are concordant in splitting *L. natalensis* into groups by drainage system, and may be indicative of more recent subdivision which is exhibited in the mitochondrial SNPs but has not yet become evident across the nuclear genome, or of gene flow between drainage systems.

Additionally, the mapping method resulted in mitogenome assemblies of high coverage – 66% on average per sample prior to the depth requirement which reduces this to an average coverage of 54%. I further combined these mappings over all three species to produce a chimeric South African *Labeobarbus* mitogenome representing ~98% of the total genome. This is a valuable resource for further work on South African *Labeobarbus*. I used the incomplete mitogenome mappings for each individual to generate maximum likelihood trees and a phylogram to test the relationship between the samples. I found a correlation between these mitogenome phylogenies and the existing mitochondrial patterns identified using isolated mitochondrial SNPs as well as by sequencing the mitochondrial cyt *b* region for *L. natalensis*. Finally, this approach combined with my previous nuclear genomic analysis allowed me to identify potential contemporary hybridisation events within populations of *L. natalensis* (Tugela to Umgeni) as well as more ancient intraspecific hybridisation (Umgeni with Mkomaas/Mbokodweni) within *L. natalensis* and interspecific hybridisation between the Orange-Vaal yellowfish *L. aeneus* and *L. kimberleyensis*. This showcases the power and utility of mitochondrial DNA comparisons and of this approach in particular. Inclusion of data such as this may greatly enhance interpretation, even for genomic studies using techniques such as RAD to identify hundreds to thousands of nuclear markers.

3.6 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 Rob Karssing. I would also like to thank Vincent Savolainen, Bengt Hansson, Ulrich Schliewen, Emmanuel Vreven, Alexander Papadopoulos, Ilkser Erdem Kiper and their research groups for insightful discussions about this project. 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) 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.

3.7 References

- Altschul SF, Madden TL, Schaffer AA, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402.
- Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews. Genetics* **17**, 81-92.
- Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: <<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>>
- Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology* **22**, 3179-3190.
- Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376.
- Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology* **13**, 729-744.
- Barkhuizen LM (2017) *Labeobarbus kimberleyensis*. The IUCN Red List of Threatened Species 2017: e.T63292A100166441. Downloaded on **07 December 2017**. <<http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T63292A100166441.en>>
- Bloomer P, Bills IR, van der Bank FH, *et al.* (2007) Multidisciplinary investigation of differences and potential hybridization between two yellowfish species *Labeobarbus kimberleyensis* and *L. aeneus* from the Orange-Vaal system. Follow-up study 2004-2007. Yellowfish Working Group.

- Bloomer P, Naran D (2006) Identification of conservation units of two yellowfish species *Labeobarbus kimberleyensis* and *L. aeneus*. Orange-Vaal yellowfish genetics pilot study report for the Yellowfish Working Group and AngloGold Ashanti Ltd, pp. 1-27.
- Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1**, 171-182.
- Catchen JM, Hohenlohe P, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.
- Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology* **9**, 1657-1660.
- Cruaud A, Gautier M, Galan M, *et al.* (2014) Empirical assessment of RAD sequencing for interspecific phylogeny. *Molecular Biology and Evolution* **31**, 1272-1274.
- DaCosta JM, Sorenson MD (2014) Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE* **9**, e106713.
- Davey JW, Cezard T, Fuentes-Utrilla P, *et al.* (2013) Special features of RAD sequencing data: implications for genotyping. *Molecular Ecology* **22**, 3151-3164.
- Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.
- Dowling TE, Tibbets CA, Minckley W, Smith GR, McEachran J (2002) Evolutionary relationships of the plagiopeterns (Teleostei: Cyprinidae) from cytochrome *b* sequences. *Copeia* **2002**, 665-678.
- Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**, 1969-1973.
- Eaton DA, Spriggs EL, Park B, Donoghue MJ (2017) Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology* **66**, 399-412.
- Eccles D (1986) Development of the gut in the South African cyprinid fish *Barbus aeneus* (Burchell). *African Zoology* **21**, 165-169.
- Estoup A, Largiadere C, 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.
- Gaigher I (1976) The reproduction of *Barbus cf. kimberleyensis* (Pisces, Cyprinidae) in the Hardap dam, South West Africa. *African Zoology* **11**, 97-110.
- Gleason LU, Burton RS (2016) Genomic evidence for ecological divergence against a background of population homogeneity in the marine snail *Chlorostoma funebris*. *Molecular Ecology* **25**, 3557-3573.
- Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160-174.
- Herrera S, Watanabe H, Shank TM (2015) Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Molecular Ecology* **24**, 673-689.
- Hunter AMS (2009) 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, Pietermaritzburg (South Africa).
- Jeffries DL, Copp GH, Lawson Handley L, *et al.* (2016) Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Molecular Ecology* **25**, 2997-3018.
- Jezkova T, Riddle BR, Card DC, *et al.* (2015) Genetic consequences of postglacial range expansion in two codistributed rodents (genus *Dipodomys*) depend on ecology and genetic locus. *Molecular Ecology* **24**, 83-97.

- Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772-780.
- Lajbner Z, Šlechtová V, Šlechta V, *et al.* (2009) Rare and asymmetrical hybridization of the endemic *Barbus carpathicus* with its widespread congener *Barbus barbus*. *Journal of Fish Biology* **74**, 418-436.
- Levin DA (1983) Polyploidy and novelty in flowering plants. *The American Naturalist* **122**, 1-25.
- Lischer H, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298-299.
- Macher JN, Rozenberg A, Pauls SU, *et al.* (2015) Assessing the phylogeographic history of the montane caddisfly *Thremma gallicum* using mitochondrial and restriction-site-associated DNA (RAD) markers. *Ecology and Evolution* **5**, 648-662.
- Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* **17**, 240-248.
- Mulder P (1973) Aspects on the ecology of *Barbus kimberleyensis* and *Barbus holubi* in the Vaal River. *African Zoology* **8**, 1-14.
- Mulder P, van Vuuren N, Ferreira J, van der Bank F (1990) A preliminary biochemical comparison of two species of the genus *Barbus* from the Vaal River system. *Water S.A.* **16**, 147-150.
- Oellermann LK, Skelton PH (1990) Hexaploidy in yellowfish species (*Barbus*, Pisces, Cyprinidae) from southern Africa. *Journal of Fish Biology* **37**, 105-115.
- Oswald JA, Overcast I, Mauck WM, Andersen MJ, Smith BT (2017) Isolation with asymmetric gene flow during the nonsynchronous divergence of dry forest birds. *Molecular Ecology* **26**, 1386-1400.
- Pante E, Abdelkrim J, Viricel A, *et al.* (2015) Use of RAD sequencing for delimiting species. *Heredity* **114**, 450-459.
- Paris JR, Stevens JR, Catchen JM (2017) Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution* **8**, 1360-1373.
- Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7**, e37135.
- Puckett EE, Etter PD, Johnson EA, Eggert LS (2015) Phylogeographic analyses of American black bears (*Ursus americanus*) suggest four glacial refugia and complex patterns of post-glacial admixture. *Molecular Biology and Evolution* **32**, 2338-2350.
- Pujolar JM, Jacobsen M, Als TD, *et al.* (2014) Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology* **23**, 2514-2528.
- Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574.
- Skelton PH (1986) Fish of the Orange-Vaal system. In: *The Ecology of River Systems* (eds. Davies BR, Walker KF), pp. 143-162. Dr W. Junk Publishers, Dordrecht, The Netherlands.
- Skelton PH (2001) *A Complete Guide to the Freshwater Fishes of Southern Africa* Struik, Cape Town.
- Skelton PH, Bills R (2008) An Introduction to African Yellowfish and to this Report. In: *Technical Report on the State of Yellowfishes in South Africa 2007* (eds. Impson ND, Bills IR, Wolhuter L). Water Research Commission, Pretoria.
- Streicher JW, Devitt TJ, Goldberg CS, *et al.* (2014) Diversification and asymmetrical gene flow across time and space: lineage sorting and hybridization in polytypic barking frogs. *Molecular Ecology* **23**, 3273-3291.
- Tamura K, Stecher G, Peterson D, Filipowski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* **30**, 2725-2729.

- Tómasson T, Cambray J, Jackson P (1984) Reproductive biology of four large riverine fishes (Cyprinidae) in a man-made lake, Orange River, South Africa. *Hydrobiologia* **112**, 179-195.
- Truong HT, Ramos AM, Yalcin F, *et al.* (2012) Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS ONE* **7**, e37565.
- Tsigenopoulos CS, Kasapidis P, Berrebi P (2010) Phylogenetic relationships of hexaploid large-sized barbs (genus *Labeobarbus*, Cyprinidae) based on mtDNA data. *Molecular Phylogenetics and Evolution* **56** 851-856.
- Van de Peer Y, Mizrahi E, Marchal K (2017) The evolutionary significance of polyploidy. *Nature Reviews Genetics* **18**, 411-424.
- van Vuuren N, Mulder P, Ferreira J, van der Bank FH (1989) The identification of hybrids of *Barbus aeneus* X *B. kimberleyensis* and *Labeo capensis* X *L. umbratus* in Hardap Dam, SWA/Namibia. *Madoqua* **16**, 27-34.
- Vreven EJ, Musschoot T, Snoeks J, Schliewen UK (2016) The African hexaploid Torini (Cypriniformes: Cyprinidae): review of a tumultuous history. *Zoological Journal of the Linnean Society* **177**, 231-305.
- Yang L, Sado T, Hirt MV, *et al.* (2015) Phylogeny and polyploidy: resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). *Molecular Phylogenetics and Evolution* **85**, 97-116.
- Zhan S, Glick L, Tsigenopoulos C, Otto S, Mayrose I (2014) Comparative analysis reveals that polyploidy does not decelerate diversification in fish. *Journal of Evolutionary Biology* **27**, 391-403.

3.8 Supporting Information

Table S3.1 Detailed sampling information for all samples of *L. aeneus*, *L. kimberleyensis* and *L. natalensis* used in this study.

Sample	Species	Library	Locality	Drainage	Co-ordinates	Date	Collector/s
IRB-005	<i>L. aeneus</i>	Orange-Vaal 2	Noordoewer	Orange	28° 44' 50.0" S 17° 36' 32.0" E	21-10-2006	R. Bills
IRB-010	<i>L. kimberleyensis</i>	Orange-Vaal 2	Noordoewer	Orange	28° 44' 50.0" S 17° 36' 32.0" E	21-10-2006	R. Bills
IRB-018	<i>L. aeneus</i>	Orange-Vaal 2	Noordoewer - Felix Unite Camp	Orange	28° 41' 19.0" S 17° 33' 20.0" E	20-10-2006	R. Bills
IRB-028	<i>L. kimberleyensis</i>	Orange-Vaal 2	Gamkab	Orange	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006	R. Bills
IRB-029	<i>L. kimberleyensis</i>	Orange-Vaal 2	Gamkab	Orange	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006	R. Bills
IRB-031	<i>L. aeneus</i>	Orange-Vaal 2	Gamkab	Orange	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006	R. Bills
IRB-041	<i>L. kimberleyensis</i>	Orange-Vaal 2	Gamkab	Orange	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006	R. Bills
IRB-044	<i>L. aeneus</i>	Orange-Vaal 2	Gamkab	Orange	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006	R. Bills
IRB-047	<i>L. kimberleyensis</i>	Orange-Vaal 2	Gamkab	Orange	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006	R. Bills
IRB-048	<i>L. aeneus</i>	Orange-Vaal 2	Gamkab	Orange	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006	R. Bills
IRB-081	<i>L. kimberleyensis</i>	Orange-Vaal 2	Houms River	Orange	28° 51' 05" S 18° 36' 52" E	18-10-2006	R. Bills
IRB-083	<i>L. aeneus</i>	Orange-Vaal 2	Houms River	Orange	28° 51' 05" S 18° 36' 52" E	18-10-2006	R. Bills
IRB-089	<i>L. aeneus</i>	Orange-Vaal 2	Houms River	Orange	28° 51' 05" S 18° 36' 52" E	18-10-2006	R. Bills
IRB-097	<i>L. kimberleyensis</i>	Orange-Vaal 2	Houms River	Orange	28° 51' 05" S 18° 36' 52" E	18-10-2006	R. Bills
IRB-098	<i>L. aeneus</i>	Orange-Vaal 2	Houms River	Orange	28° 51' 05" S 18° 36' 52" E	18-10-2006	R. Bills
IRB-215	<i>L. aeneus</i>	Orange-Vaal 2	Houms River	Orange	28° 51' 05" S 18° 36' 52" E	18-10-2006	R. Bills
IRB-225	<i>L. aeneus</i>	Orange-Vaal 2	Noordoewer	Orange	28° 45' 20.0" S 17° 36' 46.0" E	20-10-2006	R. Bills
IRB-257	<i>L. kimberleyensis</i>	Orange-Vaal 2	Noordoewer	Orange	28° 45' 20.0" S 17° 36' 46.0" E	20-10-2006	R. Bills
IRB-261	<i>L. kimberleyensis</i>	Orange-Vaal 2	Houms River	Orange	28° 51' 05" S 18° 36' 52" E	19-10-2006	R. Bills
IRB-266	<i>L. kimberleyensis</i>	Orange-Vaal 2	Houms River	Orange	28° 51' 05" S 18° 36' 52" E	19-10-2006	R. Bills
IRB-277	<i>L. aeneus</i>	Orange-Vaal 2	Gamkab rapids at camp	Orange	28° 14' 30.0" S 17° 20' 49.0" E	23-10-2006	R. Bills
IRB-284	<i>L. aeneus</i>	Orange-Vaal 2	Gamkab rapids at camp	Orange	28° 14' 30.0" S 17° 20' 49.0" E	23-10-2006	R. Bills
IRB-359	<i>L. kimberleyensis</i>	Orange-Vaal 2	Gamkab	Orange	28° 14' 30.0" S 17° 20' 49.0" E	23-10-2006	R. Bills
IRB-401	<i>L. kimberleyensis</i>	Orange-Vaal 2	Orange Mouth	Orange	28° 36' 02.9" S 16° 27' 23.5" E	26-10-2006	R. Bills
IRB-403	<i>L. aeneus</i>	Orange-Vaal 2	Orange Mouth	Orange	28° 36' 02.9" S 16° 27' 23.5" E	26-10-2006	R. Bills
IRB-447	<i>L. aeneus</i>	Orange-Vaal 2	Orange Mouth	Orange	28° 36' 02.9" S 16° 27' 23.5" E	26-10-2006	R. Bills
LkA06	<i>L. kimberleyensis</i>	Orange-Vaal 2	Aliwal North, weir 1 km upstream of N6 road bridge	Orange	30° 40' 45" S 26° 43' 11" E	22-01-2004	R. Bills
LaA13	<i>L. aeneus</i>	Orange-Vaal 2	Aliwal North, weir 1 km upstream of N6 road bridge	Orange	30° 40' 45" S 26° 43' 11" E	22-01-2004	R. Bills
LaA15	<i>L. aeneus</i>	Orange-Vaal 1	Aliwal North, weir 1 km upstream of N6 road bridge	Orange	30° 40' 45" S 26° 43' 11" E	22-01-2004	R. Bills
LaA22	<i>L. aeneus</i>	Orange-Vaal 1	upstream of N6 road bridge	Orange	30° 40' 45" S 26° 43' 11" E	22-01-2004	R. Bills
LaO02	<i>L. aeneus</i>	Orange-Vaal 1	Onseepkans farm	Orange	28° 44' 31" S 19° 20' 07" E	25-01-2004	R. Bills
LaO05	<i>L. aeneus</i>	Orange-Vaal 1	Onseepkans farm	Orange	28° 44' 31" S 19° 20' 07" E	25-01-2004	R. Bills
LaO08	<i>L. aeneus</i>	Orange-Vaal 1	Onseepkans farm	Orange	28° 44' 31" S 19° 20' 07" E	25-01-2004	R. Bills
LaO10	<i>L. aeneus</i>	Orange-Vaal 1	Onseepkans farm	Orange	28° 44' 31" S 19° 20' 07" E	25-01-2004	R. Bills
LaO11	<i>L. aeneus</i>	Orange-Vaal 1	Onseepkans farm	Orange	28° 44' 31" S 19° 20' 07" E	25-01-2004	R. Bills
LaO13	<i>L. aeneus</i>	Orange-Vaal 1	Pella Drift upper site	Orange	28° 57' 39" S 19° 09' 50" E	26-01-2004	R. Bills
LaO17	<i>L. aeneus</i>	Orange-Vaal 1	Pella Drift upper site	Orange	28° 57' 39" S 19° 09' 50" E	26-01-2004	R. Bills
LkO06	<i>L. kimberleyensis</i>	Orange-Vaal 1	Pella Drift upper site	Orange	28° 57' 39" S 19° 09' 50" E	28-01-2004	R. Bills
LkO07	<i>L. kimberleyensis</i>	Orange-Vaal 1	Pella Drift upper site	Orange	28° 57' 39" S 19° 09' 50" E	28-01-2004	R. Bills
LnBL003	<i>L. natalensis</i>	KZN 1	Black Umfolozi River	Umfolozi	27° 56' 18.7" S 31° 12' 34.0" E	28-11-2007	R.J. Karssing, M. Nkosi & N. Rivers-Moore
LnBL004	<i>L. natalensis</i>	KZN 1	Black Umfolozi River	Umfolozi	27° 56' 18.7" S 31° 12' 34.0" E	28-11-2007	R.J. Karssing, M. Nkosi & N. Rivers-Moore
LnBU002	<i>L. natalensis</i>	KZN 1	Buffalo River - Tayside	Tugela	28° 03' 34.5" S 30° 22' 24.8" E	27-11-2007	R.J. Karssing, M. Nkosi & N. Rivers-Moore
LnBU008	<i>L. natalensis</i>	KZN 1	Buffalo River - Tayside	Tugela	28° 03' 34.5" S 30° 22' 24.8" E	27-11-2007	R.J. Karssing, M. Nkosi & N. Rivers-Moore
LnKN017	<i>L. natalensis</i>	KZN 1	Umzimkhulu, 2km above Polela junction	Umzimkhulu	29° 57' 36.0" S 29° 41' 24.0" E	12-05-2003	R. Arderne
LnKN018	<i>L. natalensis</i>	KZN 1	Umzimkhulu, 2km above Polela junction	Umzimkhulu	29° 57' 36.0" S 29° 41' 24.0" E	12-05-2003	R. Arderne
LnKN212	<i>L. natalensis</i>	KZN 1	Lions River	Tugela	29° 25' 12.0" S 30° 00' 36.0" E	17-08-2003	T. Wilkinson & H. Plank
LnKN315	<i>L. natalensis</i>	KZN 1	Mkomaas, below Josephine's bridge	Mkomaas	30° 00' 36.0" S 30° 09' 00.0" E	30-08-2003	H. Plank
LnKZ223	<i>L. natalensis</i>	KZN 1	Mkuze River, Farm Verdukt (444)	Umfolozi	27° 40' 48.0" S 31° 26' 24.0" E	22-08-2003	H.E. Filter
LnMB007	<i>L. natalensis</i>	KZN 1	Mbokodweni River, Durban Metro	Mbokodweni	30° 00' 33.3" S 30° 53' 24.7" E	17-05-2006	R.J. Karssing
LnMB008	<i>L. natalensis</i>	KZN 1	Mbokodweni River, Durban Metro	Mbokodweni	30° 00' 33.3" S 30° 53' 24.7" E	17-05-2006	R.J. Karssing
LnMB009	<i>L. natalensis</i>	KZN 1	Mbokodweni River, Durban Metro	Mbokodweni	30° 00' 33.3" S 30° 53' 24.7" E	17-05-2006	R.J. Karssing
LnMB010	<i>L. natalensis</i>	KZN 1	Mbokodweni River, Durban Metro	Mbokodweni	30° 00' 33.3" S 30° 53' 24.7" E	17-05-2006	R.J. Karssing
LnMB011	<i>L. natalensis</i>	KZN 1	Mbokodweni River, Durban Metro	Mbokodweni	30° 00' 33.3" S 30° 53' 24.7" E	17-05-2006	R.J. Karssing
LnMD003	<i>L. natalensis</i>	KZN 1	Mdloti River, North Coast, Durban	Umgeni	29° 36' 09.6" S 31° 00' 31.5" E	28-03-2006	R.J. Karssing
LnMK001	<i>L. natalensis</i>	KZN 1	Mkomazi River	Mkomaas	30° 08' 17.3" S 30° 40' 28.4" E	16-05-2006	R.J. Karssing
LnMK002	<i>L. natalensis</i>	KZN 1	Mkomazi River	Mkomaas	30° 08' 17.3" S 30° 40' 28.4" E	16-05-2006	R.J. Karssing
LnMK003	<i>L. natalensis</i>	KZN 1	Mkomazi River	Mkomaas	30° 08' 17.3" S 30° 40' 28.4" E	16-05-2006	R.J. Karssing
LnMK004	<i>L. natalensis</i>	KZN 1	Mkomazi River	Mkomaas	30° 08' 17.3" S 30° 40' 28.4" E	16-05-2006	R.J. Karssing
LnML004	<i>L. natalensis</i>	KZN 1	Molweni River, Durban Metro	Umgeni	29° 45' 06.3" S 30° 53' 21.1" E	11-05-2006	R.J. Karssing
LnMO007	<i>L. natalensis</i>	KZN 1	Mooi River, Winterhoek Weir	Tugela	29° 02' 58.4" S 30° 18' 13.8" E	04-05-2006	R.J. Karssing
LnTU002	<i>L. natalensis</i>	KZN 1	DWAF gauging weir	Tugela	28° 39' 09.9" S 29° 02' 36.7" E	23-03-2006	R.J. Karssing
LnWU008	<i>L. natalensis</i>	KZN 1	White Umfolozi River, Lenjanedrif, Vryheid	Umfolozi	27° 56' 06.1" S 30° 57' 18.7" E	28-11-2007	R.J. Karssing, M. Nkosi & N. Rivers-Moore

Table S3.2 *In silico* restriction digestion of the *Labeobarbus intermedius* mitogenome with the restriction enzymes *Mlu*CI and *Nla*III to predict expected coverage of this approach in South African *Labeobarbus*. Gap size is calculated by comparing the value of 'Below' for that position to the value of 'Above' for the previous position. If Below is greater, the gap is equal to Below - Above.

