Received: 5 May 2017 | Accepted: 17 December 2017

DOI: 10.1002/ece3.3821

ORIGINAL RESEARCH

WILEY Ecology and Evolution Open Access

Exploring the phylogeography of a hexaploid freshwater fish by RAD sequencing

Cora Sabriel Stobie | Carel J. Oosthuizen | Michael J. Cunningham | Paulette Bloomer

Molecular Ecology and Evolution Programme, Department of Genetics, University of Pretoria, Pretoria, South Africa

Correspondence
Cora Sabriel Stobie and Paulette Bloomer, Department of Genetics, University of Pretoria, Pretoria, South Africa.
Emails: cora.stobie@gmail.com; paulette.bloomer@up.ac.za

Funding information
Genomics Research Institute, University of Pretoria; National Research Foundation, Grant/Award Number: 77240

Abstract

The KwaZulu-Natal yellowfish (*Labeobarbus natalensis*) is an abundant cyprinid, endemic to KwaZulu-Natal Province, South Africa. In this study, we developed a single-nucleotide polymorphism (SNP) dataset from double-digest restriction site-associated DNA (ddRAD) sequencing of samples across the distribution. We addressed several hidden challenges, primarily focusing on proper filtering of RAD data and selecting optimal parameters for data processing in polyploid lineages. We 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. We 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.

KEYWORDS
genotyping-by-sequencing, polyploidy, population genomics, population history

1 | INTRODUCTION

The Cyprinidae is the largest freshwater fish family, comprising approximately 80% of all freshwater fish species in temperate zones (Naran, Skelton, & Villet, 2007) and including over 2,400 species (de Graaf, Nagelkerke, Palstra, & Sibbing, 2010; Swartz, Mwale, & Hanner, 2008). 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 hybridization between a tetraploid ancestor in Torini and a diploid ancestor of *Cyprinion* followed by autopolyploidization, resulting in the current hexaploid lineage (2N = ±150 chromosomes) sometime prior to their colonization of Africa *c.* 13 mya (Oellermann & Skelton, 1990; Tsigenopoulos, Kasapidis, & Berrebi, 2010; Yang et al., 2015). This lineage has subsequently speciated into 125 valid species (Vreven, Musschoot, Snoeks, & Schliewen, 2016).

Seven species of *Labeobarbus* exist in southern Africa. Five of these (*Labeobarbus aeneus*, *Labeobarbus capensis*, *Labeobarbus*

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2018 The Authors. *Ecology and Evolution* published by John Wiley & Sons Ltd.

kimberleyensis, *L. natalensis*, and *Labeobarbus 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, Goodman, & Nkosi, 2007) which were later colonized 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, Bills, Chakona, Coetzer, & Weyl, 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' geographic range (Palumbi, 2003; Smith & Bermingham, 2005). Various phylogeographic studies have been conducted on cyprinids (Durand, Tsigenopoulos, Ünlü, & Berrebi, 2002; Machordom & Doadrio, 2001) although few of these have explored the South African branches of the family (but see Chakona, Swartz, & Gouws, 2013; Chakona, Malherbe, Gouws, & Swartz, 2015; Chakona & Skelton, 2017; Swartz, Skelton, & Bloomer, 2007, 2009; Swartz, Chakona, Skelton, & Bloomer, 2014; van der Walt, Swartz, Woodford, & Weyl, 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. Unpublished data). 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. Unpublished data). 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 (Perera, Ratnayake-Perera, & Proches, 2011; Rivers-Moore et al., 2007). 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, we decided to use a reduced representation approach, restriction site-associated DNA (RAD) sequencing (Baird et al., 2008; Miller, Dunham, Amores, Cresko, & Johnson, 2007), to understanding genomic diversity in *L. natalensis*. This method is popular and has been used in many studies since its inception (Figure S1). RAD sequencing has been used, particularly in fish, to identify population divergence (Boehm, Waldman, Robinson, & Hickerson, 2015; Ferchaud & Hansen, 2016; Larson et al., 2014), for SNP identification in polyploid fish (Hohenlohe, Amish, Catchen, Allendorf, & Luikart, 2011; Ogden et al., 2013; Palti et al., 2014), in phylogeographic studies (Macher et al., 2015; Reitzel, Herrera, Layden, Martindale, & Shank, 2013), for QTL analysis (Gagnaire, Normandeau, Pavey, & Bernatchez, 2013; Houston et al., 2012; Yoshizawa et al., 2015), for linkage mapping (Brieuc, Waters, Seeb, & Naish, 2014; Henning, Lee, Franchini, & Meyer, 2014), in hybridization studies (Hand et al., 2015; Lamer et al., 2014; Pujolar et al., 2014), for exploration of genome architecture and evolution (Brawand et al., 2014; Kai et al., 2014; Waples, Seeb, & Seeb, 2016), and in phylogenetic analyses (Gonen, Bishop, & Houston, 2015; Wagner et al., 2013). 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).

Double-digest RAD (ddRAD) sequencing (Peterson, Weber, Kay, Fisher, & Hoekstra, 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 artifacts; however, as mutations in restriction sites may result in underestimation of diversity due to allele dropout (Arnold, Corbett-Detig, Hartl, & Bomblies, 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, Grau, & Seeb, 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, Waples, Seeb, & Seeb, 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, 2013). Recently, McKinney

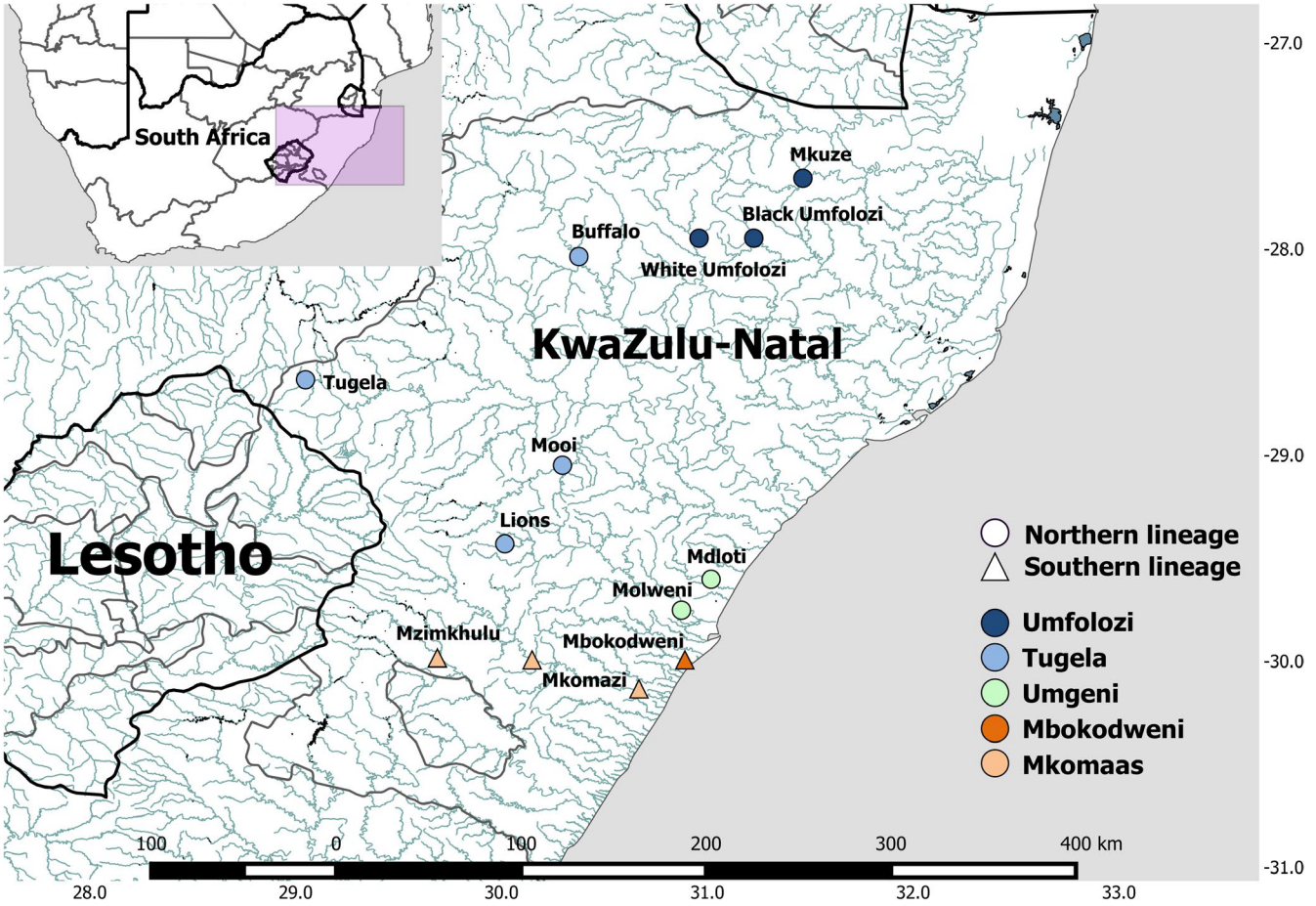

FIGURE 1 Distribution of samples in this study across KwaZulu-Natal with reference to a map of South Africa (top left). River names are indicated at each sampling site. The color of each sampling site corresponds to the putative population identified in this study. The shape of the symbol indicates an 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)

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

2 | MATERIALS AND METHODS

2.1 | Sample collection and DNA extraction

Samples were collected from 13 localities across the KwaZulu-Natal Province, South Africa (Figure 1; Table S1) between March 2003 and November 2007. 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, Fritsch, & Maniatis, 1989).

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 (Table S1).

2.2 | Library preparation and sequencing

Samples were divided into two libraries, which were digested with the restriction enzymes *Nla*III (CATG/) and *Mlu*CI (/AATT) following the double-digest paired-end protocol of 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.

2.3 | 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, Amores, Hohenlohe, Cresko, & Postlethwait, 2011; Catchen, Hohenlohe, Bassham, Amores, & Cresko, 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 (Figure S2). Quality of the bases was assessed for each sample both before and after trimming using the program FASTQC (Andrews, 2010; Figures S3 and S4). 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. We 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. Nonspecific matches were ignored.

2.4 | Single-nucleotide polymorphism 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, 2013). We chose to use only our 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, Stevens, and Catchen (2017). Briefly, we constructed plots of the average read depth across samples while varying the minimum stack depth parameter *−m* (Figure 2). We then compared the number of SNPs, assembled loci, and polymorphic loci for each sample and across samples using the 80% sample representation cutoff 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 (Figures S5 and S6). 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 $n = M + 1$ (Table S2). 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, we identified optimal parameters for this dataset to be *−m* 5 *−M* 1 *−n* 0.

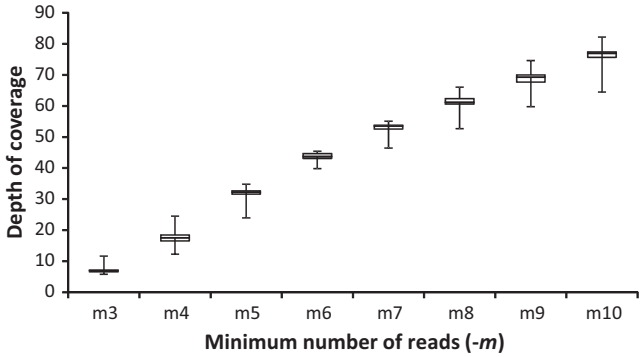

FIGURE 2 Box-and-whisker plot showing the distribution of the mean depth of coverage across all 23 samples (plus two replicates) as the value for the minimum number of reads (*−m*) is varied from 3 to 10, as per Paris et al. (2017). Whiskers here indicate the maximum and minimum values across the samples. Paris et al. (2017) recommend a depth of coverage of >25×, 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). We removed potential paralogs that had been merged by identifying loci that possessed more than two haplotypes within a single sample. We compared this haplotype approach with the HDPLOT method of McKinney et al. (2017). A blacklist was constructed in POPULATIONS for loci identified as mitochondrial or paralogous.

We retained loci from POPULATIONS that were scored in at least 60% of all individuals. We used a minor allele frequency (MAF) filter of 0.04 to filter out singleton SNPs that may mask population structure (Rodríguez-Ezpeleta et al., 2016; Roesti, Salzburger, & Berner, 2012), 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, we filtered our 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, we 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, which was used in all downstream analyses unless specified otherwise.

TABLE 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

2.5 | 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, Stephens, & Donnelly, 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. Location was specified as a prior. We 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, Regnaut, and Goudet (2005). CLUMPP 1.1.2 (Jakobsson & Rosenberg, 2007) was used to visualize the data.

In addition to STRUCTURE, RADPAINTER and FINERADSTRUCTURE (Malinsky, Trucchi, Lawson, & Falush, 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, Borsa, Chikhi, Raufaste, & Bonhomme, 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 loci under selection 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 all loci, and excluding outlier loci. Finally, we identified a prevalent *Hin*dIII 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.

2.6 | 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. We first tested six basic scenarios of a basal split separating one population from all others or a basal polytomy (Figure S7A). We then tested 24 ladder-like scenarios of successive divergence (Figure S7B). We also tested ten scenarios where the ancestral population split into two major lineages which then diverged into five regional populations (Figure S7C). Finally, we compared the best supported scenarios from the above tests as well as two variants that model the Umgeni population as a product of admixture (Figure S7D). 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 (π = 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, Beerli, Nickerson, & Edwards, 2003).

3 | RESULTS

3.1 | Library preparation and sequencing

Illumina 90-bp paired-end sequencing produced over 137 million reads (Table 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%–94.81%.

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). 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.

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. Reads contaminated by adapter pollution were filtered with custom scripts using regular expressions based on restriction enzyme recognition sequences. We observed a high degree of incomplete enzyme digestion, with 9% of reads containing more than one of the same restriction site.