Position	Above (+80bp)	Below (-80bp)	Enzyme	Recognition sequence	Gaps
14	94	-62	NlaIII	AATT	0
64	144	-12	MluCI	CATG	0
76	156	0	MluCI	CATG	0
116	196	40	MluCI	CATG	0
232	312	156	NlaIII	AATT	0
382	462	306	NlaIII	AATT	0
545	625	469	NlaIII	AATT	7
770	850	694	MluCI	CATG	69
827	907	751	MluCI	CATG	0
1039	1119	963	NlaIII	AATT	56
1076	1156	1000	MluCI	CATG	0
1082	1162	1006	NlaIII	AATT	0
1315	1395	1239	MluCI	CATG	77
1511	1591	1435	MluCI	CATG	40
1553	1633	1477	MluCI	CATG	0
1563	1643	1487	NlaIII	AATT	0
1755	1835	1679	MluCI	CATG	36
1763	1843	1687	MluCI	CATG	0
1772	1852	1696	NlaIII	AATT	0
2185	2265	2109	NlaIII	AATT	257
2305	2385	2229	MluCI	CATG	0
2422	2502	2346	MluCI	CATG	0
2710	2790	2634	NlaIII	AATT	132
2794	2874	2718	MluCI	CATG	0
2801	2881	2725	NlaIII	AATT	0
2882	2962	2806	NlaIII	AATT	0
2939	3019	2863	NlaIII	AATT	0
3176	3256	3100	NlaIII	AATT	81
3218	3298	3142	NlaIII	AATT	0
3243	3323	3167	NlaIII	AATT	0
3317	3397	3241	NlaIII	AATT	0
3641	3721	3565	NlaIII	AATT	168
3783	3863	3707	MluCI	CATG	0
3916	3996	3840	NlaIII	AATT	0
4016	4096	3940	MluCI	CATG	0
4145	4225	4069	NlaIII	AATT	0
4160	4240	4084	NlaIII	AATT	0
4358	4438	4282	NlaIII	AATT	42
4427	4507	4351	NlaIII	AATT	0
4438	4518	4362	MluCI	CATG	0
4517	4597	4441	NlaIII	AATT	0
4613	4693	4537	NlaIII	AATT	0
4649	4729	4573	NlaIII	AATT	0
4721	4801	4645	NlaIII	AATT	0
4886	4966	4810	MluCI	CATG	9
4904	4984	4828	NlaIII	AATT	0
5177	5257	5101	NlaIII	AATT	117
5634	5714	5558	NlaIII	AATT	301
5736	5816	5660	NlaIII	AATT	0
5751	5831	5675	MluCI	CATG	0
5916	5996	5840	NlaIII	AATT	9
5961	6041	5885	NlaIII	AATT	0
5982	6062	5906	NlaIII	AATT	0
6162	6242	6086	NlaIII	AATT	24
6234	6314	6158	NlaIII	AATT	0
6244	6324	6168	MluCI	CATG	0
6408	6488	6332	NlaIII	AATT	8
6471	6551	6395	NlaIII	AATT	0
6543	6623	6467	NlaIII	AATT	0
6580	6660	6504	MluCI	CATG	0
6783	6863	6707	MluCI	CATG	47
6855	6935	6779	NlaIII	AATT	0
6935	7015	6859	NlaIII	AATT	0
7014	7094	6938	NlaIII	AATT	0
7024	7104	6948	NlaIII	AATT	0
7043	7123	6967	NlaIII	AATT	0
7121	7201	7045	NlaIII	AATT	0
7165	7245	7089	MluCI	CATG	0
7257	7337	7181	MluCI	CATG	0
7263	7343	7187	MluCI	CATG	0
7286	7366	7210	NlaIII	AATT	0
7373	7453	7297	NlaIII	AATT	0
7412	7492	7336	NlaIII	AATT	0
7433	7513	7357	NlaIII	AATT	0
7454	7534	7378	NlaIII	AATT	0
7475	7555	7399	NlaIII	AATT	0
7550	7630	7474	MluCI	CATG	0

7582	7662	7506	NlaIII	AATT	0
7811	7891	7735	NlaIII	AATT	73
7963	8043	7887	NlaIII	AATT	0
7988	8068	7912	NlaIII	AATT	0
8113	8193	8037	MluCI	CATG	0
8136	8216	8060	NlaIII	AATT	0
8170	8250	8094	NlaIII	AATT	0
8182	8262	8106	NlaIII	AATT	0
8227	8307	8151	NlaIII	AATT	0
8365	8445	8289	MluCI	CATG	0
8554	8634	8478	NlaIII	AATT	33
8812	8892	8736	MluCI	CATG	102
8840	8920	8764	MluCI	CATG	0
8937	9017	8861	NlaIII	AATT	0
9230	9310	9154	MluCI	CATG	137
9415	9495	9339	MluCI	CATG	29
9462	9542	9386	NlaIII	AATT	0
9661	9741	9585	NlaIII	AATT	43
9676	9756	9600	NlaIII	AATT	0
9700	9780	9624	NlaIII	AATT	0
9859	9939	9783	NlaIII	AATT	3
9940	10020	9864	NlaIII	AATT	0
9961	10041	9885	NlaIII	AATT	0
10047	10127	9971	MluCI	CATG	0
10060	10140	9984	MluCI	CATG	0
10220	10300	10144	NlaIII	AATT	4
10324	10404	10248	MluCI	CATG	0
10388	10468	10312	MluCI	CATG	0
10570	10650	10494	MluCI	CATG	26
10682	10762	10606	NlaIII	AATT	0
10930	11010	10854	MluCI	CATG	92
11131	11211	11055	NlaIII	AATT	45
11174	11254	11098	MluCI	CATG	0
11219	11299	11143	NlaIII	AATT	0
11281	11361	11205	MluCI	CATG	0
11312	11392	11236	NlaIII	AATT	0
11390	11470	11314	NlaIII	AATT	0
11414	11494	11338	NlaIII	AATT	0
11533	11613	11457	MluCI	CATG	0
11540	11620	11464	NlaIII	AATT	0
11825	11905	11749	NlaIII	AATT	129
11855	11935	11779	MluCI	CATG	0
11966	12046	11890	NlaIII	AATT	0
12065	12145	11989	NlaIII	AATT	0
12202	12282	12126	NlaIII	AATT	0
12265	12345	12189	NlaIII	AATT	0
12307	12387	12231	NlaIII	AATT	0
12344	12424	12268	NlaIII	AATT	0
12506	12586	12430	NlaIII	AATT	6
12544	12624	12468	MluCI	CATG	0
12551	12631	12475	NlaIII	AATT	0
12640	12720	12564	NlaIII	AATT	0
12863	12943	12787	NlaIII	AATT	67
13085	13165	13009	NlaIII	AATT	66
13202	13282	13126	NlaIII	AATT	0
13289	13369	13213	NlaIII	AATT	0
13397	13477	13321	NlaIII	AATT	0
13502	13582	13426	NlaIII	AATT	0
13630	13710	13554	MluCI	CATG	0
13715	13795	13639	NlaIII	AATT	0
13769	13849	13693	NlaIII	AATT	0
13988	14068	13912	NlaIII	AATT	63
14161	14241	14085	MluCI	CATG	17
14300	14380	14224	NlaIII	AATT	0
14402	14482	14326	NlaIII	AATT	0
14408	14488	14332	NlaIII	AATT	0
14495	14575	14419	NlaIII	AATT	0
14528	14608	14452	MluCI	CATG	0
14592	14672	14516	NlaIII	AATT	0
14657	14737	14581	MluCI	CATG	0
14807	14887	14731	NlaIII	AATT	0
15023	15103	14947	NlaIII	AATT	60
15071	15151	14995	NlaIII	AATT	0
15337	15417	15261	NlaIII	AATT	110
15368	15448	15292	NlaIII	AATT	0
15379	15459	15303	MluCI	CATG	0
15505	15585	15429	NlaIII	AATT	0
15631	15711	15555	NlaIII	AATT	0
15659	15739	15583	NlaIII	AATT	0
15696	15776	15620	NlaIII	AATT	0
15715	15795	15639	NlaIII	AATT	0
15769	15849	15693	NlaIII	AATT	0
15851	15931	15775	NlaIII	AATT	0
15888	15968	15812	NlaIII	AATT	0
15948	16028	15872	MluCI	CATG	0
15952	16032	15876	MluCI	CATG	0
16034	16114	15958	NlaIII	AATT	0
16055	16135	15979	NlaIII	AATT	0

16091	16171	16015	NlaIII	AATT	0
16106	16186	16030	MluCI	CATG	0
16232	16312	16156	NlaIII	AATT	0
16248	16328	16172	NlaIII	AATT	0
16256	16336	16180	NlaIII	AATT	0
16480	16560	16404	NlaIII	AATT	68
16496	16576	16420	NlaIII	AATT	0
16526	16606	16450	NlaIII	AATT	0
16541	16621	16465	NlaIII	AATT	0
Sum					2653
% Coverage					0,83992

Table S3.3 Final step of parameter optimisation as per Paris *et al.* (2017) for *L. natalensis* where statistics are generated for different values of $-n$ as it is varied from the default parameter set of $-m$ 5 $-M$ 3 $-n$ 2. These statistics are generated using the script COUNT_FIXED_SNPS.PY as well as the populations module of STACKS where SNPs are called only if they occur in at least 80% of samples ($-r$ 0.8). Although Paris *et al.* (2017) suggest selecting the dataset with the greatest number of polymorphic loci ($-n = 1$ here), I opted to select a value of $-n = -M = 2$ due to some of the technical aspects of the data (such as prevalence of overlapping paired reads, adapter pollution, and many paralogous loci as a result of polyploidy).

count_fixed_snps.py		$m5M2n1$ ($-n = -M - 1$)	$m5M2n2$ ($-n = -M$)	$m5M2n3$ ($-n = -M + 1$)
Heterozygous SNPs per sample	MKO003	1648	1724	1738
	WUR008	1095	1129	1151
	MBO010	846	908	909
	MBO009	680	701	695
	BLB004	868	865	853
	KNA017	125	136	146
	MKO004	764	767	771
	TUG002	728	760	769
	BLA004	648	630	622
	KNT212	443	451	443
	MDL003	585	597	596
	BUF008	828	817	810
	MOO007	474	486	485
	MBO011	647	661	653
	KZF223	356	364	358
	MBO008	560	558	566
	BUF002	519	508	498
	KNU018	442	446	439
	KNB017	1695	1667	1637
	MOL004	530	542	543
	BLU003	386	382	382
	MKO002	281	270	274
	MKO001	702	702	699
	MBO007	171	169	171
	KNK315	313	296	285
	Variable sites across population	16334	16536	16493
	Heterozygous SNPs in the catalog	17061	20049	21123
	SNPs across all samples and in catalog	16743	19161	20061
	Fixed SNPs only in the catalog	409	2625	3568
Populations ($-r = 0.8$)	Number of loci	1184	1059	1023
	Number of polymorphic loci	661	613	595
	Number of SNPs	1541	1408	1376

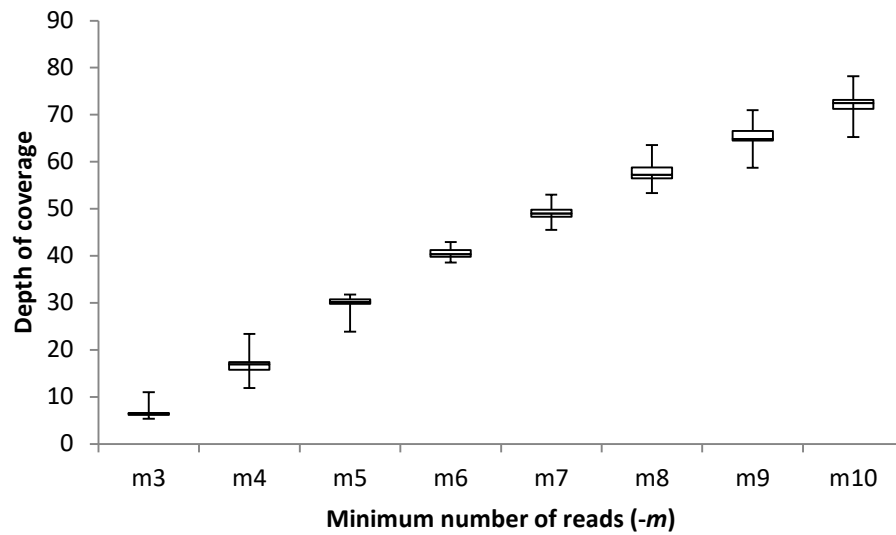


Fig. S3.1 Mean depth of coverage across all 23 samples of *L. natalensis* (plus two replicates) as the minimum number of reads required to call a stack (-m) is varied. This methodology is as per Paris *et al.* (2017). The default parameter set used was -m 5 -M 3 -n 2. Whiskers here indicate the 10th and 90th percentiles.

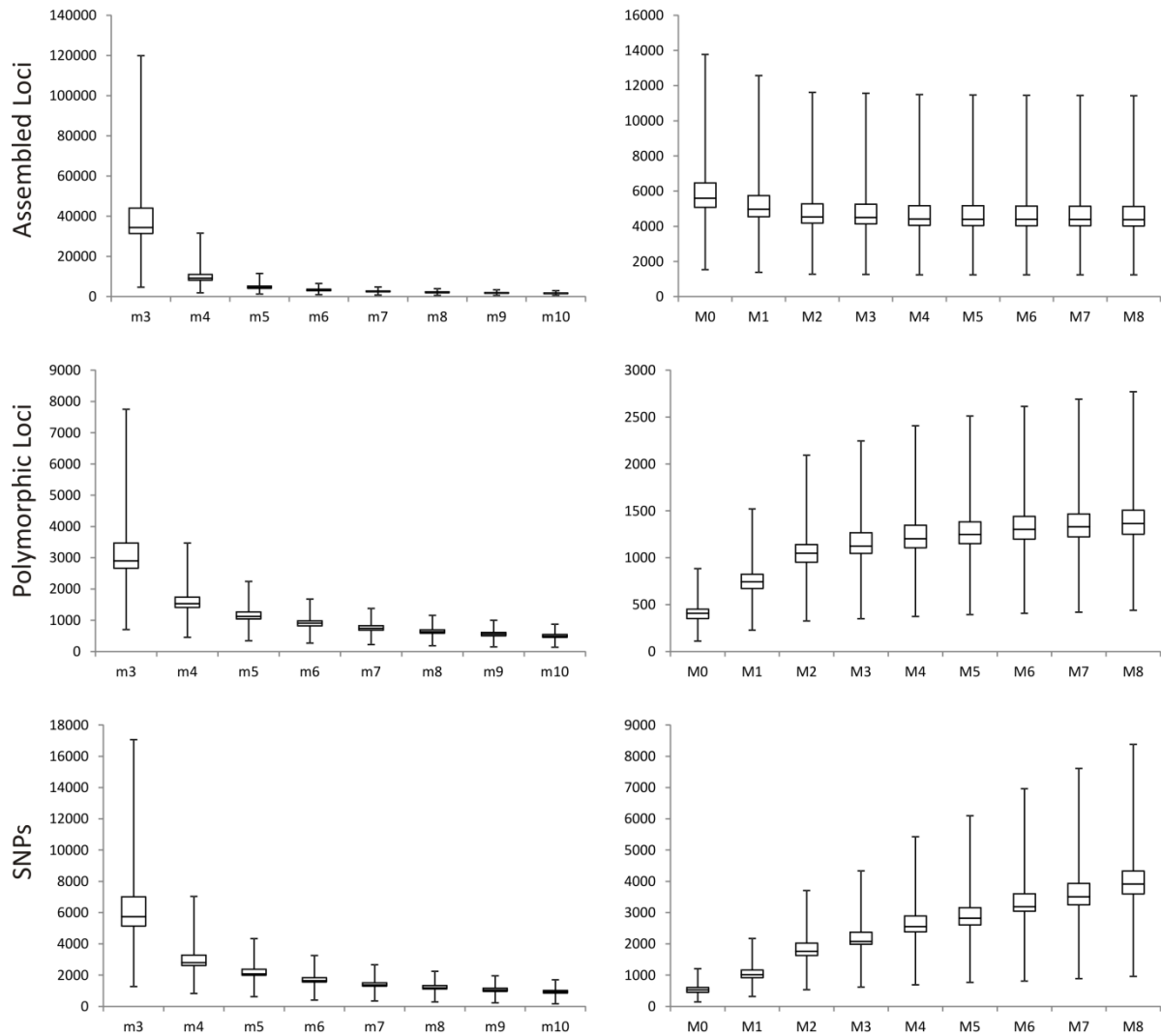


Fig. S3.2 Parameter optimisation protocol for *L. natalensis* as per Paris *et al.* (2017) showing the number of assembled loci, polymorphic loci, and SNPs as different values are tested for $-m$ and $-M$ from the default parameter set of $-m$ 5 $-M$ 3 $-n$ 2. The distribution given by the box-and-whisker plots illustrate the results across all 23 samples (plus two replicates), where the whiskers indicate minimum and maximum values. The trend for $-M$ begins to reach a plateau at $-M = 2$.

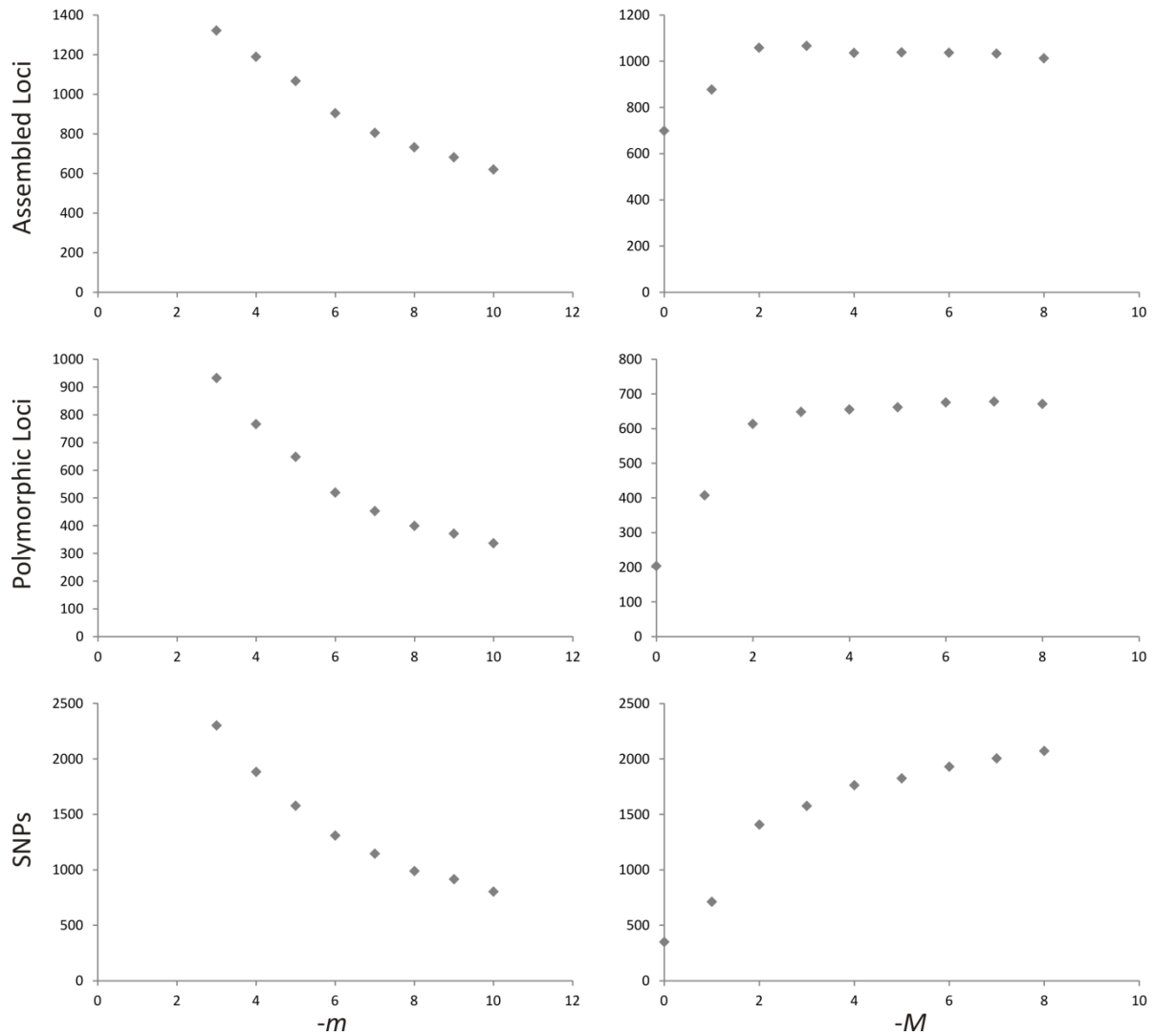


Fig. S3.3 Extension of the parameter optimisation protocol by Paris *et al.* (2017) for *L. natalensis* where the number of assembled loci, polymorphic loci, and SNPs are recorded at the population level after enforcing a threshold that loci have to occur within at least 80% of the samples in the population. Different values of $-m$ and $-M$ were tested in this way using a default parameter set of $-m$ 5 $-M$ 3 $-n$ 2. A plateau was again observed for $-M$ starting at $-M = 2$.