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). The two methods for identification of potential paralogs, excess haplotypes per individual and HDPLOT, produced very different results (see Discussion below). These methods identified 2,435 and 463

TABLE 2 Mapping of RAD sequencing reads to the *Cyprinus carpio* nuclear (GCF_000951615.1) genome

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

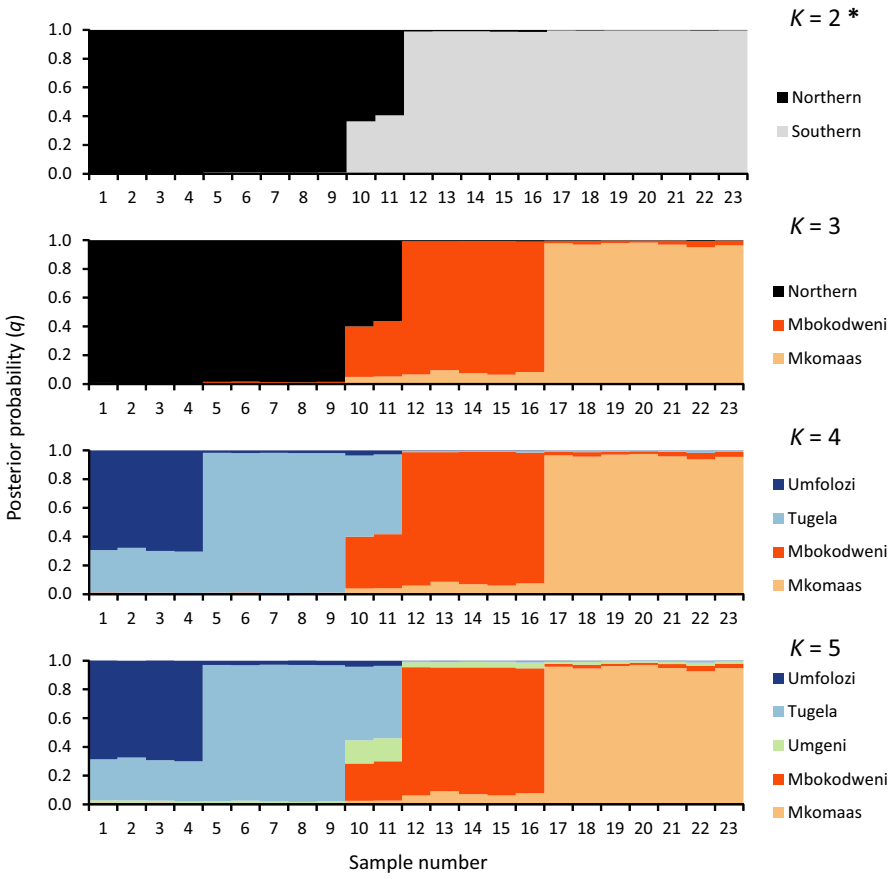

FIGURE 3 STRUCTURE analysis with K = 2–5 using 723 high-quality filtered loci. Each individual (indicated as columns along the X-axis) is probabilistically assigned (probability of assignment q on the Y-axis) to one of the inferred genetic clusters. Location was specified as prior. K = 2 was recovered as having the most support with the Evanno method (Evanno et al., 2005). CLUMPP was used to produce this representation from 20 replicates

TABLE 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; SNP, single-nucleotide polymorphism.

loci, respectively (Figure S8). We 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 summarized in Table 3 and distribution of SNPs across individuals in Figure S8.

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 (Figure 3;

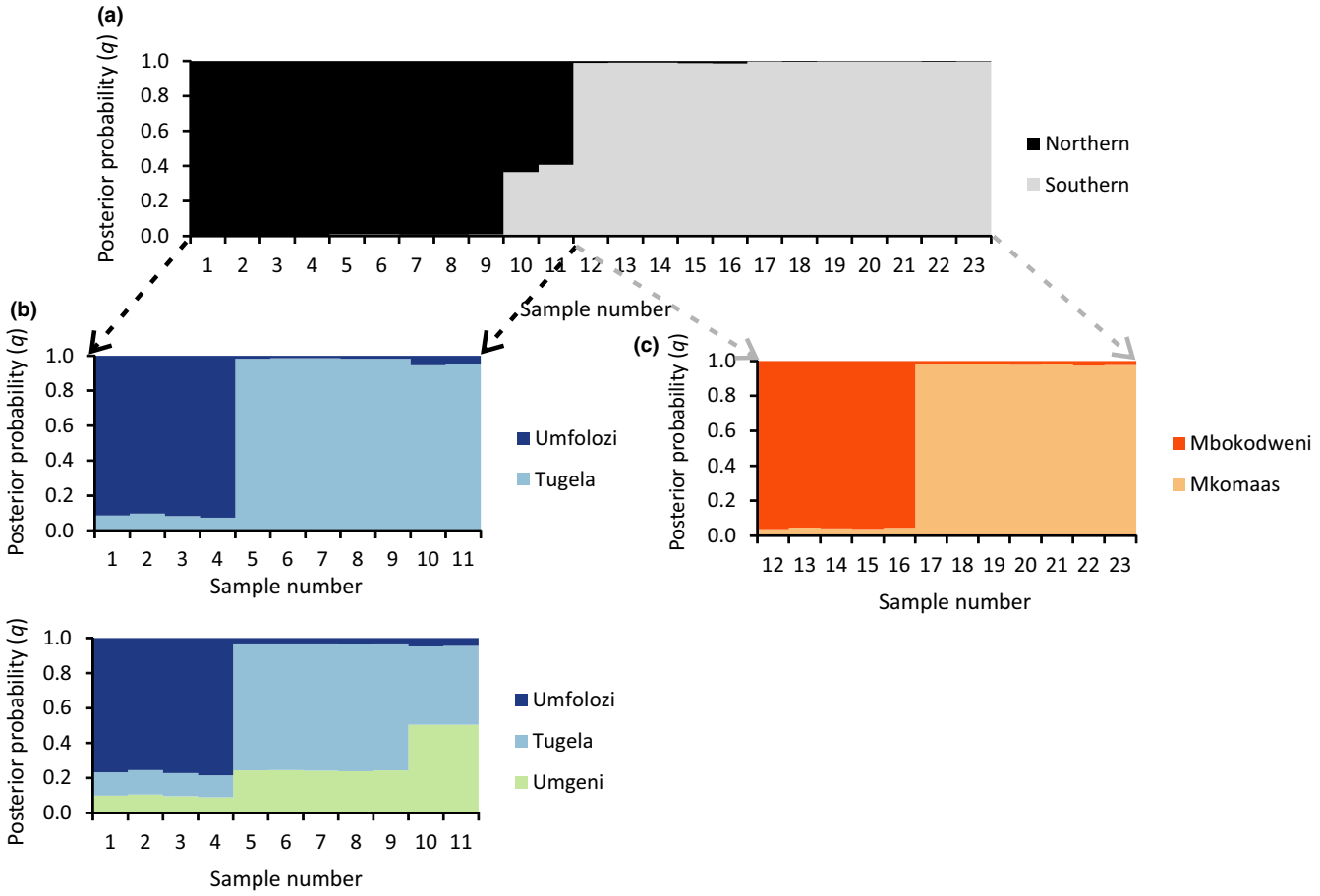

FIGURE 4 Hierarchical STRUCTURE results for the northern and southern lineages. Each individual (indicated as columns along the *X*-axis) is probabilistically assigned (probability of assignment *q* on the *Y*-axis) to one of the inferred genetic clusters. (a) Primary run of STRUCTURE for $K = 2$ on 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

Figures S10 and S11). 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 (Figure 4) showed further subdivision into five populations: Umfolozi, Tugela, Umgeni, Mbokodweni, and Mkomaas, from north to south. These results were not well supported (Figures S12–S15). The two samples from the Mzimkhulu system, near the southern distribution limit, were not distinguished from those in the adjacent Mkomaas, and we 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).

FineRADstructure was used to generate a co-ancestry matrix and tree (Figure 5) showing five populations—the Umfolozi, Tugela, Mkomaas, Mbokodweni, and Umgeni. Further support for these populations was shown in PCA and FCA plots from FineRADstructure (Figure 6) and Genetix (Figure 7). The plots also display variance within populations. The Umgeni was shown to be similar to both neighboring 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). We identified 24 loci as potential F_{ST} outliers. However, we did not find any difference in genetic signal when comparing these to other loci, and so all loci were combined for downstream analyses. Removing *Hind*III satellite loci was found to produce similar results, leading to their retention in the final dataset.

The number of private alleles varied from 35 in Umfolozi to 57 in the Mkomaas population, except Umgeni which had only six private alleles (Table 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

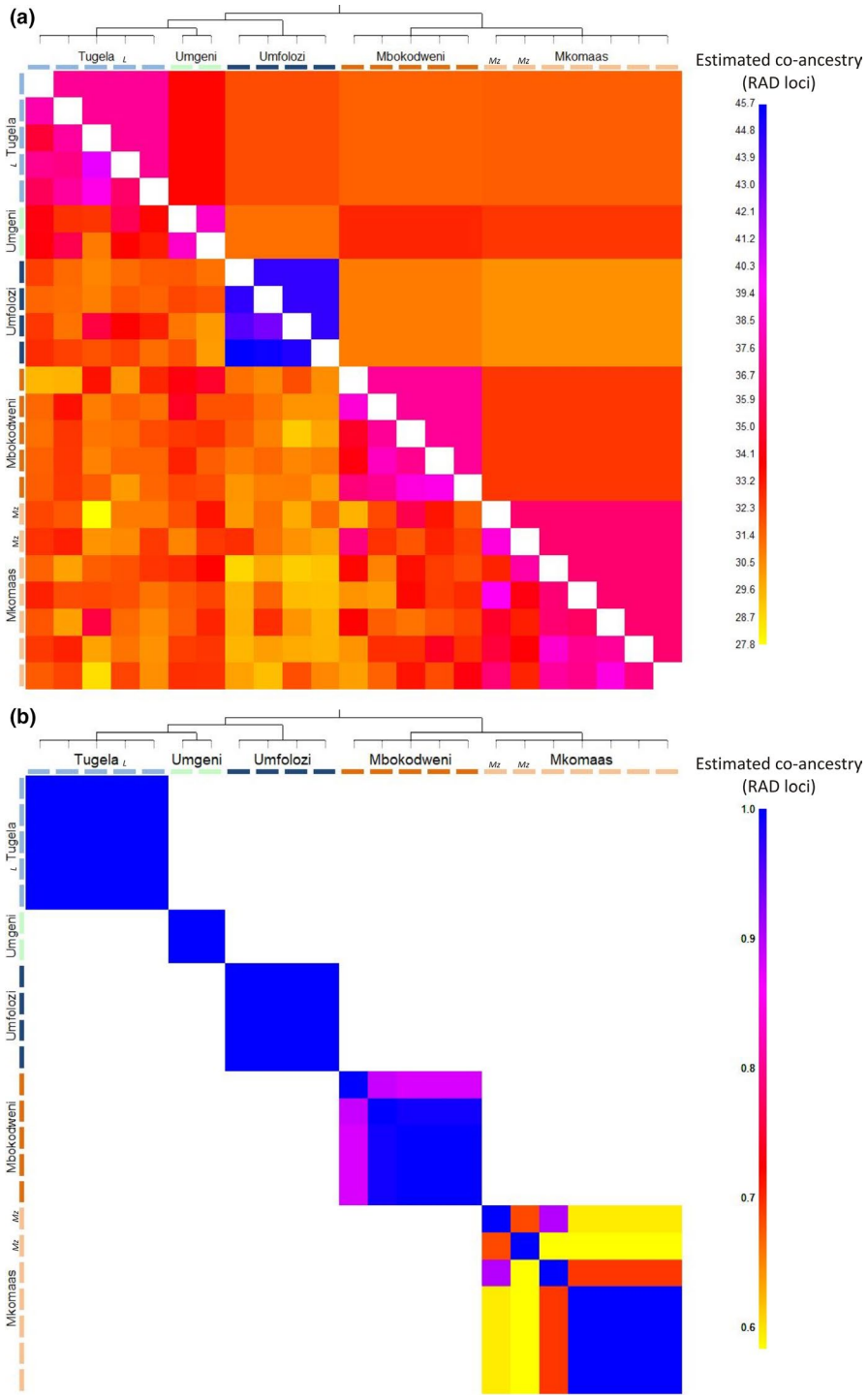

FIGURE 5 (a) FINERADSTRUCTURE co-ancestry matrix, indicating pairwise co-ancestry between individuals. The lower diagonal indicates raw copy numbers whereas the upper diagonal shows aggregated copy numbers. Individuals clustering into populations are indicated by clustering in the accompanying tree and along the diagonal of the plot. The sample labeled "*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

77 alleles which are not found in the north. Extending this approach, we removed the potentially admixed Umgeni lineage and categorized each allele as private within a single population, shared between two or more populations only, or shared between all populations (Figure 8).