Chapter 4: Investigating hybridisation and introgression of the Orange-Vaal yellowfish species *Labeobarbus aeneus* and *L. kimberleyensis*

Abstract

Loss of physical isolation between co-occurring species is a common consequence of habitat loss caused by anthropogenic activities. However, identification of the resulting hybridisation events between these species is often challenging, particularly in polyploid genera. In this study, genetic differentiation, hybridisation and introgression of two sympatric yellowfish species (*Labeobarbus aeneus* and *L. kimberleyensis*) were investigated using a combination of nuclear and mitochondrial genomic sequences extracted using Restriction-site Associated DNA (RAD) sequencing. A bias of in-frame coding regions due to the choice of restriction enzymes was observed, and library bias manifesting as genetic differentiation between the two libraries, each containing samples of both species. Despite being morphologically and ecologically distinct, low levels of differentiation against a background of genetic homogeneity were observed between and within the two species. Nonetheless, one sample was found to have been misidentified, and another was determined to be a recent hybrid between the two species. Mitogenome analysis indicated a distinct clade for each species, with five samples possessing mitogenome haplotypes different to the species to which they were morphologically assigned. At a nuclear level, the differentiation between species was found to be driven by loci potentially under selection against a background of relative homogeneity for neutral loci. No substructuring within species was identified, except for possible mitogenome capture in *L. aeneus* from the upper Orange River. Although hybridisation was found to be potentially unidirectional using mitogenome mappings, nuclear SNPs suggested bidirectional gene flow with a slight bias for the direction observed in mitogenome analyses. In conclusion, I show

several lines of evidence that hybridisation and introgression are occurring between two genetically-distinct yellowfish species.

4.1 Introduction

The family Cyprinidae, carps and minnows, is the largest and most diverse fish family with over 3,100 currently-recognised species found across the world (Eschmeyer & Fong 2017). This diversity is also expressed at a karyotypic level, as the family includes numerous examples of diploid, tetraploid, and hexaploid lineages, which have arisen multiple times throughout its history (Yang *et al.* 2015). One of these hexaploid genera, *Labeobarbus*, is endemic to Africa and has recently become a popular target for genetic research due to its complicated genetic history, popularity as a game and subsistence fish, as well as its abundance and diversity of species with 125 currently recognised (Yang *et al.* 2015; Vreven *et al.* 2016). This genus was recently grouped into the newly-erected tribe Torini, comprising the genera *Tor*, *Carasobarbus* and *Neolissochilus* from Asia and Africa (Yang *et al.* 2015). The ploidy levels of the tribe members combined with their current distributions suggest an advent of allopolyploidy between a tetraploid member of the tribe, prior to the divergence of *Carasobarbus* and *Labeobarbus*, and a diploid ancestor of the *Cyprinion* genus, resulting in the current hexaploid lineage (Fig. 1.1) (Oellermann & Skelton 1990; Yang *et al.* 2015).

There are currently seven species of *Labeobarbus* or “yellowfish” found in South Africa (Vreven *et al.* 2016). The most popular and abundant of these is arguably the smallmouth Orange-Vaal yellowfish, *L. aeneus*, which co-occurs across its distribution with the largemouth Orange-Vaal yellowfish, *L. kimberleyensis* (Gaigher 1976; van Vuuren *et al.* 1989; Skelton 2001). Although they co-occur, the two species fill very different ecological roles - *L. aeneus* is an abundant generalist feeder, whereas *L. kimberleyensis* is a top-level piscivore (Mulder 1973; Eccles 1986). The two species also

differ in colour and morphometry (Skelton 2001; Bloomer *et al.* 2007), although it has been noted that it can be challenging to differentiate between them on external morphological characters alone (Mulder *et al.* 1990). Previous research has also shown differences in gut sizes (Kruger & Mulder 1973; Eccles 1986) as well as preferred habitat and home range size (O'Brien *et al.* 2013), which is to be expected given their differing diets (Mulder 1973; O'Brien *et al.* 2013). These species also differ in abundance - *L. aeneus* is currently listed as a species of Least Concern, whereas *L. kimberleyensis* is recognised as Near-Threatened (Barkhuizen 2017; Impson 2017). The latter species' extended period of development to reproductive maturity (6 years in males, 8 in females) is thought to contribute to differences in abundance (Mulder 1973; Skelton 2001).

Despite these differences between the Orange-Vaal yellowfish species, numerous reports have been made over the last few decades of suspected hybrids appearing (Gaigher 1976; Eccles 1986; van Vuuren *et al.* 1989; Mulder *et al.* 1990). These hybrids are usually sighted in man-made environments such as dams (Gaigher 1976; Eccles 1986; van Vuuren *et al.* 1989; Mulder *et al.* 1990), where it is possible that the species do not have access to their preferred breeding grounds and are forced to use the same area. The occurrence of hybridisation in riverine habitat was assessed in a previous study using a combination of morphometrics, mitochondrial haplotypes and allozymes (Bloomer *et al.* 2007). Although consistent morphological differences were found between the two species (with further divergence between upper and lower Orange River *L. aeneus*), no allozymes were identified as diagnostic markers between the two species and substantial mitochondrial allele sharing was observed (Bloomer *et al.* 2007). These genetic observations were at odds with the distinct morphological split observed in that study but could be explained either by extensive introgression and/or sharing of ancestral polymorphisms among two recently-diverged species (Bloomer *et al.* 2007), or divergence during development into two different ecotypes due to environmental factors in a process similar to “shoot carp / tobi koi” (Nakamura & Kasahara 1955;

Nakamura & Kasahara 1956; Nakamura & Kasahara 1961; Wohlfarth 1977; Eccles 1986). “Shoot carp” occur when carp are grown together at particular densities and food is of a specific size or insufficient for all individuals within a population, resulting in some of the carp developing at a much more rapid rate and often demonstrating cannibalism (Nakamura & Kasahara 1955; Nakamura & Kasahara 1956; Nakamura & Kasahara 1961). If this is the case here then it may suggest that the Orange-Vaal yellowfishes are a single species that differentiates early in development into two ecotypes, or on a smaller scale, fish that appear as “intermediates” between the two species could merely be *L. aeneus* “shoot” fish, which would appear more like *L. kimberleyensis* (Eccles 1986).

It is important to determine whether introgression is occurring within the Orange-Vaal yellowfishes for their effective conservation management. The difference in IUCN status for each species places even more of an emphasis on this, as the rarer *L. kimberleyensis* could be at risk of extinction by hybridisation (Rhymer & Simberloff 1996). Because hybrids have almost solely been recorded in dams, this may indicate that the habitat modification caused by the man-made structures has resulted in a breakdown of physical isolation between the two species in turn promoting hybridization events (Rhymer & Simberloff 1996). This may imply that hybridisation is only occurring in these areas, meaning that the majority of the Orange-Vaal river system should retain the two distinct species as reproductive isolation is maintained. However, because of the admixture occurring in the dams followed by dispersal and mating of hybrid offspring, one would anticipate further ramifications throughout the system.

A useful step in identifying hybridisation and introgression genetically, following morphological assessment of individuals or characterisation of parental populations using genomic markers, is to assess mitochondrial haplotypes (Rhymer & Simberloff 1996). This also allows the identification of directional hybridisation, hybridisation occurring only from the one species into the other (Rhymer & Simberloff 1996). It is imperative to also assess the nuclear genome as the mitochondrion can

sometimes demonstrate a different history. Introgression of mitochondrial DNA occurs more frequently and in extreme cases mitochondrial introgression with a complete lack of nuclear introgression have been recorded (Rhymer & Simberloff 1996; Ballard & Whitlock 2004). However, nuclear genetic analyses within these species are complicated by the hexaploid nature of the *Labeobarbus* genome.

Double-digest Restriction-site Associated DNA (ddRAD) sequencing is a reduced-representation approach that utilises a pair of restriction sites to cut throughout the nuclear genome (Peterson *et al.* 2012). The fragments are then size-selected, and both ends are sequenced. I previously used this technology to investigate the phylogeography of *L. natalensis* and found that most fragments sequenced using this approach behaved as diploid loci (Chapter 2). This therefore represents an effective methodology to bypass the complications caused by hexaploidy in this genus, although careful filtering and optimisation is paramount. In this chapter I investigate hybridisation and introgression between *L. aeneus* and *L. kimberleyensis* in natural riverine habitats throughout the Orange River using nuclear SNPs and mitogenome markers identified using ddRAD sequencing.

4.2 Materials and methods

4.2.1 Sample collection and DNA extraction

Samples were collected from ten sites along the Orange River between January 2004 and October 2006 by Roger Bills, with the majority of these sites on the South Africa/Namibia border (Fig. 4.1; Supporting Information Table S4.1). Muscle or fin tissue was obtained and stored in 96% ethanol at 4 °C. DNA was extracted using the DNeasy Blood and Tissue Extraction Kit (Qiagen) following the manufacturer's protocol. Quality and concentration of the samples was assessed using a

GeneQuant™ *pro* RNA/DNA spectrophotometer (Amersham Biosciences, Freiberg, Germany), Qubit® 2.0 Fluorometer (Invitrogen) and agarose gel electrophoresis. Thirty-nine samples were selected for double-digest RAD sequencing at Beijing Genomics Institute Hong Kong Co., Limited (BGI, Hong Kong) as per Peterson *et al.* (2012).

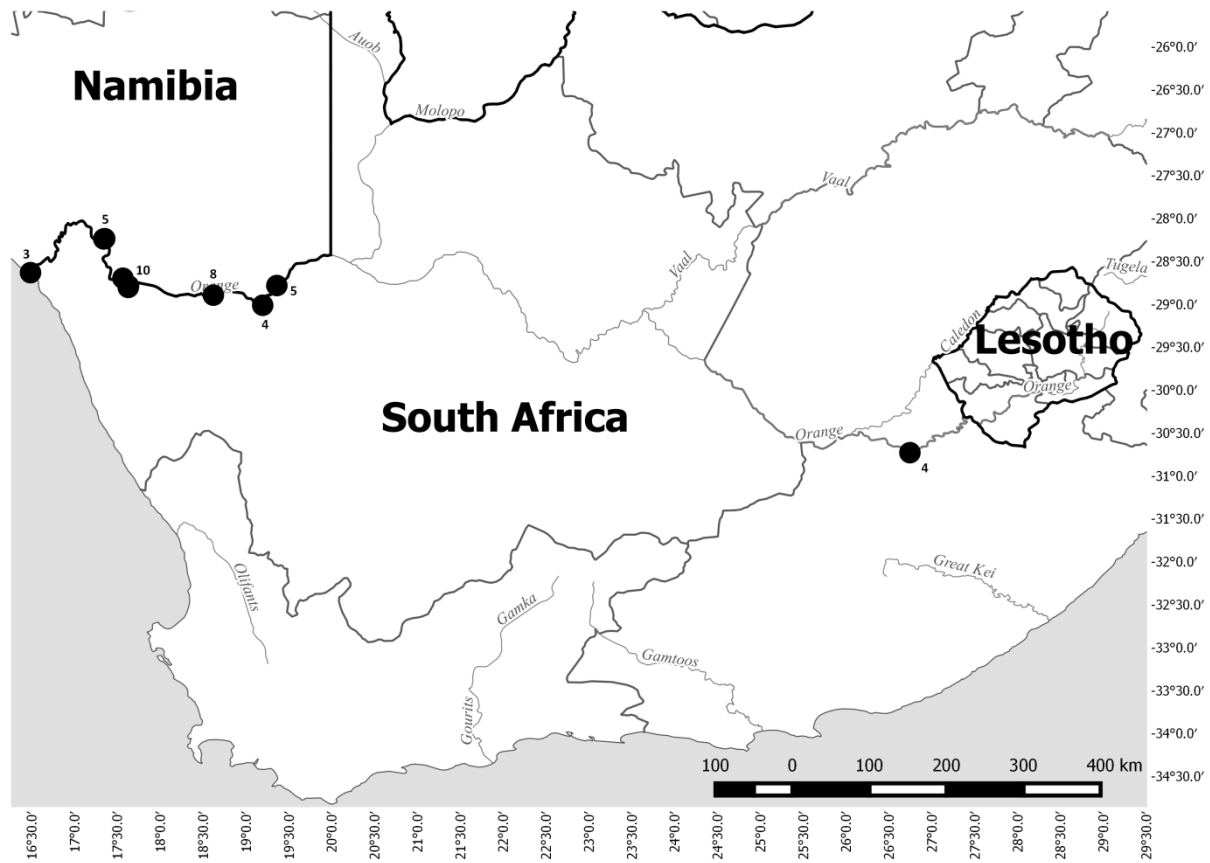


Fig. 4.1 Sampling localities of *L. aeneus* and *L. kimberleyensis* across South Africa. The number associated with each sampling site indicates how many samples were sequenced from that locality.

4.2.2 Library preparation and sequencing

The samples were sequenced in two separate libraries (Supporting Information Table S4.1) at different times. Samples were digested with restriction enzymes *Nla*III (CATG/) and *Mlu*CI (/AATT)

following the double-digest paired-end protocol of Peterson *et al.* (2012). Five samples were replicated (barcoded and sequenced twice as separate samples) as controls (Supporting Information Table S4.1). Libraries underwent an additional size-selection step of 300-500 bp prior to sequencing. Library 1 underwent 90 bp paired-end sequencing in a single lane of an Illumina HiSeq 2000 (Illumina Inc., USA), whereas library 2 was subjected to 100 bp paired-end sequencing in a single lane of an Illumina HiSeq 4000 (Illumina Inc., USA). Reads were screened for low quality (quality value ≤ 5 (E)) and demultiplexed, followed by adapter and barcode trimming.

4.2.3 Bioinformatics pipeline and quality control

Reads were cleaned in STACKS (Catchen *et al.* 2011; Catchen *et al.* 2013) using PROCESS_RADTAGS. The following additional flags were specified: *-r* (rescue RAD tags), *-c* (remove uncalled bases) and *-q* (remove low quality bases). To remove sequencing error and retain comparable results across libraries, reads were trimmed to a uniform 80 bp after plotting a SNP density spectrum to visualise potential error (Supporting Information Fig. S4.1). Read base quality was assessed using FASTQC (Andrews 2010) (Supporting Information Figs. S4.2-S4.3). Only reads that remained paired after processing were retained. Potential traces of adapter pollution were filtered from library 1 using custom regular expressions - library 2 appeared free of this type of error.

A potential bias in coding regions (see results and discussion below) was assessed by performing mappings in CLC GENOMICS WORKBENCH 7.0.4 (CLC Inc., Aarhus, Denmark) against the common carp (*Cyprinus carpio*) nuclear genome and coding DNA sequences (CDS) (GCF_000951615.1), and by assessing the transition/transversion ratio of SNPs reported by the pipeline. Mappings were performed with default parameters except for the length and identity of matches which were both raised to 90% to enforce highly similar matches. Non-specific matches were ignored. The reads

chosen for the mappings contained at least one SNP as identified in STACKS using parameters of $-m$ 5 $-M$ 2 $-n$ 1 prior to any filtering or optimisation methods.

I next assessed how many of the reads resulting from the SNP identification pipeline prior to filters were mapped in-frame starting with an overlap between the start site of the gene coding sequence and the *Nla*III restriction site. This was achieved by trimming the first base of the reads followed by trimming down to a uniform length of 78 bp per read. Likewise, the CDS was also trimmed to 78 bp beginning with the start codon. Trimmed SNP reads were then mapped to the trimmed CDS enforcing a match length of 100%, and identity of 90% (with non-specific matches ignored). The trimmed SNP reads were mapped to the untrimmed CDS with the same mapping parameters.

4.2.4 SNP discovery

Due to issues with adapter pollution and overlapping fragments primarily observed within library 1, I opted to use only Read 1 fragments (*Nla*III cut-site) for SNP identification. The wrapper program `DENOVO_MAP.PL` in STACKS was used to identify SNPs. Optimal parameters for the dataset were identified using the method of Paris *et al.* (2017). Briefly, the depth of coverage across all samples was assessed (Fig. 4.2) to determine the optimal minimum coverage parameter ($-m$). The number of SNPs, assembled loci and polymorphic loci were then compared for each sample and across all samples using the 80% sample representation rule presented in Paris *et al.* (2017) as the minimum coverage parameter ($-m$) and distance allowed between stacks ($-M$) were varied from defaults of $-m$ 5 $-M$ 2 (Supporting Information Figs. S4-S5). The parameter distance allowed between stacks to build catalog loci ($-n$) was assessed by comparing the number of polymorphic loci observed for $-n = -M - 1$, $-n = -M$, and $-n = -M + 1$ (Paris *et al.* 2017) (Supporting Information Table S4.2). The optimal

parameters for this dataset were therefore found to be $-m\ 8\ -M\ 2\ -n\ 1$. The parameter `--max_locus_stacks` was set to 7 to allow adequate binning for a hexaploid organism (J. Catchen personal communication). Additional flags `-t` and `-d` were specified to remove or break up highly repetitive tags and to resolve over-merged tags.

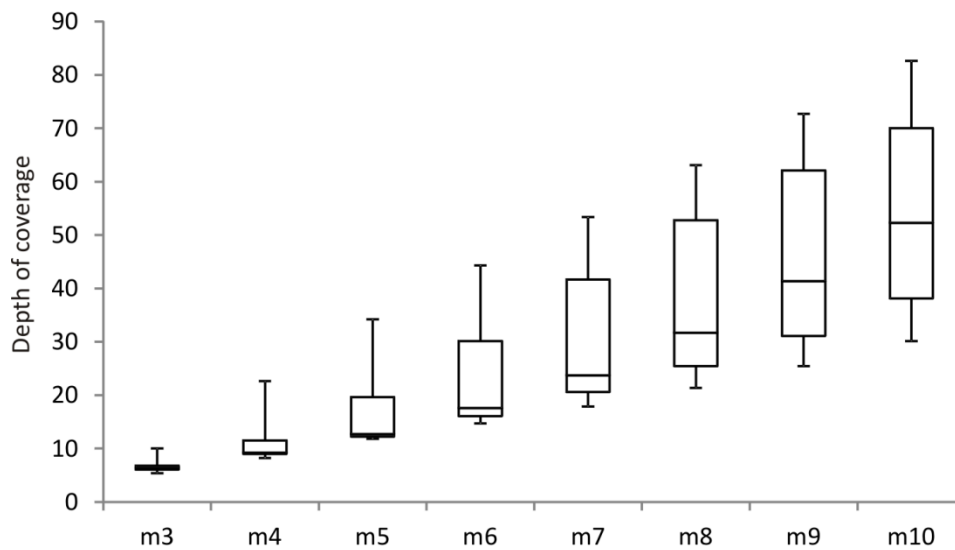


Fig. 4.2 Mean depth of coverage across all samples (including replicates) as the minimum read depth ($-m$) was varied as per Paris *et al.* (2017). Whiskers indicate the maximum and minimum values. The default parameter set was $-m\ 5\ -M\ 2\ -n\ 1$. The optimal value of $-m$ yielding mean depth of coverage $>25\times$ (Paris *et al.* 2017) was found to be $-m = 8$.

Mitochondrial reads were identified by using nucleotide BLAST (Altschul *et al.* 1997) to match ($E \leq 1 \times 10^{-20}$) against the available mitogenomes for five related genera in the same way as specified in Chapters 2 & 3. Potentially paralogous loci were identified by filtering loci with an excess of haplotypes (>2) within a single individual and by filtering loci which were genotyped as heterozygotes in all samples. Loci were further filtered to include only SNPs called within at least 50% of the samples in both populations (populations specified as the species to which the sample

was morphologically assigned). SNPs with a minor allele frequency less than 0.067 were removed to filter out singleton SNPs which may confound tests of genetic structure and differentiation (Chapter 2; Roesti *et al.* 2012; Rodríguez-Ezpeleta *et al.* 2016). One SNP per locus was selected based on maximising the count of the minor allele. Where multiple SNPs had the same minor allele count at a locus, preference was given to the first SNP represented across the most samples. However, priority was given to loci demonstrating private alleles for one of the populations. Loci passing these filters were whitelisted and rerun through POPULATIONS, resulting in 577 SNPs.

4.2.5 Mitogenome read mapping

In Chapter 3 I employed two novel methods of mitogenome reconstruction using the ddRAD reads from the 39 Orange-Vaal yellowfish, as well as 23 KwaZulu-Natal yellowfish (*L. natalensis*). SNPs were identified using these two different methods with haplotype networks, a constant coalescent chronogram and a phylogram constructed from these data used to determine the relationship between samples. Detailed methods are available in Chapter 3. Because these results are a useful independent analysis for potential introgression, particularly when compared to nuclear results obtained using the same ddRAD-seq methodology, I report on the major findings from this section even though detailed methods and results were covered in a previous chapter.

4.2.6 Population genetic parameters, structure, and assessment of nuclear hybridisation

Genetic structure was assessed in STRUCTURE 2.3.4 (Pritchard *et al.* 2000) using 100,000 burn-in chains and 500,000 MCMC chains with 20 iterations at $K = 1-4$. The result files were processed using STRUCTURE HARVESTER (Earl & VonHoldt 2011) and CLUMPP 1.1.2 (Jakobsson & Rosenberg 2007).

The optimal value for K was determined by the Evanno method (Evanno *et al.* 2005). The data were visualised spatially with a Factorial Correspondence Analysis (FCA) in GENETIX 4.05.2 (Belkhir *et al.* 2004).