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 5) were significant at $p = .01$. All pairwise comparisons with Umgeni yielded negative nonsignificant values, probably because this sample comprised two individuals with indications of admixture. The highest pairwise F_{ST} values were recorded for the Umfolozi drainage

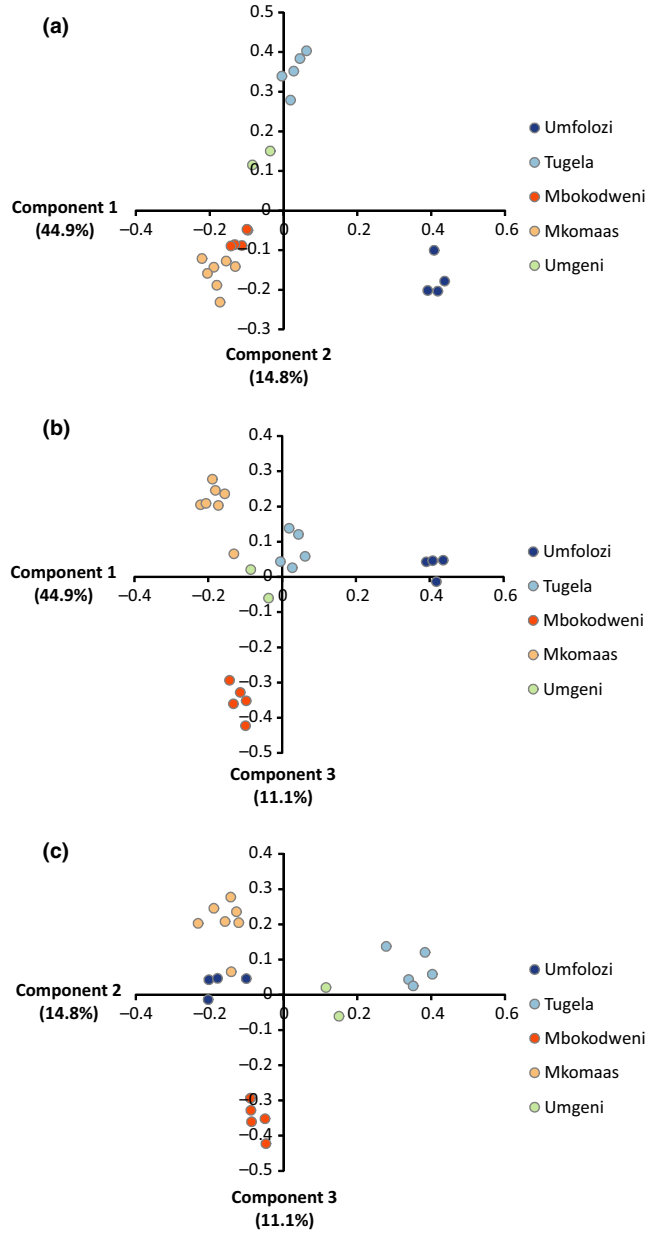

FIGURE 6 Principal components analysis plots indicating the distribution of individuals into populations according to the first three principal components identified in FINERADSTRUCTURE, accounting for 70.8% of the eigenvalues. Component 1 splits Umfolozi from the rest. Component 2 splits Tugela (and 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

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).

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 (Figure 9). 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 (Figure S7D). The effective population size is estimated to lie between 87,500 and 875,000 individuals.

4 | DISCUSSION

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 zebrafish, *Danio rerio*, genome (38.6%) (Zhou, Bizzaro, & Marx, 2004) giving confidence that these data were not particularly biased by our choice of AT-rich restriction sites (contrary to Campagna, Gronau, Silveira, Siepel, & Lovette, 2015; DaCosta & Sorenson, 2014). The relatively high transition/transversion ratio of 2.04 after filtering for paralogs and mitochondrial loci may indicate a bias toward genic regions, as SNPs occur more frequently as transitions in exons than in introns (Park, Yu, Mun, & Lee, 2010). This value may also reflect effective filters to reduce sequencing error in the final SNP dataset (Pujolar et al., 2013; Rašić, Filipović, Weeks, & Hoffmann, 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 (Figure S2) 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 (Figure S2), with further errors removed by filtering for singleton allele SNPs 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*. As the hexaploid lineage *L. natalensis* has 150 chromosomes (Oellermann & Skelton, 1990) versus 100 in the tetraploid lineage *C. carpio* (Ráb, Pokorný, & Roth, 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 toward 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 nontarget fragments may be favored 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.

4.2 | Bioinformatics and SNP discovery

A paired-end ddRAD sequencing approach was used 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 (Boehm et al., 2015; Macher et al., 2015; Willing, Dreyer, & Van Oosterhout, 2012).

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

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 nontarget fragments is 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 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, we were able to obtain loci of sufficient coverage (>25×) 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 cutoff should be favored in analyses (Mastretta-Yanes et al., 2015). However, setting the cutoff value too high would result in allele dropout, leading to further errors (Mastretta-Yanes et al., 2015). Here, we 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 us to generate intuitive graphical representations of how parameters in STACKS influenced our dataset (Figure 2; Figures S5–S7; Table S2). From these plots, we 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, we employed two further approaches: identifying loci with more than two haplotypes in a sample and the HDPLOT 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 HDPLOT method was difficult to interpret due to the sparsity of samples (Figure S8), 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 HDPLOT 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, we 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 HDPLOT 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 our best efforts at isolating paralogous loci, the F_{IS} output by STACKS reported negative values which are indicative of paralogous loci retained in our 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 (Buerkle & Gompert, 2013; Chattopadhyay, Garg, & Ramakrishnan, 2014; DaCosta & Sorenson, 2014; Huang & Knowles, 2014; Rubin, Ree, & Moreau, 2012; Wagner et al., 2013), but this has been argued against in other studies (Henning et al., 2014). Initially, we opted for a conservative approach to minimize missing data and thereby reduce uncertainty in population analyses. However, we 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

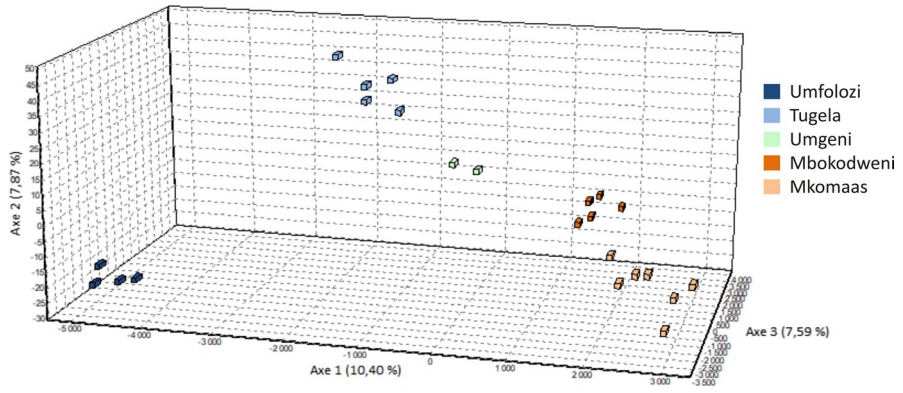

FIGURE 7 Factorial correspondence analysis plot showing the clustering of individuals into the five populations according to the first three components, accounting for 25.86% of the variance in the data

TABLE 4 Summary data produced by the POPULATIONS program in STACKS for variant positions (top) and all positions (bottom)

	Pop	*N*	Pvt	Sites	% Poly sites	*P*	H_O	H_E	F_{IS}	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

Populations are as follows: 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.

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.