Several analyses indicated that one of the samples (IRB-083A) had been misidentified when it was morphologically assigned as *L. aeneus*, as it clustered closely with samples of *L. kimberleyensis* (see discussion below). I therefore removed this sample when testing differentiation and calculating summary statistics between species. Genetic differentiation between populations was estimated by calculating pairwise F_{ST} between the two species in ARLEQUIN 3.5.1.2 (Excoffier & Lischer 2010). Summary statistics were calculated in POPULATIONS. The effect of loci potentially under selection between the two species was assessed using BAYESCAN 2.1 (Foll & Gaggiotti 2008) to identify F_{ST} outlier SNPs. A False Discovery Rate (FDR) of 0.05 and default parameters was used. The results of running STRUCTURE with and without these F_{ST} outlier SNPs was compared. Two additional sets of SNPs were identified with the same parameters and filters as above except that SNPs were required to be called in 50% of one of the two populations, or in 50% of all samples combined as a single population. These two datasets allowed me to test whether more relaxed SNP selection criteria reflected the same results, particularly for loci under selection. A third additional dataset was produced using the same parameters and filters as above but restricted to just samples from library 2 and requiring SNPs to be represented across 80% of samples from both populations to test whether between-library effects were responsible for the results observed.

Finally, the degree of hybridisation of samples was assessed using GENECLASS2 2.0.h (Piry *et al.* 2004) and MIGRATE-N 3.6.11 (Beerli 2009). GENECLASS2 was used to assign samples to one of the two species and to detect first generation migrants which are equivalent to first generation hybrids in this scenario. Default parameters were used with the Bayesian methods of Rannala & Mountain (1997). Two separate runs were conducted to test the effect of clustering the potentially

misidentified sample in each species group. MIGRATE-N was used to calculate the rate of migration between species after removing the potentially misidentified sample. The program was run with default parameters to produce a preliminary overview of the data. Default parameters were selected, with priors of $\theta = 0 - 0.100000$ and $M = 0 - 1000$.

4.3 Results

4.3.1 Library preparation, sequencing and bioinformatics

Sequencing of the first library resulted in 83.5 million reads (average 7,587,820 reads) whereas the second library gave 335.8 million reads (average 10,494,963 reads) (Table 3.1). Post-filtering for adapter pollution and quality I retained 79,133,404 reads for library 1, and 335,325,196 for library 2. GC-content was similar to that observed for *L. natalensis* (Chapter 2), and quality of both Orange-Vaal libraries was marginally improved over that for *L. natalensis*. Differences in GC-content are expected due to differing representation of repetitive regions between the runs, and may not indicate biological differences between the libraries.

The SNP density spectrum (Supporting Information Fig. S4.1) revealed a bias in SNPs at every third position from position 7 onwards for the forward read only. Although more reads mapped to the entire nuclear genome (296,578) than to coding sequences alone (CDS: 38,753), the proportion of each covered by mappings was 0.01 and 0.04 respectively. The transition/transversion ratio prior to filtering and using default parameters was found to be 2.05. Furthermore, I observed 1,092 CDS read mappings where the coding start site matched the *Nla*III restriction site, out of 27,807 total CDS read mappings.

4.3.2 SNP discovery

The SNP identification process and filtering steps used are summarised in Table 4.1. Using only Read 1 fragments resulted in the identification of 230,241 loci and 146,451 SNPs after optimisation of parameters in STACKS. Further filtering for mitochondrial reads (93), potentially paralogous loci with either more than two haplotypes within a single individual (16,536) or SNPs genotyped as heterozygotes across all samples (50,243), loci occurring in less than 50% of both populations (225,635), and SNPs with a minor allele frequency less than 0.067 (12,490) resulted in 621 SNPs. Selecting a single SNP per locus based on allele frequencies reduced this to the final dataset of 577 SNPs.

Table 4.1 SNP identification process summary and filtering steps to obtain the final SNP dataset.

Total raw reads	419,304,844
Processed single reads	179,399,987
Loci assembled	230,241
SNPs identified	146,451
Less filters:	
Mitochondrial loci	93
Potentially paralogous loci	16,536
Loci represented across <50% individuals	225,635
SNPs with MAF < 0.067	12,490
SNPs with $H_o > 0.99$	50,243
SNPs passing all filters	621
One SNP per locus based on number of minor alleles	577

MAF = minor allele frequency; H_o = Observed heterozygosity

4.3.3 Mitogenome read mapping

Haplotype networks produced using both mitochondrial SNP identification approaches revealed substantial differences between the two species, albeit with unidirectional gene flow between them

as shown by the haplotypes of samples morphologically identified as one species (almost always *L. aeneus*) clustering within the other species' haplogroup. This was shown even more clearly in the constant coalescent chronogram and phylogram in which five samples that had been morphologically assigned to *L. aeneus* consistently grouped in the *L. kimberleyensis* clade, indicating potential mitochondrial introgression between these species. Three of these five samples accounted for all of the representatives of *L. aeneus* from the upper Orange River sampling locality at Aliwal North. These results are summarised in Fig. 4.3B.

4.3.4 Population genetic parameters, structure and assessment of nuclear hybridisation

Genetic structure analysis using the Evanno method determined an optimal value of $K = 3$. However, there was no visible structure at this level. I therefore conclude that the split at $K = 2$ which corresponded with the two species in this study is most likely (Fig. 4.3C). However, sample 083A, which had been morphologically identified as *L. aeneus*, was assigned to *L. kimberleyensis*. Additionally, several samples of both species exhibited low but substantial probability of assignment (q) to a different species than based on morphological identification. The FCA plot (Fig. 4.4) showed similar patterns, where two primary clusters are visible which appear to delimit the different species. The same sample that had appeared misidentified in the STRUCTURE plot was here shown to group with *L. kimberleyensis*. However, a third cluster was also evident, which splits most of library 1 apart from the samples in library 2. This indicates that some of the signal observed on both the FCA and STRUCTURE plot may be due to the similarity within libraries.

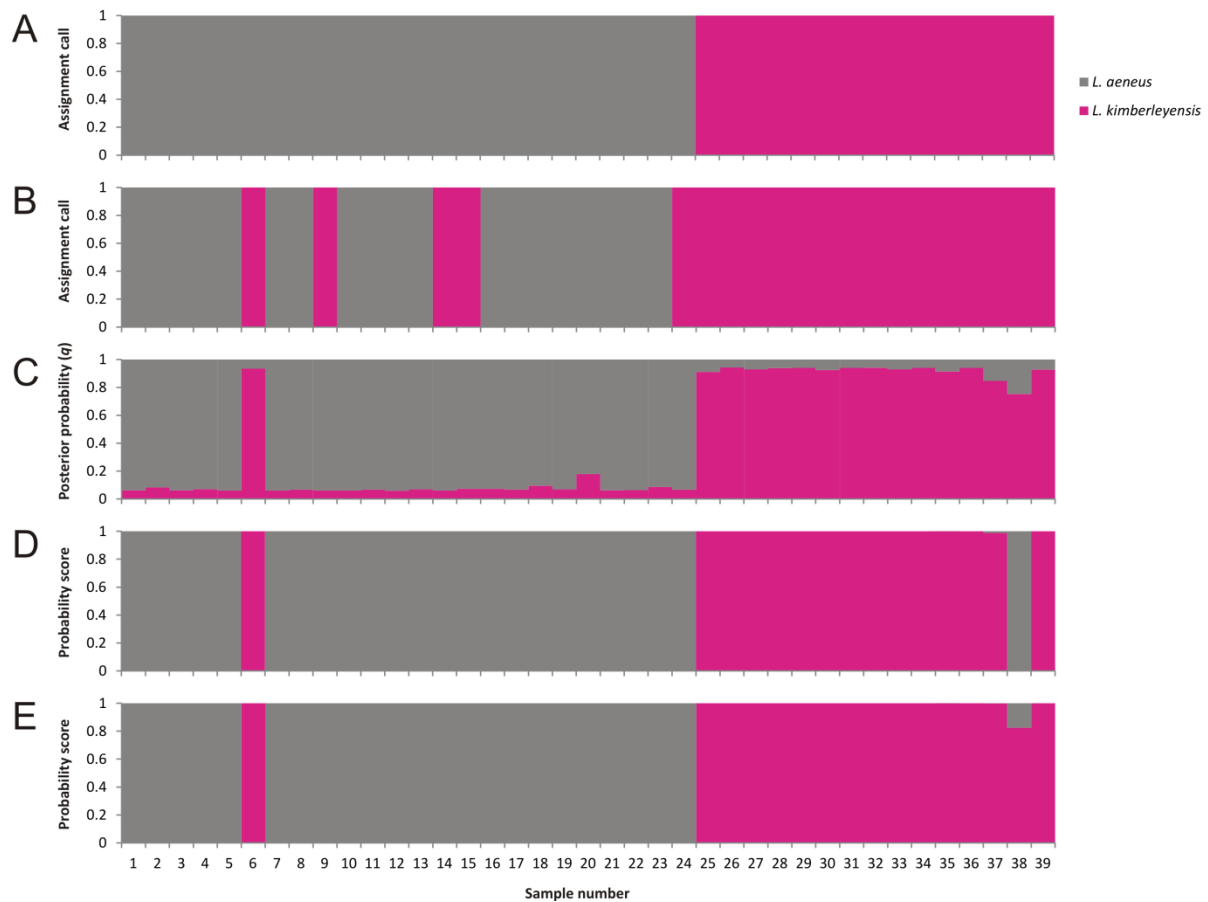


Fig. 4.3 Results of various methods of species assignment across the 39 samples in this study. A: Morphological assignment given in the field when the sample was collected. B: Mitogenome assignment based on the results of the mapping approach presented in Chapter 3. C: STRUCTURE plot showing $K = 2$ for all 39 samples across the distribution using the final dataset of 577 SNPs. Each individual is probabilistically assigned (probability of assignment q) to one of the two species. This representation was produced using CLUMPP over 20 replicates. D: GENECLASS2 probability of assignment of each sample using 577 SNPs and using the morphological assignment as the default population assignment. E: GENECLASS2 probability of assignment using the nuclear genetic STRUCTURE assignment as the default population assignment.

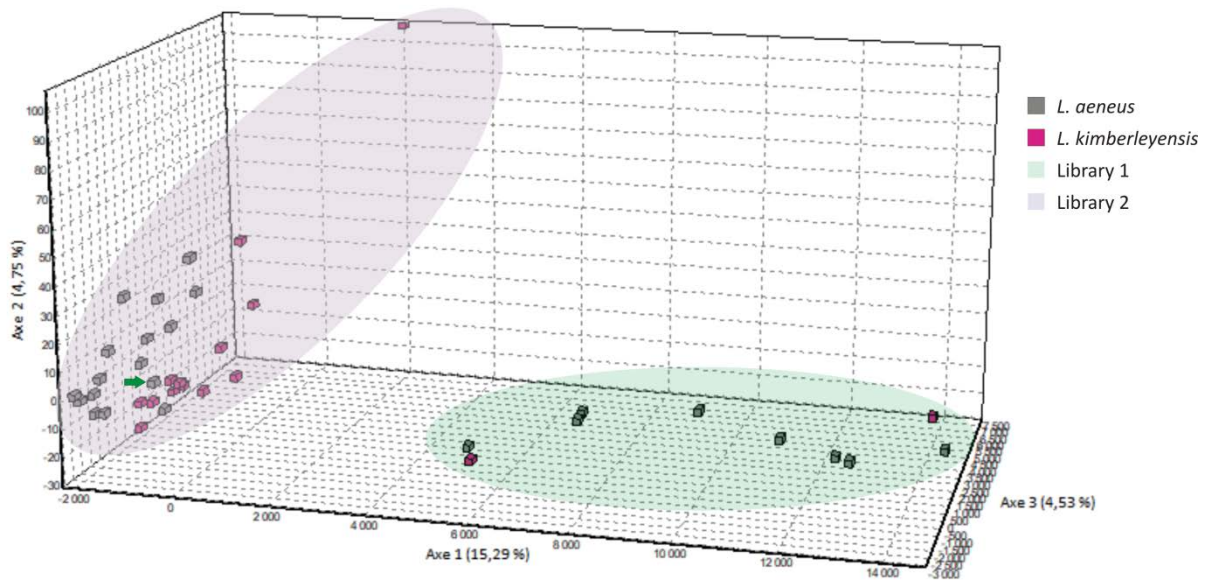


Fig. 4.4 Factorial Correspondence Analysis indicating the clustering of samples into species and libraries based on the first three components, accounting for 24.57% of the variance between samples. The two sequencing libraries are indicated. One sample from library 1 with the highest sequencing depth clustered with the samples of library 2 (green arrow).

A pairwise F_{ST} statistic of 0.016 ($p = 0.000$) was calculated between the two species, after removing the misidentified sample. Summary statistics (Table 4.2) indicated very few private alleles for each species (27 and 17). Estimates of Wright's inbreeding coefficient (F_{IS}), nucleotide diversity (π) and heterozygosity were all elevated in these species compared to the values obtained for the KwaZulu-Natal yellowfish (Chapter 2). All three of these statistics were found to be slightly greater in *L. kimberleyensis* over *L. aeneus*. However all summary statistics were highly similar between the two species.

Table 4.2 Summary statistics (from POPULATIONS) for variant positions (top) and all positions (bottom) after the potentially misidentified sample was removed from the analysis, reducing SNPs from 577 to 497.

	Pop	N	Pvt	Sites	% Poly Sites	P	H_o	H_e	F_{is}	π
Variant sites	<i>Aen</i>	15.5	27	497	96.6	0.7500	0.486	0.325	-0.318	0.336
	<i>Kim</i>	11.0	17	497	94.6	0.7440	0.505	0.325	-0.336	0.341
All sites	<i>Aen</i>	17.1	27	39,760	1.21	0.9969	0.0061	0.0041	-0.004	0.0042
	<i>Kim</i>	12.2	17	39,760	1.18	0.9968	0.0063	0.0041	-0.004	0.0043

Pop = populations; N = average samples genotyped at each locus; Pvt = number private alleles; Sites = number of sites ; % Poly Loci = percentage of polymorphic sites; P = average major allele frequency; H_o = average observed heterozygosity; H_e = average expected heterozygosity; F_{is} = average Wright's inbreeding coefficient; π = mean nucleotide diversity; *Aen* = *L. aeneus*; *Kim* = *L. kimberleyensis*.

Although 32 loci were identified as highly differentiated outliers, only four were retained in the final dataset and were hence removed following outlier testing. Removal of just these four loci resulted in a loss of genetic structure, as shown by the comparison of STRUCTURE plots with and without these loci (Supporting Information Fig. S4.6A). Nucleotide BLAST analysis of these four loci revealed two of them are found in the same genic region which forms a monomer that is repeated 1-35 times in one or more genes across Cyprinidae (for details see Supporting Information Table S4.3). This region may therefore be important for explaining the split observed between the two species. Investigating this further with additional SNP datasets, I found that relaxing the criteria for SNP selection allowed me to identify structure even after removing F_{ST} outlier loci (Supporting Information Fig. S4.6B-E). However, the number of STRUCTURE replicate runs showing this pattern declined after F_{ST} outlier loci were removed. A similar pattern to the primary dataset of a loss of genetic differentiation was observed when the analysis was limited to just samples from library 2 (Supporting Information Fig. S4.6F-G). I therefore opted to retain the four highly differentiated outlier loci in all analyses unless specified otherwise.

Two samples in the dataset showed signal of hybridisation or introgression with GENECLASS2 - these are samples 083A and Lk007 (Fig. 4.3D). Whereas sample 083A displayed a strong affinity with

the alternate species, LkO07 showed a much reduced signal of association. This may indicate that LkO07 is a recent hybrid in contrast to 083A potentially being a misidentified sample, or it may reflect the difference in number of loci used to assign each sample (291 vs. 507). I reran this analysis after switching sample 083A to the species to which it is genetically assigned (*L. kimberleyensis*) in contrast to its morphology, and found that this sample was no longer identified as a hybrid (Fig. 4.3E). MIGRATE-N analysis identified substantial migration (mean posterior migration of 283.3/436.4 with a maximum prior of 1000) occurring between the species even after the potentially-misidentified 083A was removed (Table 4.3; Supporting Information Fig. S4.7).

Table 4.3 Results of the MIGRATE-N assessment of migration between species after removing the misidentified sample 083A. Each column indicates the percentile (2.5%, 25.0%, 75.0% or 97.5%), mode, mean or median of the posterior distribution of the parameters θ and M .

Parameter	2.5%	25.0%	Mode	75.0%	97.5%	Median	Mean
θ_{aen}	0.00473	0.00580	0.00657	0.00733	0.00840	0.00663	0.00653
θ_{kim}	0.00267	0.00387	0.00470	0.00547	0.00667	0.00477	0.00472
$M_{kim \rightarrow aen}$	278.0	292.0	302.3	312.0	328.7	300.3	283.3
$M_{aen \rightarrow kim}$	374.7	388.0	400.3	412.7	491.3	417.0	436.4

θ = mutation-scaled effective population size, $M_{A \rightarrow B}$ = mutation-scaled immigration rate from population A into population B, *aen* = *L. aeneus*, *kim* = *L. kimberleyensis*.

4.4 Discussion

4.4.1 Library preparation, sequencing and bioinformatics

The two different libraries used in this study possessed subtly different characteristics. Although the average number of reads per sample was similar between libraries, samples from library 1 generally had moderate genome coverage with low site coverage such that a large proportion of loci were dropped or identified as absent from these samples, likely due to minimum coverage thresholds. This indicates the difference in quality between these two libraries. Although I have tried to use

parameters and filters to minimise the differences between these two libraries, I believe that the differences in depth are great enough to reflect as artefactual genetic structure, as shown by the FCA plot and some STRUCTURE runs. This has not permeated across all of my analyses, but some of the puzzling results obtained appear to have been caused by differences between libraries. This may have partly been caused by differences in experimental procedure, such as sequencing depth of the two libraries or size selection step range, which is notoriously variable even within replicates of a single sample (Mastretta-Yanes *et al.* 2015). Other researchers have reported similar library-based genetic differentiation artefacts between libraries with identical experimental procedures, even using advanced size selection equipment such as the Pippin Prep (Puritz *et al.* 2015). I would therefore conclude that the best approach to reduce the effect of between-library bias would be to sequence all samples for a particular study in a single library at the same time. Failing that, researchers must choose to either attempt to optimise parameters and filters to reduce this bias as I have here, or to restrict their analysis to only one of their libraries. I briefly explored this alternative (see below), but found similar results to my optimised approach.

I speculate that the observation of increased SNP density at every third position of the total loci correlates with ‘wobble’ mutations in third-codon in-frame positions as a result of the ‘ATG’ start site within the *NlaIII* restriction enzyme. I posit that the choice of restriction enzyme has therefore resulted in a slight but noticeable bias in targeting in-frame coding regions as a result of the start site. This is supported by the demonstration of increased proportions of mappings to the CDS compared to the entire nuclear genome and transition/transversion ratio both indicating a bias in coding regions, as transitions are more likely to occur in exonic regions (Park *et al.* 2010). Additionally, I showed that 3.9% of all CDS read mappings occur at the start site of the gene, which likely contributes towards a bias of in-frame coding reads. The most likely explanation for this would be the choice of restriction enzyme. This is a valuable result of this study, as it implies that enzyme

selection is paramount to study design and that a study can be biased towards coding regions - which are often of more interest to researchers due to the constraints upon these regions (Cooper & Brown 2008) - merely by selecting the appropriate restriction enzymes. Although other studies have demonstrated particular enzymes to produce a bias in genic regions through enriching for GC-rich regions (Amores *et al.* 2011; Arnold *et al.* 2013; DaCosta & Sorenson 2014; Gonen *et al.* 2015), this study is the first to my knowledge to show such a dramatic bias caused by the presence of 'ATG' within a restriction enzyme.

4.4.2 SNP discovery

A number of stringent filtering steps were used to produce a high-quality SNP dataset in Orange-Vaal yellowfishes. I found a lower proportion of mitochondrial loci in these species than in *L. natalensis* (Chapter 2). A large number of total SNPs were found to be heterozygous across all samples, which is indicative of paralogous loci where these are fixed for alternate states and have been merged in STACKS into a single locus. However, the major loss of loci was for those which occurred in less than half of the samples for both populations. This would suggest that many of the loci recovered were only found across a few samples, or within a single species. The former could have been caused by the choice of restriction enzyme, which cut so frequently that it is unlikely I would have completely sampled all loci across all samples, or possibly by further differences between the two libraries. Although I have demonstrated the merits of using the *Nla*III enzyme above, this would suggest that replacing *Mlu*CI with a rare-cutting enzyme may yield superior results as this would produce fewer fragments at a higher depth, presumably raising the representation across samples.

I previously showed that much of the missing data incorporated in my study with *L. natalensis* was due to locus dropout across entire populations (Chapter 2). This is undoubtedly the case here at

the species level, and it may partially explain why this particular set of parameters (requiring more than half of the samples in both species to possess the locus) resulted in comparatively few loci. However, due to the lack of structure observed within species, it is unlikely that this explains much of the missing data within species. The differences observed in sample read depth and coverage between libraries is likely responsible for much of the missing data seen here. I would therefore anticipate that limiting the analysis to just samples from library 2 would result in a lot less missing data and a corresponding increase in the number of suitable SNPs for this analysis. Because I was able to demonstrate that the current dataset was able to answer my research question I did not continue with this approach.

4.4.3 Mitogenome read mapping

The mitogenome approach revealed clear differences between the two species, with a proposed divergence approximately 1.5 mya. Within each of these lineages there is no further genetic structuring, which is expected given that these fish are migratory and occur throughout the river system.

Despite this deep split with shallow, recent divergence, I observed five samples which were morphologically identified as *L. aeneus* grouping in the chronogram, phylogram and haplotype networks with *L. kimberleyensis*. This is strongly suggestive of mitochondrial introgression, however it may not be applicable to the nuclear genome as introgression is often more likely to occur in the mitochondrion (Ballard & Whitlock 2004). Three of these samples represented all of the samples of *L. aeneus* from the upper Orange River. Therefore even though I found a lack of genetic structuring within *L. aeneus*, I observed a difference between *L. aeneus* of the upper and lower Orange River in terms of propensity to introgress. This may be expected given that previous studies have shown

genetic differences in *L. aeneus* between these regions (Bloomer & Naran 2006; Bloomer *et al.* 2007). This could be due to the presence of a known biogeographic barrier which plays a role for other fish species, Augrabies Falls, separating these two regions (Skelton 1986), and may indicate that extensive introgression has occurred in the upper Orange River. Further sampling of the region will need to be conducted to confirm this.

In the alternative “STACKS-based” SNP identification approach, the haplotype network indicated four samples again grouping incorrectly to their morphological assignment, but only three of these matched those identified using the mapping approach. The remaining three samples which were identified using only a single approach are from the low-coverage library 1 and may have clustered incorrectly due to large proportions of missing data across the relatively few SNPs used in the STACKS approach. Ancestral polymorphism may also have contributed towards this unexpected result. One of these latter three is a sample of *L. kimberleyensis* which is clustering with *L. aeneus* - the reciprocal to all other observed hybrids. I have chosen to use the haplotype assignment based on the mapping approach due to the greater amount of data and resolution identified using this approach.

This suggests that the observed introgression appears to be unidirectional, from *L. aeneus* into *L. kimberleyensis*. This is alarming as unidirectional gene flow via hybridisation into the rarer of the two species could lead to its extinction (Rhymer & Simberloff 1996). Unidirectional gene flow is indicative of reproductive isolation, either by selection against hybrids of a particular cross or by a tendency toward assortative mating (Dowling & Childs 1992), as has been observed between Carpathian and common barbels (Lajbner *et al.* 2009). Although this particular aspect warrants further investigation particularly from an ecological viewpoint, I suggest that it may be caused by partially overlapping spawning times of the Orange-Vaal yellowfishes. *Labeobarbus aeneus* spawns earlier on in the season, but occasionally completes a second spawning several months later and is conceivably reproductively active over this period, whereas *L. kimberleyensis* is still developing sexually during

the *L. aeneus* spawning season (Mulder 1973; Tómasson *et al.* 1984; O'Brien *et al.* 2013). This suggests that introgression can occur infrequently where the two species share a habitat (such as in a man-made environment like a dam where preferred habitats may not be available) and when *L. aeneus* have an extended period of reproductive activity. The abundance of the fish may be another contributing factor - *L. kimberleyensis* is rarer than *L. aeneus* and also requires several additional years to reach reproductive maturity (Mulder 1973; Skelton 2001), and both species exhibit an increasing skew in female/male sex ratio of up to 9:1 as they age (Gaigher 1976). This may suggest that there is simply more opportunity for male *L. aeneus* to fertilise *L. kimberleyensis* eggs when the two species share a spawning area and overlapping spawning times, with comparatively fewer *L. kimberleyensis* males living to this age of reproductive maturity and being prevalent at any given site.

4.4.4 Population genetic parameters, structure and assessment of hybridisation

The result of analyses investigating nuclear genetic differentiation across samples of both species from the upper and lower Orange River indicate a distinction between the two species, with no further genetic substructuring. A large proportion of the divergence between these species appears to be driven by loci under selection, as by removing a mere four F_{ST} outlier SNPs which are potentially under selection I was unable to differentiate between the two species with the current SNP dataset. This was resolved when SNP selection parameters were relaxed; however it does suggest that a disproportionate degree of genetic signal between species is being given by these loci. Due to the clear differentiation between species afforded by the mitogenome results as well as the large proportion of F_{ST} outlier SNPs which were identified as mitochondrial, this leads me to question whether the mitogenome is potentially under divergent selection as well. A system where patterns of genetic differentiation vary markedly between neutral markers and markers associated with life

history traits is found in the Atlantic cod (Hemmer-Hanson *et al.* 2012), where neutral markers split the fish into populations based on ecotype (migratory vs. stationary) but including highly differentiated outlier markers resulted in a different pattern of grouping by geographic location. A similar scenario of complex interplay between gene flow and divergence potentially driven by selection may be at work in these yellowfishes.

The number of private alleles observed between the two species for the current dataset was much lower than anticipated given the divergence at the population level for *L. natalensis* (Chapter 2). However when SNP identification constraints were relaxed many more private alleles were found (data not shown). This may suggest that the SNPs selected for the final dataset include relatively few markers which are useful for discriminating between the two species due to the method by which they were selected, i.e. by selecting only loci present in at least 50% of the samples for both populations. This may favour high-coverage loci with relatively few differences observed between species. Although the SNPs identified in this study reflected very similar summary statistics for both species, *L. kimberleyensis* was found to possess slightly greater F_{IS} , π , and heterozygosity despite its relative rarity. This may be an artefact - the values for both species are so similar that it is likely that these statistics reflect the ancestral values for both species combined. The elevated negative inbreeding values and heterozygosity indicate the presence of a large proportion of paralogous loci despite the filtering steps employed (as discussed in Chapter 2), whereby similar alternate paralogs are built together as stacks and reflect as heterozygotes across all or most individuals in which they are sampled, inflating heterozygosity. Further SNP studies in yellowfish would benefit from more extensive paralogous locus filtering - the HDPLLOT method of McKinney *et al.* (2017) was briefly examined here (Supporting Information Fig. S4.8), but was not used due to the paucity of samples and difficulty in distinguishing between clouds of loci.

Several parameter sets indicated a sizeable degree of admixture in some samples via probability of assignment in STRUCTURE, however further inspection revealed that this may rather be due to clustering of low-coverage samples (primarily from library 1) across both species into their own 'population'. This is most evident in the FCA plot which depicts a third 'population' which consists almost entirely of samples from the low-coverage library 1. This is consistent with the effect of library bias discussed above. Rerunning just the high-coverage library 2 samples in STRUCTURE indicated similar results to the primary analysis, suggesting that I can be confident that my efforts to reduce this bias are acceptable.

More specialised analysis programs GENECLASS2 and MIGRATE-N both independently suggested that hybridisation may be occurring between these species. GENECLASS2 suggested that two samples may be hybrids or misidentified (083A and LkO07), with one of these occurring in each sequencing library. The relative paucity in useable loci for sample LkO07 and the other samples in library 1 may have contributed towards the difficulty in other analysis programs to identify potential hybridisation. The current dataset appears to suggest that sample LkO07 is a fairly recent hybrid between the two species. The identification of recent hybridisation may suggest that Bloomer *et al.*'s (2007) hypothesis of recent introgression resulting in a loss of diagnostic allozyme markers between species is likely. The identification of only a single recent hybrid contrasts with my mitogenome results, which showed that five samples identified morphologically as *L. aeneus* possessed a haplotype typical of *L. kimberleyensis*. None of the remaining four samples (excluding the misidentified sample) demonstrated any nuclear introgression. This may suggest mitogenome capture during relatively ancient admixture events, potentially within the upper Orange River for *L. aeneus* in particular, accounting for three of the four samples.

Sample 083A was also identified as a potential hybrid, but this may have been caused by misassignment of the sample to *L. aeneus*. This consistent assignment to a species other than that

which it had been identified as morphologically was seen across the mitogenome analyses, nuclear STRUCTURE, nuclear FCA plot and GENECLASS2 analyses, which suggests that this sample was misidentified in the field or mixed up in the laboratory (although this is most unlikely). Although I cannot rule out the possibility that it originates from hybridization and repeated backcrossing to *L. kimberleyensis*, I find this possibility unlikely. This was confirmed when sample 083A was assigned to *L. kimberleyensis* and the analysis rerun - the sample was then reported as a standard individual of *L. kimberleyensis* with no obvious hybridisation. Conversely, sample LkO07 was more clearly identified as a hybrid sample. Additionally, an initial MIGRATE-N analysis suggested substantial migration occurring between both populations. Although this migration may be slightly biased in the direction of *L. aeneus* to *L. kimberleyensis* as suggested by the mitogenome analysis, the difference is not as substantial as the limited mitogenome results indicate. I therefore conclude that hybridisation and introgression are likely occurring between the Orange-Vaal yellowfishes across their distribution.

4.5 Conclusion: Hybridisation and introgression of Orange-Vaal yellowfishes

In this study I have shown evidence that *L. aeneus* and *L. kimberleyensis* are hybridising, likely in artificial environments such as dams. This may be due to a loss of physical isolation between the two species due to anthropogenic habitat modification. Mitogenome results suggested that several samples of *L. aeneus* possessed *L. kimberleyensis* haplotypes, potentially demonstrating unidirectional gene flow as a result of partially overlapping spawning times and locations in combination with a relative rarity of male *L. kimberleyensis* at reproductive maturity. Nuclear genome analyses indicate that although only one sample showed evidence of recent hybridisation, there were traces of substantial introgression between the two species. In contrast to mitogenome analyses, this introgression was not found to be unidirectional, although from *L. aeneus* to *L.*

kimberleyensis was slightly favoured which is itself a conservation concern. One sample was consistently assigned to a different species to which it was morphologically assigned in both mitogenome and nuclear SNP analyses. Because this sample was not identified as a potential hybrid after it was grouped in the alternate species, I conclude that it must have been misidentified, mislabelled or swapped with another sample.

I found that several markers potentially under selection may be driving a large proportion of the divergence between the species against a background of relative homogeneity. This is unexpected given the large number of morphological, behavioural and ecological differences between these two species. No further substructuring was observed within these two species, except for potential mitogenome capture of the *L. kimberleyensis* haplotype within *L. aeneus* of the upper Orange River. This may be contributing to some of the divergence observed between the upper and lower Orange *L. aeneus* previously.

The proposed “shoot carp” scenario explaining divergence between the two different species as a result of environmental and developmental effects (i.e. that the two species are different developmental morphotypes of the same species) appears unlikely here due to the presence of unique mitochondrial haplogroups separating the two species. However, this phenomenon could explain some of the putative hybrids observed in dams, as I hypothesise that *L. aeneus* exhibiting a “shoot carp” effect may appear phenotypically larger and become piscivorous. These fish exhibiting effects of “shoot carp” may therefore appear to be more similar to *L. kimberleyensis* and hence may be identified as hybrids despite not actually being a product of hybridisation. I observed no evidence of this, and find an explanation of hybridisation via breakdown of physical isolation through habitat modification as described above more likely.

Although I have identified several different lines of evidence pointing to hybridisation and introgression of these two species, I cannot rule out the possibility of shared ancestral

polymorphism between these relatively-recently diverged species (c. 1.5 mya according to this study) accounting for the patterns identified here. However I find this possibility unlikely given the strength and number of analyses suggesting hybridisation and introgression between these two species. Additionally, the consistency of results identifying Lk007 as a hybrid cast doubt on the hypothesis of shared ancestral polymorphisms as these would be randomly distributed among individuals. I suggest further investigation of this possibility using programs such as IMA (Hey & Nielsen 2007) which can model these scenarios and determine the best fit scenario to produce the observed results.

4.6 References

- Altschul SF, Madden TL, Schaffer AA, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402.
- Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics* **188**, 799-808.
- Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: <<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>>
- Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology* **22**, 3179-3190.
- Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology* **13**, 729-744.
- Barkhuizen LM (2017) *Labeobarbus kimberleyensis*. The IUCN Red List of Threatened Species 2017: e.T63292A100166441. Downloaded on **07 December 2017**. <<http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T63292A100166441.en>>
- Beerli P (2009) How to use MIGRATE or why are Markov chain Monte Carlo programs difficult to use? In: *Population Genetics for Animal Conservation* (eds. Bertorelle G, Bruford MW, Hauffe HC, Rizzoli A, Vernesi C), pp. 42-79. Cambridge University Press, Cambridge UK.
- Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2004) GENETIX 4.05, logiciel sous Windows pour la génétique des populations. *Laboratoire génome, populations, interactions, CNRS UMR 5171, Université Montpellier II, Montpellier*, 1996-2004.
- Bloomer P, Bills IR, van der Bank FH, *et al.* (2007) Multidisciplinary investigation of differences and potential hybridization between two yellowfish species *Labeobarbus kimberleyensis* and *L. aeneus* from the Orange-Vaal system. Follow-up study 2004-2007. Yellowfish Working Group.

- Bloomer P, Naran D (2006) Identification of conservation units of two yellowfish species *Labeobarbus kimberleyensis* and *L. aeneus*. In: Orange-Vaal yellowfish genetics pilot study report for the YWG and AngloGold Ashanti Ltd, pp. 1-27.
- Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1**, 171-182.
- Catchen JM, Hohenlohe P, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.
- Cooper GM, Brown CD (2008) Qualifying the relationship between sequence conservation and molecular function. *Genome Research* **18**, 201-205.
- DaCosta JM, Sorenson MD (2014) Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE* **9**, e106713.
- Dowling TE, Childs MR (1992) Impact of hybridization on a threatened trout of the southwestern United States. *Conservation Biology* **6**, 355-364.
- Earl DA, VonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetic Resources* **4**, 359-361.
- Eccles D (1986) Development of the gut in the South African cyprinid fish *Barbus aeneus* (Burchell). *African Zoology* **21**, 165-169.
- Eschmeyer WN, Fong JD (2017) Species by Family/Subfamily. Electronic version accessed **19 November 2017**.
<<http://researcharchive.calacademy.org/research/ichthyology/catalog/SpeciesByFamily.asp>>
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.
- Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.
- Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics* **180**, 977-993.
- Gaigher I (1976) The reproduction of *Barbus cf. kimberleyensis* (Pisces, Cyprinidae) in the Hardap dam, South West Africa. *African Zoology* **11**, 97-110.
- Gonen S, Bishop SC, Houston RD (2015) Exploring the utility of cross-laboratory RAD-sequencing datasets for phylogenetic analysis. *BMC Research Notes* **8**, 299.
- Hemmer-Hanson J, Nielsen EE, Therkildsen NO *et al.* (2013) A genomic island linked to ecotype divergence in Atlantic cod. *Molecular Ecology* **22**, 2653-2667.
- Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences USA* **104**, 2785-2790.
- Impson D (2017) *Labeobarbus aeneus*. The IUCN Red List of Threatened Species 2017: e.T63289A100162776. Downloaded on **10 December 2017**.
<<http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T63289A100162776.en>>
- Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.
- Kruger E, Mulder P (1973) Gut length and food habits of fish - a note. *Newsletter of the Limnological Society South Africa* **20**, 1-7.
- Lajbner Z, Šlechtová V, Šlechta V, *et al.* (2009) Rare and asymmetrical hybridization of the endemic *Barbus carpathicus* with its widespread congener *Barbus barbus*. *Journal of Fish Biology* **74**, 418-436.

- Mastretta-Yanes A, Arrigo N, Alvarez N, *et al.* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources* **15**, 28-41.
- McKinney GJ, Waples RK, Seeb LW, Seeb JE (2017) Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources* **17**, 656-669.
- Mulder P (1973) Aspects on the ecology of *Barbus kimberleyensis* and *Barbus holubi* in the Vaal River. *African Zoology* **8**, 1-14.
- Mulder P, van Vuuren N, Ferreira J, van der Bank F (1990) A preliminary biochemical comparison of two species of the genus *Barbus* from the Vaal River system. *Water S.A.* **16**, 147-150.
- Nakamura N, Kasahara S (1955) A study on the phenomenon of the Tobi-Koi or shoot carp. I. On the earliest stage at which the shoot carp appears. *Bulletin of the Japanese Society for the Science of Fish* **21**, 73-76.
- Nakamura N, Kasahara S (1956) A study on the phenomenon of the Tobi-Koi or shoot carp. II. On the effect of particle size and quantity of food. *Bulletin of the Japanese Society for the Science of Fish* **21**, 1022-1024.
- Nakamura N, Kasahara S (1961) A study on the phenomenon of the Tobi-Koi or shoot carp IV. Effects of adding a small number of larger individuals to the experimental batches of carp fry and culture density upon the occurrence of shoot carp. *Bulletin of the Japanese Society for the Science of Fish* **27**, 958-962.
- O'Brien GC, Jacobs F, Cronje L, Wepener V, Smit NJ (2013) Habitat preferences and movement of adult yellowfishes in the Vaal River, South Africa. *South African Journal of Science* **109**, 1-8.
- Oellermann LK, Skelton PH (1990) Hexaploidy in yellowfish species (*Barbus*, Pisces, Cyprinidae) from southern Africa. *Journal of Fish Biology* **37**, 105-115.
- Paris JR, Stevens JR, Catchen JM (2017) Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution* **8**, 1360-1373.
- Park S, Yu H-J, Mun J-H, Lee S-C (2010) Genome-wide discovery of DNA polymorphism in *Brassica rapa*. *Molecular Genetics and Genomics* **283**, 135-145.
- Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7**, e37135.
- Piry S, Alapetite A, Cornuet J-M, *et al.* (2004) GENECLASS2: a software for genetic assignment and first-generation migrant detection. *Journal of Heredity* **95**, 536-539.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.
- Puritz JB, Hollenbeck CM, Gold JR (2015) Fishing for selection, but only catching bias: library effects in double-digest RAD data in a non-model marine species. Poster presented at: *Plant and Animal Genomes XXIII*, 10-15 January 2015, San Diego, CA.
- Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences* **94**, 9197-9201.
- Rhymer JM, Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics* **27**, 83-109.
- Rodríguez-Ezpeleta N, Bradbury IR, Mendibil I, *et al.* (2016) Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection. *Molecular Ecology Resources* **16**, 991-1001.
- Roesti M, Salzburger W, Berner D (2012) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology* **12**, 94.

- Skelton PH (1986) Fish of the Orange-Vaal system. In: *The Ecology of River Systems* (eds. Davies BR, Walker KF), pp. 143-162. Springer: The Netherlands.
- Skelton PH (2001) *A Complete Guide to the Freshwater Fishes of Southern Africa*. Struik, Cape Town.
- Tómasson T, Cambray J, Jackson P (1984) Reproductive biology of four large riverine fishes (Cyprinidae) in a man-made lake, Orange River, South Africa. *Hydrobiologia* **112**, 179-195.
- van Vuuren N, Mulder P, Ferreira J, van der Bank FH (1989) The identification of hybrids of *Barbus aeneus* X *B. kimberleyensis* and *Labeo capensis* X *L. umbratus* in Hardap Dam, SWA/Namibia. *Madoqua* **16**, 27-34.
- Vreven EJ, Musschoot T, Snoeks J, Schliewen UK (2016) The African hexaploid Torini (Cypriniformes: Cyprinidae): review of a tumultuous history. *Zoological Journal of the Linnean Society* **177**, 231-305.
- Wohlfarth G (1977) Shoot carp. *Bamidgeh* **29**, 35-39.
- Yang L, Sado T, Hirt MV, *et al.* (2015) Phylogeny and polyploidy: resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). *Molecular Phylogenetics and Evolution* **85**, 97-116.

4.7 Supporting Information

Table S4.1 Sampling information for *L. aeneus* and *L. kimberleyensis* used in this study. Species assignment was based on morphological identification. The division of samples into libraries for sequencing (Orange-Vaal 1 / Orange-Vaal 2) is also indicated. Replicated samples are indicated by *rep* in parentheses following the sample name. The allocated sample number (#) refers to the number in STRUCTURE plots and other results. All samples were collected by I. Roger Bills (SAIAB).