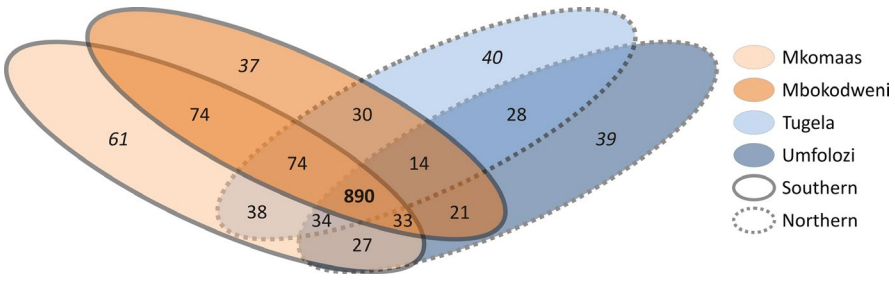

FIGURE 8 Venn diagram illustrating the association of alleles between four populations (Mkomaas, Mbokodweni, Tugela, and Umfolozi) after samples from the potentially admixed Umgeni population are removed, resulting in 720 diallelic loci. Values in italics indicate private alleles for each of the four populations, whereas the value in bold indicates alleles found in all populations

TABLE 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

*$p < .01$.

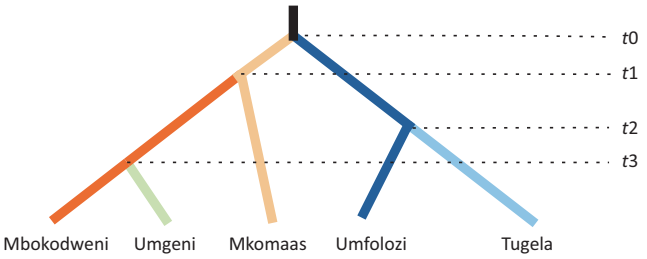

FIGURE 9 Scenario determined as most likely by DIYABC scenario testing using the logistic regression approach. Divergence times are indicated on the right. Divergence points are not drawn to scale. The node *t*0 was fixed in time as the oldest point, whereas all other nodes were allowed to vary in relation to one another

the signal of phylogeographic structure. This phenomenon has been briefly described in the literature (Campagna et al., 2015; Huang & Knowles, 2014; Takahashi, Nagata, & Sota, 2014; Wagner et al., 2013) and should be viewed as a motivation to include more missing data to reduce potential biases from only examining highly conserved regions. However, our final datasets retained relatively low levels of missing data—19.7% within our 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 we observe are 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.

We 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 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, we 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 MAF set to remove singleton alleles.

Cyprinid genomes include extensive repetitive regions (Henkel et al., 2012) such as the *Hin*dIII satellite (Datta, Dutta, & Mandal, 1988; Tseng, Chiang, & Wang, 2008), which is prevalent in our 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 our data exceeds that of all satellites across the *C. carpio* genome (2.46%) (Xu et al., 2011). A *Hin*dIII 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 neighboring populations through molecular biological processes such as biased gene conversion (Elder & Turner, 1994). Similar results were recorded in *A. paradoxus* (Tseng et al., 2008) for the satellite sequence to which our *Labeobarbus* satellite matches. In agreement with this, we analyzed a set of *Hin*dIII satellite SNPs independently to our neutral SNPs and found similar results of genetic structure between the populations identified in this study (data not shown). We similarly tested whether the 24 loci potentially under selection identified through BAYESCAN influenced our STRUCTURE results, but found that we obtained the same results whether we excluded these loci or not. As a result, both satellite loci and loci potentially under selection were retained in our final dataset.

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-recognized 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 (Alexander, Harrison, Fairbanks, & Navarro, 2004; Perera et al., 2011; Seymour, De Klerk, Channing, & Crowe, 2001). 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, Peer, & Perissinotto, 2015) and in vertebrate fauna overall (Perera et al., 2011).

The division into two broad lineages dominated our 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 neighboring Tugela and Mbokodweni populations in every spatial analysis performed (Figures 6 and 7) which again suggests admixture.

The single sample from the upper Umgeni system (Lions River) consistently grouped closely with those from the neighboring Tugela system (Figures 3–7). 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. We 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 neighboring Mkomaas River samples as a single lineage, despite grouping as a distinct mtDNA lineage (Bloomer et al. Unpublished data). 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 (Figures 5 and 6) results.

We were not able to include samples from the southernmost limit of distribution, the Mtamvuna River. Similarly, the Mbokodweni River has two neighboring river systems of similar size, the Umlazi and Illovo Rivers which were unsampled, and may contribute to the unexpectedly high diversity found within this population. 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, Amatigulu, 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 our 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. Unpublished data). 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.

4.4 | Population history

Similar levels of support were received for three models in our scenario testing with DIYABC. The model with the most support matches the results previously observed in our 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 we cannot rule out the possibility of an Mkomaas ancestral group (Figure S7).

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 we 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.

5 | CONCLUSION

In this study, we used the ddRAD sequencing approach to reduce the genome complexity of a South African endemic hexaploid fish. Our SNP identification protocol was optimized 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 allele SNPs which masked genetic structure, and retaining satellite loci and loci potentially under selection which both showed similar results to putatively neutral loci. We demonstrated that although a moderate level of missing data was observed, it was due to locus dropout caused by lineage-specific mutations in one of the two restriction sites. We used our final dataset of 723 SNPs to characterize that two major lineages—northern and southern—diverge into five regional populations—Umfolozi, Tugela, Umgeni, Mbokodweni, and Mkomaas—across the distribution. We 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 our 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 we used together with the resources established in this study will aid in combating the dearth of genetic data available for *Labeobarbus* and other cyprinids.

ACKNOWLEDGMENTS

This study is dedicated to the memory of our colleague, the late Rob Karssing. We 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. We would also like to thank Vincent Savolainen, Bengt Hansson, Ulrich Schliewen, Emmanuel Vreven, Alexander Papadopulos, Ilkser Erdem Kiper, and their research groups for their contributions to this project. Thank you also to 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.

DATA ACCESSIBILITY

Raw sequence data, processed input files for the final dataset (STRUCTURE, GENEPOP, FINERADSTRUCTURE, ARLEQUIN, GENETIX, DIYABC, GESTE/BAYESCAN, and VCF), and regular expressions for removing adapter pollution are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.g00b7.

AUTHOR CONTRIBUTIONS

P.B. and C.J.O. designed the research project. P.B., C.J.O., and M.J.C. supervised the research progress. C.S.S. performed research, developed custom scripts, analyzed and interpreted the data, and led the writing of the manuscript. M.J.C. assisted with many of the technical aspects of the project. All authors contributed to the final draft of the manuscript.

ORCID

Paulette Bloomer http://orcid.org/0000-0002-6357-0153
Cora Sabriel Stobie https://orcid.org/0000-0003-0742-1476

REFERENCES

Alexander, G., Harrison, J., Fairbanks, D., & Navarro, R. (2004). Biogeography of the frogs of South Africa, Lesotho and Swaziland. In L. R. Minter, M. Burger, J. Harrison, H. H. Braack, & P. J. Bishop (Eds.), *Atlas and red data book of the frogs of South Africa, Lesotho and Swaziland. SI/MAB series #9* (pp. 31–47). Washington, DC: Smithsonian Institution.

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

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

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

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

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, I. R., van der Bank, F. H., Villet, M. H., Jones, N., & Walsh, G. (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.

Boehm, J., Waldman, J., Robinson, J. D., & Hickerson, M. J. (2015). Population genomics reveals seahorses (*Hippocampus erectus*) of the western mid-Atlantic Coast to be residents rather than vagrants. *PLoS One*, *10*, e0116219. https://doi.org/10.1371/journal.pone.0116219

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., … Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, *513*, 375–381. https://doi.org/10.1038/nature13726

Brieuc, M. S. O., Waters, C. D., Seeb, J. E., & Naish, K. A. (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. https://doi.org/10.1534/g3.113.009316

Brumfield, R. T., Beerli, P., Nickerson, D. A., & Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, *18*, 249–256. https://doi.org/10.1016/S0169-5347(03)00018-1

Buerkle, C. A., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: How low should we go? *Molecular Ecology*, *22*, 3028–3035. https://doi.org/10.1111/mec.12105

Cambray, J., Bills, R., Chakona, A., Coetzer, W., & Weyl, O. (2017). *Labeobarbus natalensis*. *The IUCN red list of threatened species 2017: e.T63294A100168851*. Retrieved from http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T63294A100168851.en

Campagna, L., Gronau, I., Silveira, L. F., Siepel, A., & Lovette, I. J. (2015). Distinguishing noise from signal in patterns of genomic divergence in a highly polymorphic avian radiation. *Molecular Ecology*, *24*, 4238–4251. https://doi.org/10.1111/mec.13314

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

Catchen, J. M., Hohenlohe, P., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*, 3124–3140. https://doi.org/10.1111/mec.12354

Chakona, A., Malherbe, W. S., Gouws, G., & Swartz, E. R. (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. https://doi.org/10.1080/15627020.2015.1021164

Chakona, A., & Skelton, P. H. (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. https://doi.org/10.3897/zookeys.657.11076

Chakona, A., Swartz, E. R., & 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. https://doi.org/10.1371/journal.pone.0070953

Chattopadhyay, B., Garg, K. M., & Ramakrishnan, U. (2014). Effect of diversity and missing data on genetic assignment with RAD-Seq markers. *BMC Research Notes*, *7*, 841. https://doi.org/10.1186/1756-0500-7-841

Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., … Estoup, A. (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. https://doi.org/10.1093/bioinformatics/btt763

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

Datta, U., Dutta, P., & Mandal, R. K. (1988). Cloning and characterization of a highly repetitive fish nucleotide sequence. *Gene*, *62*, 331–336. https://doi.org/10.1016/0378-1119(88)90570-7

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

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

de Graaf, M., Nagelkerke, L. A. J., Palstra, A. P., & Sibbing, F. A. (2010). Experimental evidence for the biological species status in Lake Tana's *Labeobarbus* flock (Cyprinidae). *Animal Biology*, *60*, 183–193. https://doi.org/10.1163/157075610X491725

Durand, J. D., Tsigenopoulos, C. S., 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. https://doi.org/10.1006/mpev.2001.1040

Earl, D. A., & VonHoldt, B. M. (2011). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetic Resources*, *4*, 359–361. https://doi.org/10.1007/s12686-011-9548-7

Elder, J. F., & Turner, B. J. (1994). Concerted evolution at the population level: Pupfish *Hin*dIII satellite DNA sequences. *Proceedings of the National Academy of Science of the United States of America*, *91*, 994–998. https://doi.org/10.1073/pnas.91.3.994

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. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Everett, M. V., Grau, E. D., & Seeb, J. E. (2011). Short reads and non-model 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. https://doi.org/10.1111/j.1755-0998.2010.02969.x

Excoffier, L., & Lischer, H. E. L. (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. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Ferchaud, A. L., & Hansen, M. M. (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. https://doi.org/10.1111/mec.13399

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. https://doi.org/10.1534/genetics.108.092221

Gagnaire, P.-A., Normandeau, E., Pavey, S. A., & 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. https://doi.org/10.1111/mec.12127

Gonen, S., Bishop, S. C., & Houston, R. D. (2015). Exploring the utility of cross-laboratory RAD-sequencing datasets for phylogenetic analysis. *BMC Research Notes*, *8*, 299. https://doi.org/10.1186/s13104-015-1261-2

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. https://doi.org/10.4102/koedoe.v57i1.1324

Hand, B. K., Hether, T. D., Kovach, R. P., Muhlfeld, C. C., Amish, S. J., Boyer, M. C., … Luikart, G. (2015). Genomics and introgression: Discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Current Zoology*, *61*, 146–154. https://doi.org/10.1093/czoolo/61.1.146

Henkel, C. V., Dirks, R. P., Jansen, H. J., Forlenza, M., Wiegertjes, G. F., Howe, K., … Spaink, H. P. (2012). Comparison of the exomes of common carp (*Cyprinus carpio*) and zebrafish (*Danio rerio*). *Zebrafish*, *9*, 59–67. https://doi.org/10.1089/zeb.2012.0773

Henning, F., Lee, H. J., 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. https://doi.org/10.1111/mec.12860

Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W., & 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. https://doi.org/10.1111/j.1755-0998.2010.02967.x

Hohenlohe, P. A., Day, M. D., Amish, S. J., Miller, M. R., Kamps-Hughes, N., Boyer, M. C., … Luikart, G. (2013). Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*, *22*, 3002–3013. https://doi.org/10.1111/mec.12239

Houston, R. D., Davey, J. W., Bishop, S. C., Lowe, N. R., Mota-Velasco, J. C., Hamilton, A., … Taggart, J. B. (2012). Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. *BMC Genomics*, *13*, 244. https://doi.org/10.1186/1471-2164-13-244

Huang, H., & Knowles, L. L. (2014). Unforeseen consequences of excluding missing data from Next-Generation Sequences: Simulation study of RAD sequences. *Systematic Biology*, *65*, 357–365. https://doi.org/10.1093/sysbio/syu046

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*, 1801–1806. https://doi.org/10.1093/bioinformatics/btm233

Kai, W., Nomura, K., Fujiwara, A., Nakamura, Y., Yasuike, M., Ojima, N., … Ototake, M. (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. https://doi.org/10.1186/1471-2164-15-233

Karssing, R. J. (2008). Status of the KwaZulu-Natal yellowfish *Labeobarbus natalensis* (Castelnau, 1861). In N. D. Impson, I. R. Bills, & L. Wolhuter (Eds.), *Technical report on the state of yellowfishes in South Africa 2007* (pp. 31–45). Pretoria, South Africa: Water Research Commission.