Sample	#	Species	Library	Locality	Co-ordinates	Date
IRB-005A4	1	<i>L. aeneus</i>	Orange-Vaal 2	Noordoewer	28° 44' 50.0" S 17° 36' 32.0" E	21-10-2006
IRB-010K4	25	<i>L. kimberleyensis</i>	Orange-Vaal 2	Noordoewer	28° 44' 50.0" S 17° 36' 32.0" E	21-10-2006
IRB-018A3 (<i>rep</i>)	2	<i>L. aeneus</i>	Orange-Vaal 2	Noordoewer - Felix Unite Camp	28° 41' 19.0" S 17° 33' 20.0" E	20-10-2006
IRB-028K5	26	<i>L. kimberleyensis</i>	Orange-Vaal 2	Gamkab	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006
IRB-029K5 (<i>rep</i>)	27	<i>L. kimberleyensis</i>	Orange-Vaal 2	Gamkab	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006
IRB-031A5	3	<i>L. aeneus</i>	Orange-Vaal 2	Gamkab	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006
IRB-041K5	28	<i>L. kimberleyensis</i>	Orange-Vaal 2	Gamkab	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006
IRB-044A5	4	<i>L. aeneus</i>	Orange-Vaal 2	Gamkab	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006
IRB-047K5 (<i>rep</i>)	29	<i>L. kimberleyensis</i>	Orange-Vaal 2	Gamkab	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006
IRB-048A5	5	<i>L. aeneus</i>	Orange-Vaal 2	Gamkab	28° 14' 45.0" S 17° 21' 00.0" E	22-10-2006
IRB-081K1	30	<i>L. kimberleyensis</i>	Orange-Vaal 2	Houms River	28° 51' 05" S 18° 36' 52" E	18-10-2006
IRB-083A1	6	<i>L. aeneus</i>	Orange-Vaal 2	Houms River	28° 51' 05" S 18° 36' 52" E	18-10-2006
IRB-089A1	7	<i>L. aeneus</i>	Orange-Vaal 2	Houms River	28° 51' 05" S 18° 36' 52" E	18-10-2006
IRB-097K1	31	<i>L. kimberleyensis</i>	Orange-Vaal 2	Houms River	28° 51' 05" S 18° 36' 52" E	18-10-2006
IRB-098A1	8	<i>L. aeneus</i>	Orange-Vaal 2	Houms River	28° 51' 05" S 18° 36' 52" E	18-10-2006
IRB-215A1	9	<i>L. aeneus</i>	Orange-Vaal 2	Houms River	28° 51' 05" S 18° 36' 52" E	18-10-2006
IRB-225A2	10	<i>L. aeneus</i>	Orange-Vaal 2	Noordoewer	28° 45' 20.0" S 17° 36' 46.0" E	20-10-2006
IRB-257K2	32	<i>L. kimberleyensis</i>	Orange-Vaal 2	Noordoewer	28° 45' 20.0" S 17° 36' 46.0" E	20-10-2006
IRB-261K1	33	<i>L. kimberleyensis</i>	Orange-Vaal 2	Houms River	28° 51' 05" S 18° 36' 52" E	19-10-2006
IRB-266K1	34	<i>L. kimberleyensis</i>	Orange-Vaal 2	Houms River	28° 51' 05" S 18° 36' 52" E	19-10-2006
IRB-277A7	11	<i>L. aeneus</i>	Orange-Vaal 2	Gamkab rapids at camp	28° 14' 30.0" S 17° 20' 49.0" E	23-10-2006
IRB-284A7	12	<i>L. aeneus</i>	Orange-Vaal 2	Gamkab rapids at camp	28° 14' 30.0" S 17° 20' 49.0" E	23-10-2006
IRB-359K7	35	<i>L. kimberleyensis</i>	Orange-Vaal 2	Gamkab	28° 14' 30.0" S 17° 20' 49.0" E	23-10-2006
IRB-401K11	36	<i>L. kimberleyensis</i>	Orange-Vaal 2	Orange Mouth	28° 36' 02.9" S 16° 27' 23.5" E	26-10-2006
IRB-403A11	13	<i>L. aeneus</i>	Orange-Vaal 2	Orange Mouth	28° 36' 02.9" S 16° 27' 23.5" E	26-10-2006
IRB-447A11	14	<i>L. aeneus</i>	Orange-Vaal 2	Orange Mouth	28° 36' 02.9" S 16° 27' 23.5" E	26-10-2006
LkA06 (<i>rep</i>)	39	<i>L. kimberleyensis</i>	Orange-Vaal 2	Aliwal North, weir 1 km upstream of N6 road bridge	30° 40' 45" S 26° 43' 11" E	22-01-2004
LaA13	24	<i>L. aeneus</i>	Orange-Vaal 2	Aliwal North, weir 1 km upstream of N6 road bridge	30° 40' 45" S 26° 43' 11" E	22-01-2004
LaA15	15	<i>L. aeneus</i>	Orange-Vaal 1	Aliwal North, weir 1 km upstream of N6 road bridge	30° 40' 45" S 26° 43' 11" E	22-01-2004
LaA22	16	<i>L. aeneus</i>	Orange-Vaal 1	Aliwal North, weir 1 km upstream of N6 road bridge	30° 40' 45" S 26° 43' 11" E	22-01-2004
LaO02	17	<i>L. aeneus</i>	Orange-Vaal 1	Onseepkans farm	28° 44' 31" S 19° 20' 07" E	25-01-2004
LaO05 (<i>rep</i>)	23	<i>L. aeneus</i>	Orange-Vaal 1	Onseepkans farm	28° 44' 31" S 19° 20' 07" E	25-01-2004
LaO08	18	<i>L. aeneus</i>	Orange-Vaal 1	Onseepkans farm	28° 44' 31" S 19° 20' 07" E	25-01-2004
LaO10	19	<i>L. aeneus</i>	Orange-Vaal 1	Onseepkans farm	28° 44' 31" S 19° 20' 07" E	25-01-2004
LaO11	20	<i>L. aeneus</i>	Orange-Vaal 1	Onseepkans farm	28° 44' 31" S 19° 20' 07" E	25-01-2004
LaO13	21	<i>L. aeneus</i>	Orange-Vaal 1	Pella Drift upper site	28° 57' 39" S 19° 09' 50" E	26-01-2004
LaO17	22	<i>L. aeneus</i>	Orange-Vaal 1	Pella Drift upper site	28° 57' 39" S 19° 09' 50" E	26-01-2004
LkO06	37	<i>L. kimberleyensis</i>	Orange-Vaal 1	Pella Drift upper site	28° 57' 39" S 19° 09' 50" E	28-01-2004
LkO07	38	<i>L. kimberleyensis</i>	Orange-Vaal 1	Pella Drift upper site	28° 57' 39" S 19° 09' 50" E	28-01-2004

Table S4.2 Optimisation of the STACKS parameter $-n$ using the method of Paris *et al.* (2017). Parameter sets for $-n = -M$, $-n = -M - 1$ and $-n = -M + 1$ were tested for the number of assembled loci, polymorphic loci and SNPs across all samples as well as at the population level (numbers in bold) using the script COUNT_FIXED_SNPS.PY and POPULATIONS respectively. All replicates were included, and the “80% rule” was specified for either all samples as a single population or within at least a single population when both populations were defined. Numbers in italics indicate the criteria for which the optimal dataset is chosen according to Paris *et al.* (2017).

COUNT_FIXED_SNPS.PY		<i>m8M2n1</i> ($-n = -M - 1$)	<i>m8M2n2</i> ($-n = -M$)	<i>m8M2n3</i> ($-n = -M + 1$)
Heterozygous SNPs per sample	048A5	35626	36333	36824
	047K5-A	3098	3232	3346
	098A1	2307	2384	2471
	277A7	2163	2191	2294
	LaO17	737	734	732
	010K4	23870	23423	23128
	LaO13	7171	6988	6895
	215A1	2539	2616	2727
	LaA13	10020	10116	10219
	LaO11	138	130	129
	257K2	1242	1253	1291
	LaO10	228	220	216
	LaO05-B	101	102	104
	LkO07	246	234	231
	044A5	2075	2056	2077
	018A3-B	11795	11525	11432
	LkO06	47	43	44
	447A11	7725	7705	7780
	403A11	316	333	343
	031A5	320	305	319
	261K1	501	503	516
	LaO05-A	383	377	371
	359K7	57	61	65
	LaA15	81	78	76
	028K5	5959	5748	5664
	047K5-B	496	494	503
	LaA22	52	46	45
	089A1	394	387	383
	284A7	8090	7988	7969
	LkA06-A	5174	5089	5050
	097K1	2194	2174	2159
	018A3-A	6768	6415	6239
	LaO02	98	88	82
	029K5-B	2554	2453	2442
	083A1	1763	1713	1672
	LkA06-B	2570	2528	2537
	266K1	1272	1228	1213
	041K5	1248	1193	1170
	081K1	295	298	296
	225A2	488	459	471
	401K11	1095	1052	1021
	LaO08	25	24	22
	029K5-A	6205	5950	5831
	005A4	1808	1640	1585
	Variable sites across population	161334	159909	159984
	Heterozygous SNPs in the catalog	170574	176075	181568
	SNPs across all samples and in catalog	165002	168863	173511
	Fixed SNPs only in the catalog	3668	8954	13527
POPULATIONS ($-r = 0.8$ across all samples as one population)	Number of loci	1292	1078	998
	Number of polymorphic loci	729	604	549
	Number of SNPs	1243	1013	916
POPULATIONS ($-r = 0.8$ in at least one of two populations)	Number of loci	2438	2150	2011
	Number of polymorphic loci	1409	1239	1142
	Number of SNPs	2338	2010	1847

Table S4.3 Results of a BLAST analysis of two of the outlier loci found to be potentially under selection. These SNPs were found to belong to the same gene, which is composed of multiple copies of a monomeric unit approximately 588 bp long and which appears in multiple sites with varying repeats across Cyprinidae. Due to this and the fact that a loss of genetic structure between species was observed after filtering the four outlier loci, it appears that this gene is important for differentiating between the two Orange-Vaal yellowfish species.

Species	Gene	Variant	# of repeats	Size of repeat	% Identity
<i>Sinocyclocheilus rhinoceros</i>	LOC107721013	1	2	~588 (497-496)	~82
	LOC107758209	1	2	~582 (289-288)	~83
<i>Sinocyclocheilus anshuiensis</i>	LOC107665915	1	2	~580 (72-71)	~83
<i>Sinocyclocheilus grahami</i>	flocculation protein FLO11-like (LOC107561920)	1	1	187	~84
<i>Danio rerio</i>	LOC101886224	1	35	199-588 (289-288)	72-100
		2	35	199-588	72-100
		3	35	199-588	72-100
		4	36	199-588	72-100
	LOC100330916	1	1	297	66
		2	13	294	65-66
		3	14	294	65-66
<i>Cyprinus carpio</i>	LOC109071419	1	1	552	82
	LOC109060850	1	1	~588	~82
	LOC109060848	1	2(?)	185-588	~81
	LOC109060847	1	2	~588	~81
	LOC109070844	1	2	~200-588	~81
	LOC109060849	1	2	115-588	~84
<i>Astyanax mexicanus</i>	LOC111190108	1	1	139	73
	LOC111196156	1	4	139	73

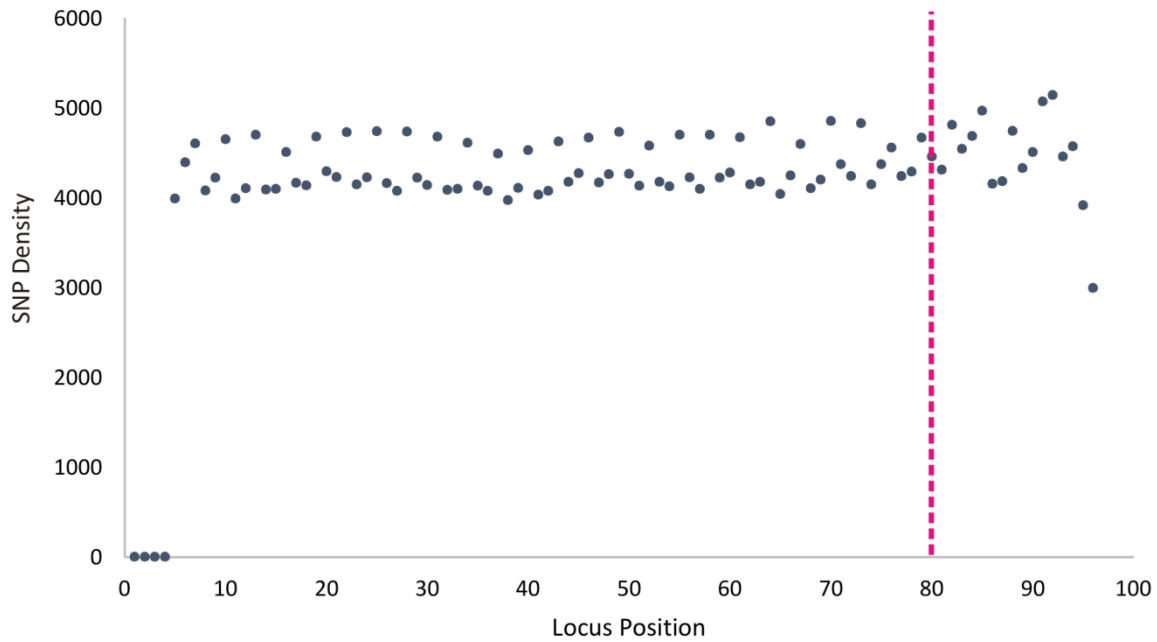


Fig. S4.1 SNP density spectrum showing the distribution of SNPs across all assembled loci (read 1 fragments only) in STACKS prior to any filtering or trimming using default parameters *-m 5 -M 2 -n 1*. The first four positions will always record an absence of SNPs as they correspond to the first restriction enzyme. An elevated number of SNPs was recorded at every third position after site 4, which I argue is due to an in-frame genic bias caused by the 'ATG' in the restriction enzyme recognition sequence (see main text). A uniform density across locus positions is indicative of low levels of sequencing error. To remove sequencing errors, I employed a size selection cut-off of 80 base pairs (magenta line).

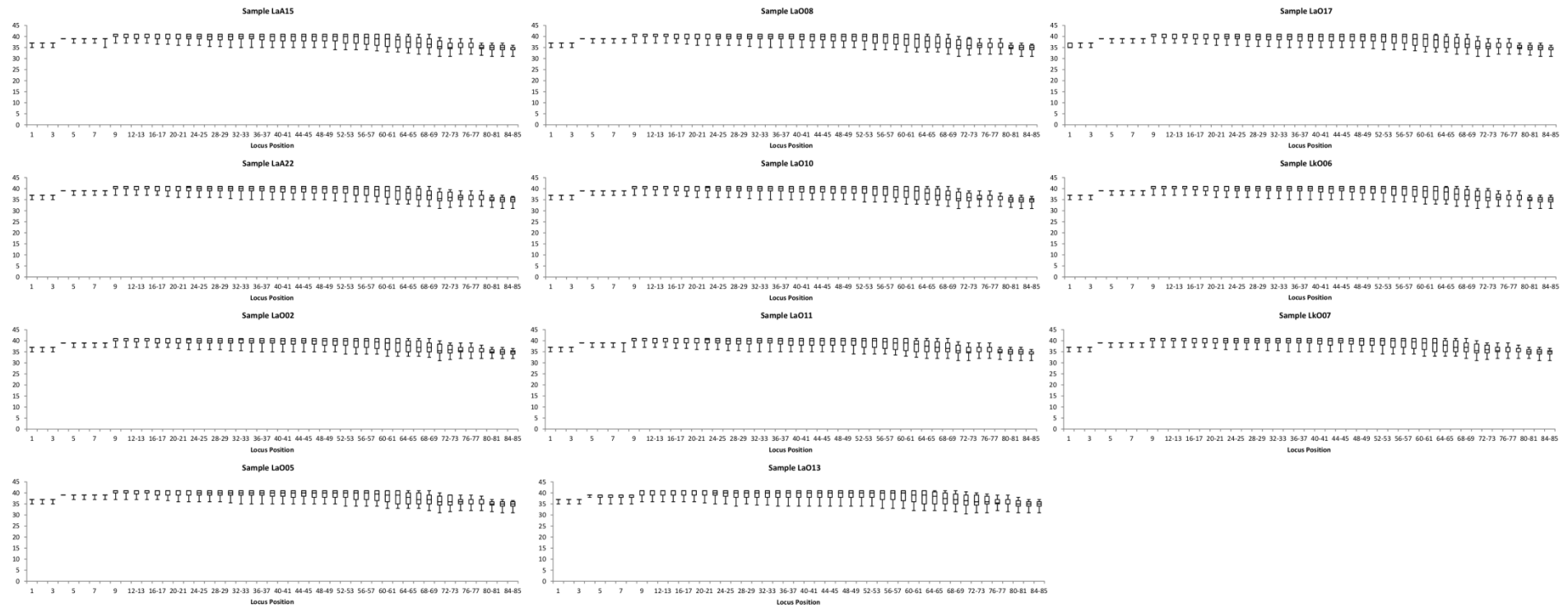


Fig. S4.2 Quality scores for each position of all reads for each sample within the first Orange-Vaal yellowfish library as per FASTQC (Andrews 2010). Whiskers here indicate 10th and 90th percentiles. The quality here was superior to that of *L. natalensis* (Chapter 2) but inferior to the second library (Fig. S4.3).

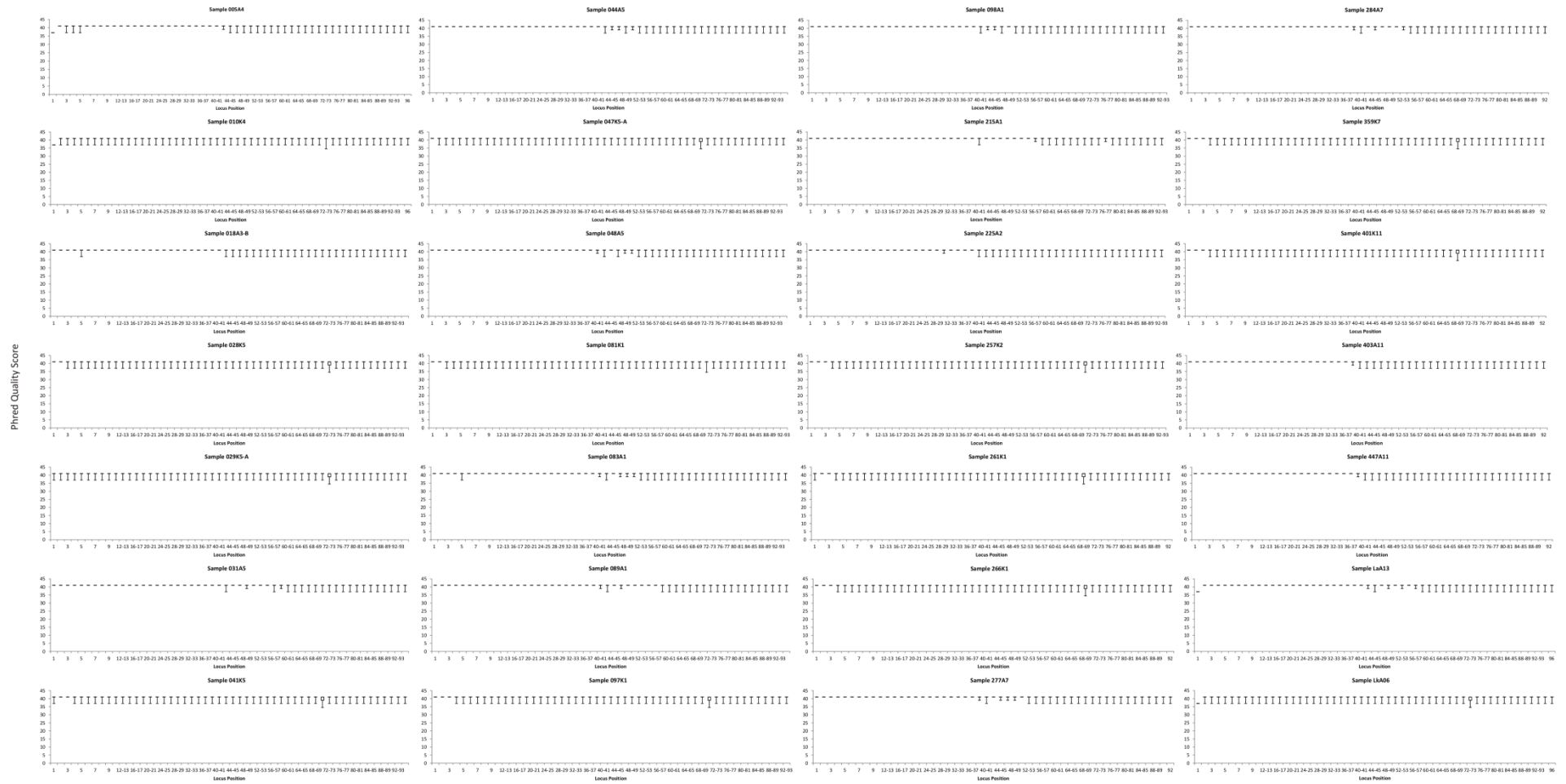


Fig. S4.3 Quality scores for each position of all reads for each sample within the second Orange-Vaal yellowfish library as per FASTQC (Andrews 2010). Whiskers indicate 10th and 90th percentiles. Quality was near-perfect for most reads across all samples, with little variance across sites.

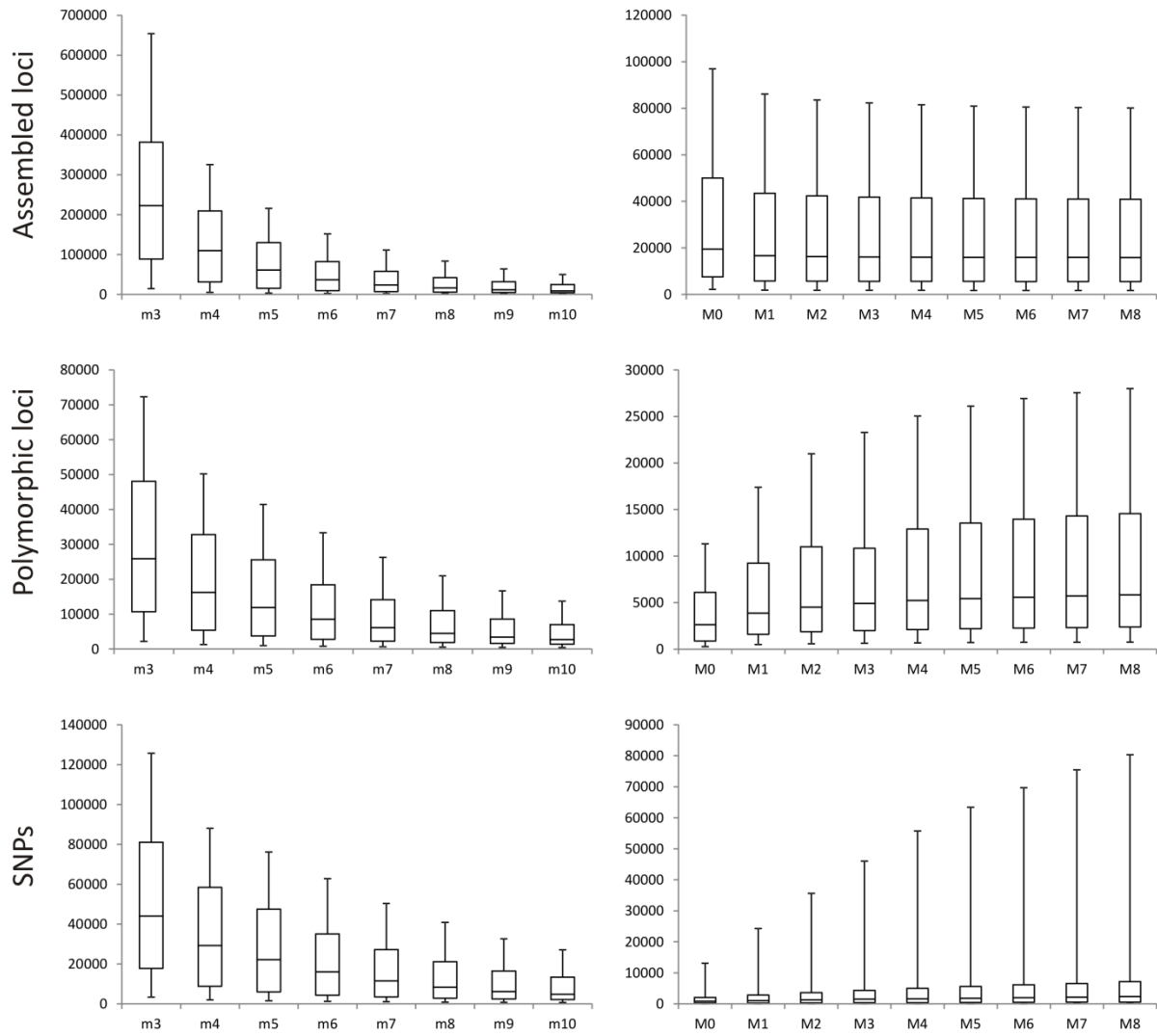


Fig. S4.4 Parameter optimisation for STACKS parameters $-M$ and $-n$ as per the method of Paris *et al.* (2017). Numbers of assembled loci, polymorphic loci and SNPs were assessed as values of $-m$ and $-M$ were varied from a default of $-m$ 5 $-M$ 2 $-n$ 1. All samples and replicates were used in this assessment.