Lamer, J. T., Sass, G. G., Boone, J. Q., Arbieva, Z. H., Green, S. J., & Epifanio, J. M. (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. https://doi.org/10.1111/1755-0998.12152

Larson, W. A., Seeb, L. W., Everett, M. V., Waples, R. K., Templin, W. D., & Seeb, J. E. (2014). Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications*, *7*, 355–369. https://doi.org/10.1111/eva.12128

Lischer, H., & Excoffier, L. (2012). PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*, 298–299. https://doi.org/10.1093/bioinformatics/btr642

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

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. https://doi.org/10.1006/mpev.2000.0876

Malinsky, M., Trucchi, E., Lawson, D., & Falush, D. (2016). RADpainter and fineRADstructure: Population inference from RADseq data. *bioRxiv*, 057711. https://doi.org/10.1101/057711

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology Resources*, *15*, 28–41. https://doi.org/10.1111/1755-0998.12291

McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (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. https://doi.org/10.1111/1755-0998.12613

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

Naran, D., Skelton, P. H., & Villet, M. H. (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. https://doi.org/10.3377/1562-7020(2006)41[178:kotrmg]2.0.co;2

Nel, J., Murray, K., Maherry, A., Petersen, C. P., Roux, D. J., Driver, A., … Smith-Adao, L. B. (2011). *Technical report for the national freshwater ecosystem priority areas project*. WRC Report 1802/2/11.

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

Ogden, R., Gharbi, K., Mugue, N., Martinsohn, J., Senn, H., Davey, J. W., … Congiu, L. (2013). Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Molecular Ecology*, *22*, 3112–3123. https://doi.org/10.1111/mec.12234

Palti, Y., Gao, G., Miller, M. R., Vallejo, R. L., Wheeler, P. A., Quillet, E., … Rexroad, C. E. 3rd (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. https://doi.org/10.1111/1755-0998.12204

Palumbi, S. R. (2003). Population genetics, demographic connectivity, and the design of marine reserves. *Ecological Applications*, *13*, 146–158. https://doi.org/10.1890/1051-0761(2003)013[0146:PGDCAT]2.0.CO;2

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

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. https://doi.org/10.1007/s00438-009-0504-0

Partridge, T., & Maud, R. (2000). Macro-scale geomorphic evolution of southern Africa. *Oxford Monographs on Geology and Geophysics*, *40*, 3–18.

Perera, S. J., 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. https://doi.org/10.4102/sajs.v107i7/8.462

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

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

Pujolar, J. M., Jacobsen, M., Als, T. D., Frydenberg, J., Magnussen, E., Jónsson, B., … Hansen, M. M. (2014). Assessing patterns of hybridization between North Atlantic eels using diagnostic single-nucleotide polymorphisms. *Heredity*, *112*, 627–637. https://doi.org/10.1038/hdy.2013.145

Pujolar, J. M., Jacobsen, M. W., Frydenberg, J., Als, T. D., Larsen, P. F., Maes, G. E., … Hansen, M. M. (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. https://doi.org/10.1111/1755-0998.12117

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

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. https://doi.org/10.1080/00087114.1989.10796950

Rašić, G., Filipović, I., Weeks, A. R., & Hoffmann, A. A. (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. https://doi.org/10.1186/1471-2164-15-275

Reitzel, A. M., Herrera, S., Layden, M. J., Martindale, M. Q., & Shank, T. M. (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. https://doi.org/10.1111/mec.12228

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, I. R., Mendibil, I., Álvarez, P., Cotano, U., & Irigoien, X. (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. https://doi.org/10.1111/1755-0998.12518

Roesti, M., Salzburger, W., & Berner, D. (2012). Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology*, *12*, 94. https://doi.org/10.1186/1471-2148-12-94

Rubin, B. E. R., Ree, R. H., & Moreau, C. S. (2012). Inferring phylogenies from RAD sequence data. *PLoS One*, *7*, e33394. https://doi.org/10.1371/journal.pone.0033394

Sambrook, J., Fritsch, E., & Maniatis, T. (1989). Extraction with phenol: Chloroform. *Molecular Cloning: A Laboratory Manual*. Second Edition. Cold spring harbor laboratory press.

Seymour, C. L., De Klerk, H. M., Channing, A., & Crowe, T. M. (2001). The biogeography of the Anura of sub-equatorial Africa and the prioritisation of areas for their conservation. *Biodiversity and Conservation*, *10*, 2045–2076. https://doi.org/10.1023/A:1013137409896

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

Skelton, P., & Bills, R. (2008). An introduction to African yellowfish and to this report. In N. D. Impson, I. R. Bills, & L. Wolhuter (Eds.), *Technical report on the state of yellowfishes in South Africa 2007* (pp. 1–14). Pretoria, South Africa: Water Research Commission.

Smith, S. A., & Bermingham, E. (2005). The biogeography of lower Mesoamerican freshwater fishes. *Journal of Biogeography*, *32*, 1835–1854. https://doi.org/10.1111/j.1365-2699.2005.01317.x

Soltis, P. S., & Soltis, D. E. (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. https://doi.org/10.1073/pnas.97.13.7051

Swartz, E. R., Chakona, A., Skelton, P. H., & 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. https://doi.org/10.1007/s10750-013-1755-7

Swartz, E. R., 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, E. R., 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. https://doi.org/10.1111/j.1365-2699.2007.01768.x

Swartz, E. R., Skelton, P. H., & 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. https://doi.org/10.1016/j.ympev.2008.10.017

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. https://doi.org/10.1016/j.ympev.2014.07.016

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. https://doi.org/10.1111/j.1095-8649.2008.01749.x

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

van der Walt, K.-A., Swartz, E. R., 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. https://doi.org/10.4102/koedoe.v59i1.1375

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

Wagner, C. E., Keller, I., Wittwer, S., Selz, O. M., Mwaiko, S., Greuter, L., … Seehausen, O. (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. https://doi.org/10.1111/mec.12023

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. https://doi.org/10.1111/1755-0998.12394

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. https://doi.org/10.1371/journal.pone.0042649

Xu, P., Li, J., Li, Y., Cui, R., Wang, J., Wang, J., … Sun, X. (2011). Genomic insight into the common carp (*Cyprinus carpio*) genome by sequencing analysis of BAC-end sequences. *BMC Genomics*, *12*, 188. https://doi.org/10.1186/1471-2164-12-188

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

Yoshizawa, M., Robinson, B. G., Duboué, E. R., Masek, P., Jaggard, J. B., O'Quin, K. E., … Keene, A. C. (2015). Distinct genetic architecture underlies the emergence of sleep loss and prey-seeking behavior in the Mexican cavefish. *BMC Biology*, *13*, 1. https://doi.org/10.1186/s12915-015-0119-3

Zhang, B.-D., Xue, D.-X., Wang, J., Li, Y. L., Liu, B. J., & Liu, J. X. (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. https://doi.org/10.1111/1755-0998.12476

Zhou, Y., Bizzaro, J. W., & Marx, K. A. (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. https://doi.org/10.1186/1471-2164-5-95

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Stobie CS, Oosthuizen CJ, Cunningham MJ, Bloomer P. Exploring the phylogeography of a hexaploid freshwater fish by RAD sequencing. *Ecol Evol*. 2018;8:2326–2342. https://doi.org/10.1002/ece3.3821