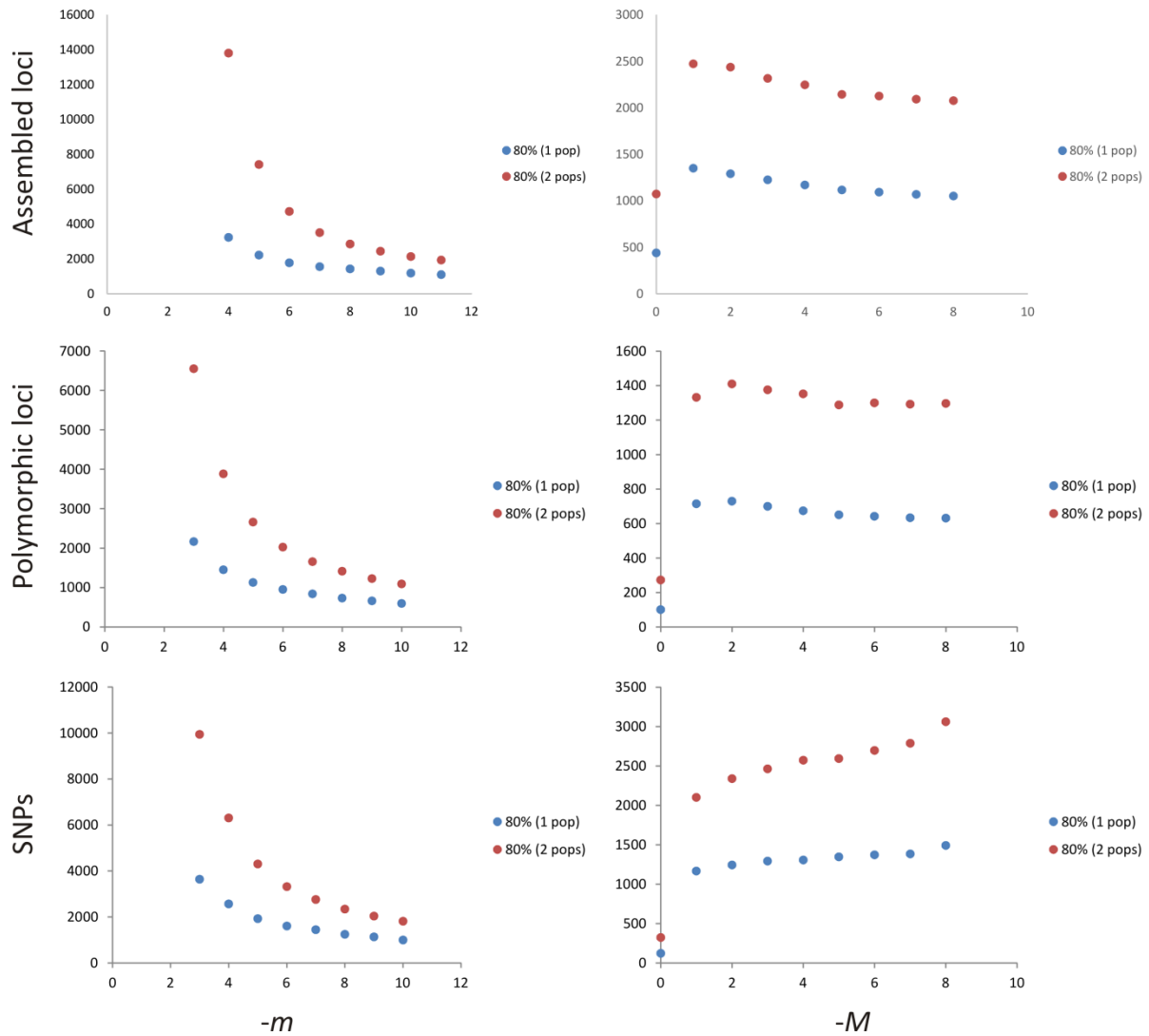


Fig. S4.5 Parameter optimisation for STACKS parameters $-M$ and $-n$ at the population level as per the method of Paris *et al.* (2017). Numbers of assembled loci, polymorphic loci and SNPs were assessed as values of $-m$ and $-M$ was varied from a default of $-m$ 5 $-M$ 2 $-n$ 1. All samples and replicates were used in this assessment. I used Paris *et al.* (2017)'s "80% rule" for both samples as a single population (1 pop) or for that cut-off within at least one of two defined populations (2 pops). The two sets reported similar results, although the latter possessed a comparative abundance of SNPs and loci due to the relaxation of constraints.

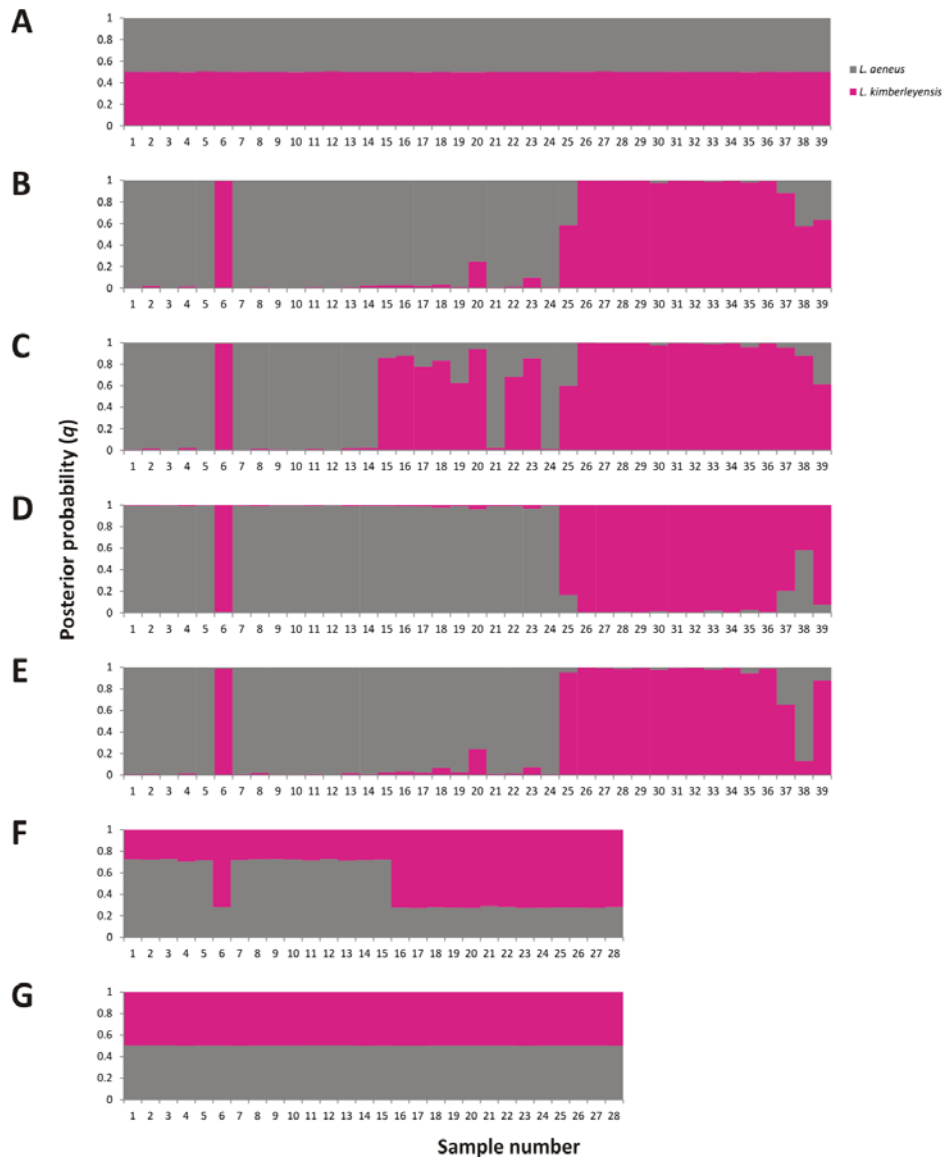


Fig. S4.6 STRUCTURE plots generated for each of the additional datasets tested in this study. Where not specified otherwise, all parameters in both STACKS and STRUCTURE are the same as for the primary analysis in the main study. Each plot was produced using CLUMPP from 10 replicates. A: primary dataset with four F_{ST} outlier loci removed, showing a complete loss of genetic differentiation between the two species. B: Additional dataset using the $-r = 0.5$ cut-off across all samples as a single population (50% of all samples have to possess a SNP for it to be called). C: Dataset from B after F_{ST} outlier loci have been removed. This plot now mistakenly groups samples from the second library together with the *L. kimberleyensis* population, presumably due to a loss of genetic differentiation and library bias (see main text). D: Second additional dataset where the $-r = 0.5$ cut-off is set for at least one of the two populations (50% of the samples within at least one of two populations are required to possess a SNP for it to be called). E: The same dataset as D but with F_{ST} outlier loci removed. Here little difference between results is observed. F: Third additional dataset where only samples from library 2 are run through STACKS with all the same parameters except $-r = 0.8$ in both populations. G: The same dataset as F but after F_{ST} outlier loci have been removed. Again a loss of genetic differentiation is observed.

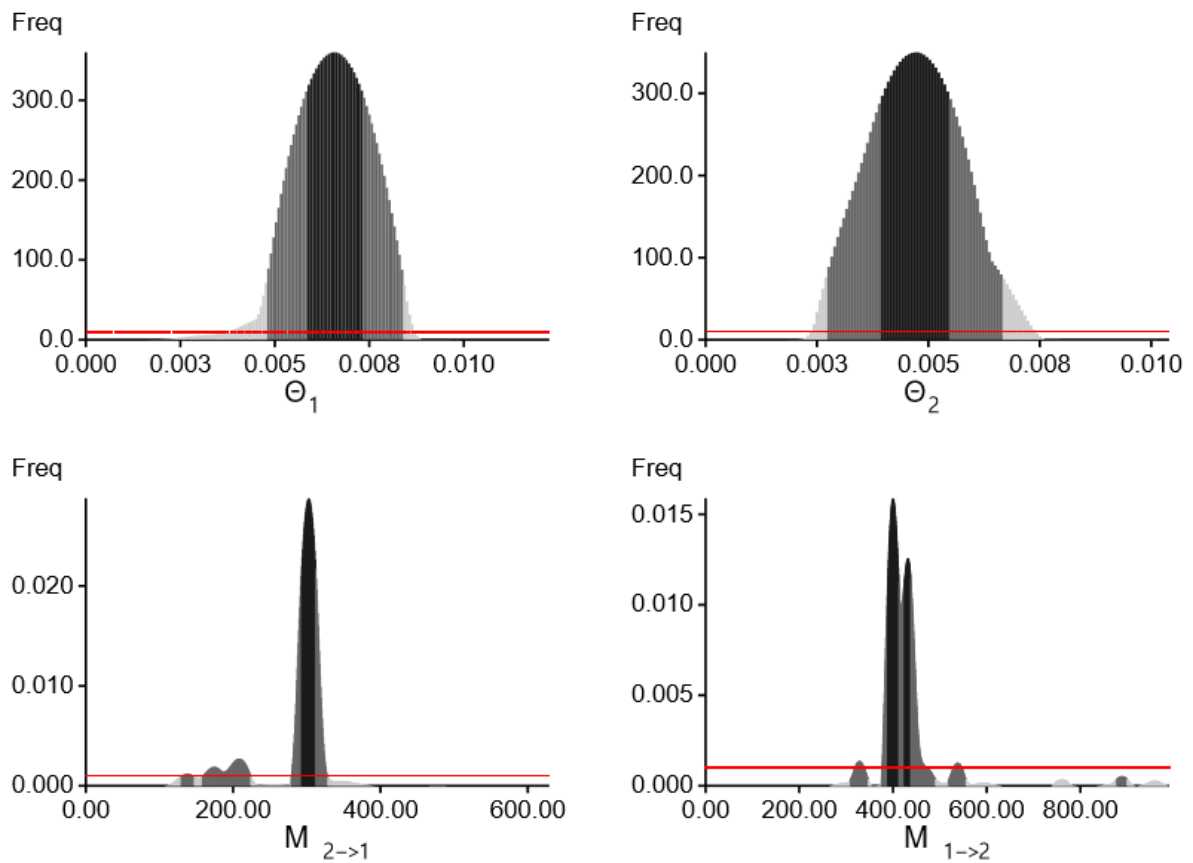


Fig. S4.7 MIGRATE-N output indicating the posterior distributions for the parameters mutation-scaled effective population size (θ) and mutation scaled immigration rate (M) between species after removing sample 083A. Population 1 refers to *L. aeneus* whereas population 2 is *L. kimberleyensis*. Associated summary statistics are indicated in Table 4.

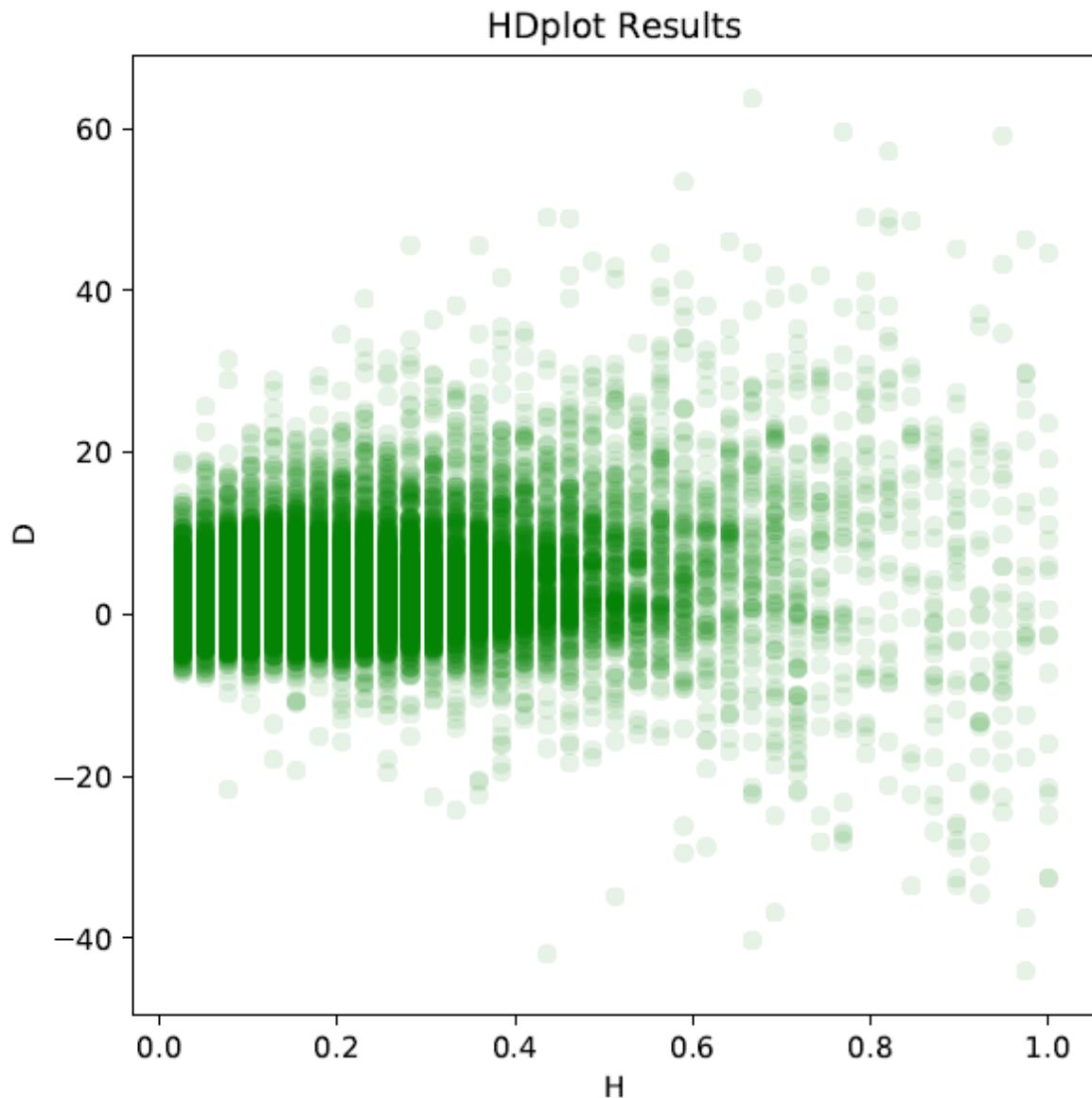


Fig. S4.8 Results of an initial screening for paralogous loci using HDPLLOT (McKinney *et al.* 2017) with the final parameter set prior to filters ($-r = 0$). The parameter H represents the proportion of heterozygotes whereas D is the deviation of alleles at a diallelic locus from a 1:1 ratio. It appears as if the clouds of duplicated loci and non-paralogous loci are directly adjacent to each other, and it therefore seems difficult to correctly isolate paralogous SNPs in this case. The relatively few samples and sequencing coverage undoubtedly contributed towards this result.

Chapter 5: Synthesis

5.1 Aims

The genus *Labeobarbus* is highly popular for recreational angling and subsistence fisheries but remains understudied. The taxonomic instability of recent years within the genus herald the difficulties in delimiting and placing species, particularly due to morphological plasticity, intraspecific variability and frequent hybridisation (Boulenger 1902; Banister 1972; Chenuil *et al.* 1999; Vreven *et al.* 2016). Two of the three species I explored in this study (*L. natalensis* and *L. aeneus*) are renowned for their wide variety of mouth morphotypes (Vreven *et al.* 2016) and *L. natalensis* shows further evidence of a ‘plastic’ morphology by historically being described as up to twelve different species (now taxonomic synonyms) due to morphological differences between its morphotypes (Crass 1964; Cambray *et al.* 2017). The aim of this study was to examine the phylogeography of three South African species of the genus using a population genomics framework. Due to complications in traditional nuclear genetic analyses caused by hexaploidy of the genus, this was achieved using SNPs identified through double-digest Restriction-site Associated DNA (ddRAD) sequencing. In Chapter 2 I sought to identify the populations of *L. natalensis* throughout its distribution to aid in conservation and management of the species, and to determine the patterns and processes acting on the species resulting in genetic structure across its distribution. Chapter 3 demonstrated a new technique I have developed whereby ddRAD sequencing data are mined for mitochondrial reads which may be used to independently assess the maternal history of a population or species. This may support (nuclear) RAD sequencing results, or unveil mitonuclear discordance which would suggest processes such as hybridisation may be at work. Finally, in Chapter 4 I assessed whether hybridisation and introgression are occurring between two sympatric species of yellowfish. I also aimed to

demonstrate that these two species are indeed genetically distinct units, as this has been questioned previously (e.g. Mulder *et al.* 1990).

5.2 Observations from this study

I was successful in answering each of my research questions. In Chapter 2 two major lineages of *L. natalensis* which I term ‘northern’ and ‘southern’ were identified and are further subdivided into five populations (Umfolozi, Tugela, Umgeni, Mbokodweni, Mkomaas). These populations are similar to but distinct from the six major haplogroups identified in previous analyses using a single mitochondrial marker (Bloomer *et al.* In prep). However, the mitogenome analysis (Chapter 3) reproduced the six major haplogroups previously identified. I therefore determined that the prediction from this question was incorrect due to a mitonuclear disparity. The difference observed between the nuclear and mitochondrial results has three potential explanations: ancestral polymorphisms at nuclear loci; the inability for this form of nuclear analysis to identify more recent divergence in the presence of substantial historical events; or gene flow occurring between these populations combined with local mitochondrial fixation between lineages. Whatever the cause, this highlights the importance of using both nuclear and mitochondrial markers when delimiting different populations or species.

I was able to mine mitochondrial read markers from ddRAD sequencing datasets by employing two different approaches (Chapter 3). This is a relatively novel method as, although other NGS datasets have been utilised to reconstruct partial or even complete mitogenomes (Li *et al.* 2014; Tang *et al.* 2014; Zhang *et al.* 2015; Crampton-Platt *et al.* 2016; Vargas *et al.* 2017), this is the first study to my knowledge to use this approach with RAD data (although several studies do at least filter for these loci - e.g. Truong *et al.* 2012; Pujolar *et al.* 2014). This represents a step forward for

research using RAD datasets as the maternal history can be assessed as an independent marker to validate nuclear signal or indicate mitonuclear discordance, which may itself herald the occurrence of processes, such as introgression, demonstrated here.

My third research question related to the genetic distinction between *L. aeneus* and *L. kimberleyensis*. I investigated this in Chapters 3 and 4 utilising mitochondrial and nuclear markers respectively. Nuclear analyses indicated that a few loci, primarily those potentially under selection, were contributing most of the signal for genetic differentiation against a background of relative homogeneity of the two species. The number of private alleles determined between species was fewer even than those found between the populations of *L. natalensis*. The morphological distinctiveness of the two species was also questioned when it was discovered that one of the samples assigned as *L. aeneus* appeared to have been misidentified. However, mitogenome results indicated a clear distinction between the two species. This may indicate that the mitogenome is also under selection. These results lead me to conclude that the two species are genetically distinct.

Hybridisation and introgression at both intra- and interspecific levels was assessed using both mitochondrial (Chapter 3) and nuclear (Chapter 2 & 4) markers, as suggested by Rhymer & Simberloff (1996). At the intraspecific level, admixture was determined to be occurring within *L. natalensis* between the northern and southern lineages at the Umgeni system. In addition, I found evidence of a translocation from the Tugela population (Mooi River) into the Umgeni system via the Mooi-Mgeni Transfer Scheme (MMTS). Although there was no evidence of admixture due to this ongoing translocation, I speculate that in the intervening years from when the samples were taken a greater extent of admixture has likely taken place. This is all at odds with my initial hypotheses, as I had not anticipated gene flow between these relatively isolated rivers.

All three species in this study may currently be experiencing interspecific hybridisation, particularly with one another - *L. natalensis* is listed as being potentially threatened by hybridisation

with introduced *L. aeneus* in the Thukela catchment (Karssing 2008), whereas *L. aeneus* and *L. kimberleyensis* are sympatric and a number of potential hybrids have been reported (Gaigher 1976; Eccles 1986; van Vuuren *et al.* 1989; Mulder *et al.* 1990). Although I did not explicitly examine the former scenario, I found no evidence of introgression between these species in my mitogenome analysis or any evidence of structure across the Tugela system in the analysis of nuclear RAD markers, which is in accordance with my hypotheses. The latter scenario was examined in great depth in Chapter 4. I found one potentially early-generation hybrid in the dataset, and further analyses indicated that a substantial proportion of gene flow (introgression) was occurring between the two species in both directions. This supports my hypotheses and the previous observations of hybridisation. Furthermore, I found some evidence that mitogenome capture of the *L. kimberleyensis* haplotype may have occurred within *L. aeneus* of the upper Orange River, which points to possible differences between *L. aeneus* of the upper and lower Orange River despite the lack of genetic structure observed within these species. The sample size for this region was limited and this aspect would benefit from future research.

5.3 RAD sequencing

Although I initially assessed the applicability of traditional nuclear markers (microsatellites, exons and introns) to phylogeographic analysis of yellowfish, I found that optimisation of these markers was difficult and complicated by the polyploidy of the genus. When polymorphic microsatellites were amplified, more genotypes than what would be expected for a diploid were observed. Additionally, introns and exons sequenced yielded additional paralogous sequences which required further cloning steps for correct genotyping - a complication which makes this form of analysis prohibitively expensive. This has since also been demonstrated by others (e.g. Yang *et al.* 2015). I

was forced to venture into the comparatively new field of population genomics as a potential solution to dealing with polyploidy by using a reduced-representation sequencing approach and bioinformatics tools to manage the occurrence of paralogs. In this regard I was partially successful - I was able to identify and remove many potential paralogous loci from the analysis, and, although I give evidence that I was not able to remove all paralogous loci (as shown by the F_{IS} estimates in Chapters 2 and 4), I was ultimately able to answer my research questions using nuclear loci.

The reduced-representation approach selected, ddRAD sequencing, is an alternative to traditional RAD sequencing which uses a pair of enzymes to cut throughout the genome instead of relying on sonication (Peterson *et al.* 2012). In this study a pair of common-cutting enzymes were selected to maximise the coverage across the genome, however I found that this was far too ambitious given the relatively low sequencing depth. This resulted in a large number of low-depth loci which were ultimately filtered from the SNP identification pipelines (as recommended in Andrews *et al.* 2016), as shown by the difference observed between the nuclear mappings in *L. natalensis* where a large proportion of the genome appears to be represented, but relatively few loci pass through the SNP identification pipeline. This means that I would have probably acquired far more SNPs if I had used a rare-cutting enzyme in combination with one of the common-cutting enzymes due to the improved depth across comparatively fewer sites. I had the opportunity to attempt this with the second and third libraries (Orange-Vaal libraries 1 and 2 in Chapters 3 & 4); however I felt that this would make the results between libraries incomparable. Despite the high coverage and low depth attained, I ensured that loci in the final datasets were of a sufficient depth for further analyses as per Paris *et al.* (2017) and Rochette & Catchen (2017).

The choice of restriction enzymes had another unanticipated side effect - I was able to demonstrate that the choice of *Nla*III as a restriction enzyme resulted in a slight but noticeable bias in in-frame coding regions, assumably due to the presence of an 'ATG' within the short restriction

site. This was most visible in a SNP density spectrum (Fig. S4.1) where every third position shows an elevated number of SNPs, although several other analyses also indicated a bias in genic regions (e.g. transition/transversion ratios in all libraries, proportion of reads mapped to CDS for *L. natalensis*). I posit that this is due to ‘wobble’ occurring at third positions of in-frame coding sequences. Although this particular finding appears to be unprecedented, similar biases have been shown for GC-rich regions of the genome in the literature (Amores *et al.* 2011; Arnold *et al.* 2013; DaCosta & Sorenson 2014; Gonen *et al.* 2015). This suggests that future ddRAD sequencing projects could utilise *Nla*III together with a rare-cutting enzyme to select for regions in the genome which are more likely to match in-frame coding regions with a depth sufficient to retain most loci for further analysis due to the rare-cutting enzyme. This is beneficial as although one of the major advantages of RAD and ddRAD sequencing is random sampling across the genome, researchers are often far more interested to acquire markers in coding sequences due to the constraints on these regions (Cooper & Brown 2008).

A negative aspect of the ddRAD sequencing libraries generated in this study, particularly for *L. natalensis*, is that a substantial proportion of adapter pollution was identified, possibly as a result of inefficient size selection. This led to some initial difficulties in the analysis, but was ultimately accounted for by filtering the reads affected by adapter pollution. This effect was not observed in library 3, which was sequenced last at the greatest coverage using two sets of size selection steps. Incomplete restriction digestion also appeared to be relatively common in the datasets, which has the property of increasing genome coverage while decreasing depth. This was useful for some applications of the study - such as partial mitogenome reconstruction - but may have undermined some of the SNP identification process by resulting in fewer loci of sufficient depth making it through the pipeline.

A further technical difficulty arising from the ddRAD sequencing protocol was library bias. The research on Orange-Vaal yellowfish relied on two libraries that were sequenced independently a year apart using the same protocol, but with different machines (Illumina HiSeq 2000 & 4000) and different samples (although both species were included in each library). In analyses using both libraries I detected consistent ‘genetic’ differences between libraries instead of a split between the two species. This may have been caused by slight differences in methodology between the two libraries, or may simply be inherent library bias as demonstrated previously in red drum where multiple libraries sequenced using the exact same methodology at the same time grouped differently (Puritz *et al.* 2015). I suspect that differences in sequencing depth between libraries may be at least partially responsible. One sample of Orange-Vaal library 1, which was sequenced to a greater degree than any other sample, clustered with the higher-coverage Orange-Vaal library 2 and the lowest-coverage sample of Orange-Vaal library 2 often grouped by itself or closely to Orange-Vaal library 1. These observations suggest that the confounding effect of library bias is best countered by completing all sequencing at the same time in the same library if possible. Failing this, strict filtering criteria should be employed to standardise the loci across libraries and minimise the effects of any bias.

One of the caveats of RAD sequencing is the occurrence of null alleles caused by mutations within the recognition sequence for the designated restriction enzymes used (Arnold *et al.* 2013; Andrews *et al.* 2016). This is exacerbated in ddRAD sequencing due to the requirement for two different restriction sites to be present (Arnold *et al.* 2013). In this study I showed evidence of this occurrence particularly within the *L. natalensis* library where locus dropout within different populations was common, and accounted for almost all of the missing data observed within the final dataset. This suggests that a degree of tolerance for missing data can and should be allowed in SNP

analyses such as these, as they are caused by real biological signal indicating differences at the population level.

5.4 Data filtering and SNP identification

I have demonstrated several results that emphasise the importance of intense data filtering during the SNP identification process. Two unexpected observations were the effect of mitochondrial reads and singleton alleles on the genetic structure observed. If unfiltered, I found that mitochondrial loci influenced the genetic structure results, allowing identification of a lineage which is differentiated using mitochondrial but not nuclear markers in *L. natalensis*. In contrast, singleton alleles were found to mask genetic structure, resulting in analyses that concluded that there was either no structure or only a northern/southern lineage split. Mitochondrial loci are not usually filtered from RAD analyses, but my results suggest that their effect should be monitored in future studies. Singleton alleles are often filtered from RAD studies by employing a MAF filter, however this is usually to preclude sequencing errors rather than to remove these loci whose low F_{ST} result in a loss of differentiation between taxa (but see Roesti *et al.* 2012; Rodríguez-Ezpeleta *et al.* 2016).

I found a surprising difference between the species investigated in this study when I assessed loci potentially under selection. Both the *L. natalensis* and Orange-Vaal libraries revealed relatively few highly differentiated outlier loci. The *L. natalensis* loci potentially under selection were found to show the same signal as putatively neutral loci and were thus included in the final dataset. However, the Orange-Vaal loci potentially under selection were found to produce much of the signal of genetic differentiation between the two species - to the extent that the species could not be differentiated in STRUCTURE (Pritchard *et al.* 2000) following their removal (Fig. S4.6). This would suggest that these outlier loci may be driving the differentiation between these species, and demonstrates the

importance of this filtering step as this and the background of relative homogeneity may otherwise not have been determined.

Another complication alluded to above, which is of particular relevance for this hexaploid genus, is how to deal with paralogous loci. I employed several methods to identify and remove paralogs: removing loci with an excess of haplotypes within at least one sample, STACKS (Catchen *et al.* 2011; Catchen *et al.* 2013) default paralog filters, and a filter to remove loci which were heterozygous for at least one SNP across all samples. Although these filters removed a substantial proportion of SNPs for all species, the F_{IS} values obtained from the summary statistics indicated that paralogous loci were still present in the final datasets despite my best efforts. This suggests that paralogous loci are resilient to these filters, and additional methods may need to be employed to remove them. I briefly explored the use of the new program HDPLLOT (McKinney *et al.* 2017) which identifies duplicated loci; however I found it difficult to implement given the sample size and read depth. Further research using this and other paralog filters is needed. Alternatively, identifying analyses which are robust to the effects of a small proportion of paralogous loci may be more reasonable as it may not be possible or feasible to remove all paralogs from a dataset.

Despite, or perhaps because of, difficulties encountered in RAD parameter optimisation and SNP identification, a number of publications have recently been released to aid these processes (e.g. Paris *et al.* 2017; Rochette & Catchen 2017). Although I did not have access to these frameworks at the start of this project, I was able to design a similar approach to that presented in both of these publications, which required minimal optimisation following the release of these papers. Due to inconsistencies between studies using RAD data, these unifying approaches are sorely needed. Growth in this field and in the methodology is most apparent from publications such as this, and underscores the importance of making data available and applicable to others.

One additional step used in this study but not found in these publications, and, in fact, contrary to a number of guides on SNP identification using STACKS is the selection of the ‘best’ SNP per locus used in Chapters 2 and 4. An anonymous reviewer of the paper resulting from the output of Chapter 2 suggested that, instead of selecting the first SNP at a locus, as is common in the literature, one should rather select each SNP based on allele frequency (if choosing a single SNP per locus is necessary for the analysis). This allowed me to establish the effect that singleton alleles were having on the dataset (as mentioned above) and improved the power of the analyses.

5.5 Assessing population structure and demography

In addition to the difficulties with mtDNA markers and singleton alleles influencing results of the tests for genetic structure mentioned above, I also experienced additional complications in resolving structure within *L. natalensis* above $K = 2$. This may be because the signal observed at this split between the two lineages overwhelms the finer-scale differentiation between populations. This trend to obtain $K = 2$ even when it should not be the case has been explored in the literature (Janes *et al.* 2017), but was resolved here by running a hierarchical STRUCTURE analysis on the two lineages identified at $K = 2$. Researchers should be aware of biases such as this to ensure that tests to counter or detect these biases are run.

I further reconstructed the demographic history of *L. natalensis* using the nuclear SNPs identified in this study. Although I was able to show support for a scenario that appeared similar to the results of my other analyses, I would predict that the real demographic history is likely more complicated than this model depicts, and the current choice of scenario is the closest approximation out of the available options. This provides a framework for further investigation of the demographic history of this species in future studies. Likewise, the estimates of long-term effective population size

calculated here for this species are extremely broad due to the ten-fold range in SNP mutation rate used. These estimates can be used as a rough guideline which can be improved on as estimates of SNP mutation rates in fish improve. However, it must be stated again that these estimates are not necessarily correlated with the contemporary census size, but rather reflect importance in terms of population diversity which may be useful for conservation management.

5.6 Conservation and management

The population genomics approach identified several areas of conservation concern for *L. natalensis* in particular, namely: that the nucleotide diversity of the smallest population which appears to be restricted to a single river is greater than that of the other, far larger populations; the Umgeni lineage appears to have its own native mitochondrial lineage and a nuclear lineage demonstrating admixture between the northern and southern lineages; and that this Umgeni lineage is currently undergoing a potential invasion by translocated Tugela individuals from the Mooi River via the MMTS. This contemporary admixture event is perhaps the most pressing aspect for conservation. Additional sampling would be beneficial to ascertain the current situation within the Umgeni system since the water transfer event has been ongoing for some time now, and to extend the sampling to the southern limit of the species distribution in the Mtamvuna River. In addition, the unexpectedly high diversity in the Mbokodweni may indicate that this population is more widely distributed than currently believed. This again emphasises the importance of further sampling, particularly in the coastal areas of the distribution such as the adjacent Umlazi and Illovo Rivers. The Orange-Vaal yellowfish would benefit from further sampling in both the upper Orange River, which was underrepresented in this study, and in the Vaal River, which was not represented at all. Divergence

between the upper and lower Orange River *L. aeneus* is a priority for further research in the effective management of these two putative populations.

Interspecific hybridisation between the Orange-Vaal yellowfish is another conservation concern; however this appears to occur relatively infrequently, possibly in areas where the two species are forced to share the same breeding grounds. Further investigation of the mechanisms by which hybridisation and introgression are occurring, as well as tests to differentiate between the competing hypotheses of ancestral polymorphism and hybridisation are urgently needed prior to the development of a management scheme to promote the genetic integrity of these species.

5.7 Future directions

This project opens the door for future research on the hexaploid fish of the genus *Labeobarbus*. Much can be done to extend the work described in this study, particularly to the other South African taxa not focussed on here (*L. marequensis*, *L. nelspruitensis*, *L. polylepis*, *L. seeberi*). The work on *L. natalensis* can be extended by increasing the sampling as discussed above to include areas of importance which were not sampled here and to ascertain the current level of hybridization in the Umgeni system. More advanced models of demographic history could be explored in this lineage, as none of the simple tested models here appeared particularly realistic. The Orange-Vaal yellowfish study could be extended to assess hybridization in the Vaal River, and include more samples from the upper Orange River. This would allow further examination of the previously-recorded observation of divergence within *L. aeneus* of the upper and lower Orange River, which I did not find here. Further investigation of the genomic components driving the divergence between these species would also be a good avenue for further research, and the highly differentiated outlier loci found in this study to be potentially driving this divergence are good candidates for additional

studies. The possibility of ancestral polymorphism resulting in the signal of introgression observed between these species should also be assessed. Above all, the future construction of a reference genome would greatly enhance all upcoming genomic studies in this genus and would allow for a better understanding of the genomic architecture of complex polyploid genomes.

5.8 Conclusion

In closing, I have investigated the phylogeography of three species of South African yellowfish using population genomics tools. I was able to delimit five populations of *L. natalensis* across the distribution separated by drainage system, which is contrary to the six mitochondrial populations identified both here and previously. I have demonstrated that mitochondrial loci are present in ddRAD sequencing datasets and can be mined to investigate mitochondrial signal. This proved to be highly useful in investigations of intra- and interspecific hybridisation. I have shown that *L. aeneus* and *L. kimberleyensis* are genetically distinct species. And finally, I identified evidence of intraspecific hybridisation within *L. natalensis* and interspecific hybridisation and introgression between *L. aeneus* and *L. kimberleyensis*. I hope that the research in this study together with the approaches taken serve as a useful guideline for future research, particularly in but not limited to the magnificent hexaploid yellowfishes.

5.9 References

- Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics* **188**, 799-808.
- Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* **17**, 81-92.

- Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology* **22**, 3179-3190.
- Banister KE. 1972. On the cyprinid fish *Barbus alluaudi* Pellegrin: a possible intergeneric hybrid from Africa. Studies on African Cyprinidae. Part I. *Bulletin of the British Museum (Natural History)*, *Zoology* **24**, 261–290.
- Bloomer P, Stobie CS, Reid K, Oosthuizen CJ (In prep) Divergent mitochondrial lineages in the KwaZulu-Natal yellowfish (*Labeobarbus natalensis*) across ancient river systems. Target journal: *BMC Evolutionary Biology*.
- Boulenger GA. 1902. Descriptions of new fishes from the collection made by Mr. E. Degen in Abyssinia. *Annals and Magazine of Natural History (Series 7)* **10**, 421–439.
- Cambray JA, Bills R, Chakona A, Coetzer W, Weyl O (2017) *Labeobarbus natalensis*. The IUCN Red List of Threatened Species 2017: e.T63294A100168851. Downloaded on **07 December 2017**. <<http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T63294A100168851.en>>
- Catchen JM, Amores A, Hohenlohe PA, Cresko WA, Postlethwait JH (2011) Stacks: Building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1**, 171-182.
- Catchen JM, Hohenlohe P, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.
- Crass RS (1964) *Freshwater Fishes of Natal*. Shuter & Shooter, Pietermaritzburg.
- Chenuil A, Galtier N, Berrebi P (1999) A test of the hypothesis of an autopolyploid vs. allopolyploid origin for a tetraploid lineage: application to the genus *Barbus* (Cyprinidae). *Heredity* **82**, 373-380.
- Cooper GM, Brown CD (2008) Qualifying the relationship between sequence conservation and molecular function. *Genome Research* **18**, 201-205.
- Crampton-Platt A, Douglas WY, Zhou X, Vogler AP (2016) Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience* **5**, 15.
- DaCosta JM, Sorenson MD (2014) Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE* **9**, e106713.
- Eccles D (1986) Development of the gut in the South African cyprinid fish *Barbus aeneus* (Burchell). *African Zoology* **21**, 165-169.
- Gaigher I (1976) The reproduction of *Barbus cf. kimberleyensis* (Pisces, Cyprinidae) in the Hardap dam, South West Africa. *African Zoology* **11**, 97-110.
- Gonen S, Bishop SC, Houston RD (2015) Exploring the utility of cross-laboratory RAD-sequencing datasets for phylogenetic analysis. *BMC Research Notes* **8**, 299.
- Janes JK, Miller JM, Dupuis JR, *et al.* (2017) The K = 2 conundrum. *Molecular Ecology* **26**, 3594-3602.
- Karssing RJ (2008) Status of the KwaZulu-Natal yellowfish *Labeobarbus natalensis* (Castelnau, 1861). In: *Technical Report on the State of Yellowfishes in South Africa 2007*. (eds. Impson ND, Bills IR, Wolhuter L), pp. 31-45. Water Research Commission, Gezina, Pretoria.
- Li S, Besenbacher S, Li Y, *et al.* (2014) Variation and association to diabetes in 2000 full mtDNA sequences mined from an exome study in a Danish population. *European Journal of Human Genetics* **22**, 1040-1045.
- McKinney GJ, Waples RK, Seeb LW, Seeb JE (2017) Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources* **17**, 656-669.
- Mulder P, van Vuuren N, Ferreira J, van der Bank F (1990) A preliminary biochemical comparison of two species of the genus *Barbus* from the Vaal River system. *Water S.A.* **16**, 147-150.
- Paris JR, Stevens JR, Catchen JM (2017) Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution* **8**, 1360-1373.

- Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7**, e37135.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.
- Pujolar JM, Jacobsen M, Als TD, *et al.* (2014) Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology* **23**, 2514-2528.
- Puritz JB, Hollenbeck CM, Gold JR (2015) Fishing for selection, but only catching bias: library effects in double-digest RAD data in a non-model marine species. Poster presented at: *Plant and Animal Genomes XXIII*, 10-15 January 2015, San Diego, CA.
- Rhymer JM, Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics* **27**, 83-109.
- Rochette NC, Catchen JM (2017) Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Protocols* **12**, 2640-2659.
- Rodríguez-Ezpeleta N, Bradbury IR, Mendibil I, *et al.* (2016) Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection. *Molecular Ecology Resources* **16**, 991-1001.
- Roesti M, Salzburger W, Berner D (2012) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology* **12**, 94.
- Tang M, Tan M, Meng G, *et al.* (2014) Multiplex sequencing of pooled mitochondrial genomes - a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research* **42**, e166.
- Truong HT, Ramos AM, Yalcin F, *et al.* (2012) Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS ONE* **7**, e37565.
- van Vuuren N, Mulder P, Ferreira J, van der Bank FH (1989) The identification of hybrids of *Barbus aeneus* X *B. kimberleyensis* and *Labeo capensis* X *L. umbratus* in Hardap Dam, SWA/Namibia. *Madoqua* **16**, 27-34.
- Vargas OM, Ortiz EM, Simpson BB (2017) Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: *Diplostephium*). *New Phytologist* **214**, 1736-1750.
- Vreven EJ, Musschoot T, Snoeks J, Schliewen UK (2016) The African hexaploid Torini (Cypriniformes: Cyprinidae): review of a tumultuous history. *Zoological Journal of the Linnean Society* **177**, 231-305.
- Yang L, Sado T, Hirt MV, *et al.* (2015) Phylogeny and polyploidy: resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). *Molecular Phylogenetics and Evolution* **85**, 97-116.
- Zhang N, Wen J, Zimmer EA (2015) Congruent deep relationships in the grape family (Vitaceae) based on sequences of chloroplast genomes and mitochondrial genes via genome skimming. *PLoS ONE* **10**, e0144701.