Received: 6 December 2021 | Revised: 22 July 2022 | Accepted: 26 July 2022

DOI: 10.1111/mec.16636

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Variable hybridization between two Lake Tanganyikan cichlid species in recent secondary contact

Alexander L. Lewanski[1] | Jimena Golcher-Benavides[1,2] | Jessica A. Rick[1,2] | Catherine E. Wagner[1,2,3]

[1]Department of Botany, University of Wyoming, Laramie, Wyoming, USA

[2]Program in Ecology, University of Wyoming, Laramie, Wyoming, USA

[3]Biodiversity Institute, University of Wyoming, Laramie, Wyoming, USA

Correspondence
Alexander L. Lewanski and Catherine E. Wagner, Department of Botany, University of Wyoming, Laramie, WY, USA.
Email: allewanski@gmail.com and catherine.wagner@uwyo.edu

Funding information
National Science Foundation, Grant/Award Number: DEB-1556963

Handling Editor: Richard J Abbott

Abstract

Closely related taxa frequently exist in sympatry before the evolution of robust reproductive barriers, which can lead to substantial gene flow. Post-divergence gene flow can promote several disparate trajectories of divergence ranging from the erosion of distinctiveness and eventual collapse of the taxa to the strengthening of reproductive isolation. Among many relevant factors, understanding the demographic history of divergence (e.g. divergence time and extent of historical gene flow) can be particularly informative when examining contemporary gene flow between closely related taxa because this history can influence gene flow's prevalence and consequences. Here, we used genotyping-by-sequencing data to investigate speciation and contemporary hybridization in two closely related and sympatrically distributed Lake Tanganyikan cichlid species in the genus *Petrochromis*. Demographic modelling supported a speciation scenario involving divergence in isolation followed by secondary contact with bidirectional gene flow. Further investigation of this recent gene flow found evidence of ongoing hybridization between the species that varied in extent between different co-occurring populations. Relationships between abundance and the degree of admixture across populations suggest that the availability of conspecific mates may influence patterns of hybridization. These results, together with the observation that sets of recently diverged cichlid taxa are generally geographically separated in the lake, suggest that ongoing speciation in Lake Tanganyikan cichlids relies on initial spatial isolation. Additionally, the spatial heterogeneity of admixture between the *Petrochromis* species illustrates the complexities of hybridization when species are in recent secondary contact.

KEYWORDS
adaptive radiation, admixture, cichlids, hybridization, Lake Tanganyika, speciation

1 | INTRODUCTION

Speciation in sexually reproducing organisms, unless reproductive isolation is instantaneous (e.g. via polyploid hybrid speciation; Köhler et al., 2010), is generally thought to involve the incremental accumulation of gene flow barriers that strengthen reproductive isolation until it is absolute (although this terminus is not always achieved; Nosil et al., 2009). However, recently divergent populations frequently exist in sympatry before the evolution of complete reproductive barriers, which can result in substantial gene flow postdating the onset of divergence. Post-divergence gene flow can represent an evolutionarily consequential interaction for the two taxa because it can promote several disparate trajectories of divergence, ranging from species collapse to the further bolstering of isolation (e.g. reinforcement) (Abbott et al., 2013).

With progressively larger and higher resolution genomic data sets, we are increasingly documenting the complexities of reproductive isolation between sympatric, closely related species. For example, evidence is mounting that the extent of hybridization can frequently vary between different co-occurring populations of the same sets of species. This spatial variability is particularly well documented in catastomid fishes in the western United States, where co-occurring populations of multiple species display markedly different hybridization outcomes (Mandeville et al., 2015, 2017). The heterogeneity of reproductive isolation also extends into the genome where intrinsic hybrid incompatibilities can be polymorphic between species (e.g. Good et al., 2008; Larson et al., 2018; reviewed in Cutter, 2012).

Among an array of relevant considerations, the demographic history of the interacting taxa can be crucial for understanding contemporary patterns, prevalence, and consequences of gene flow. Most obviously, the timing of divergence between the taxa may influence the degree of contemporary gene flow since, all else being equal, we expect reproductive isolation to increase with deeper divergence times (Coyne & Orr, 1989, 1997; Matute & Cooper, 2021). The history of gene flow during the divergence process can also influence the prevalence and impact of contemporary gene flow. Divergence can proceed in the continuous presence of gene flow, which is generally thought to occur when speciation is driven by disruptive or divergent ecological selection (Bolnick & Fitzpatrick, 2007), although other factors such as the frequency of recombination between genes under selection can also play a critical role (Pinho & Hey, 2010). In this scenario, barring a shift in the selective regime prompting divergence (Cutter & Gray, 2016), ongoing gene flow may not substantially counteract further divergence or promote collapse since existing reproductive barriers emerged despite gene flow. In contrast, in closely related taxa whose initial divergence occurred without gene flow, divergence did not require the evolution of reproductive barriers and thus gene flow may counteract divergence. Hence, closely related taxa in recent sympatry whose divergence commenced in isolation may be especially vulnerable to collapse if sympatry predated the evolution of strong reproductive barriers.

The cichlid fishes of East Africa, and especially those in the great lakes of Victoria, Tanganyika and Malawi, represent an exciting group for studying the causes and consequences of post-divergence gene flow, including how gene flow affects the speciation process and broader patterns of contemporary diversity. In each of these lakes, cichlids have independently evolved into spectacular species flocks (Kocher, 2004; Kornfield & Smith, 2000) during which they manifested some of the fastest known vertebrate diversification rates (McCune, 1997). Post-divergence gene flow has also featured prominently in the history of these cichlid radiations and has substantially shaped their evolution (Meier, Marques, et al., 2017; Malinsky et al., 2018; Irisarri et al., 2018; Svardal et al., 2020; Ronco et al., 2021; reviewed in Svardal et al., 2021).

Although the East African cichlid radiations are replete with recently diverged forms, intriguing discrepancies exist between the radiations regarding the spatial distribution of this incipient diversity. In Lake Tanganyika, closely related colour variants are rarely sympatric and instead replace each other along disjunct expanses of rocky shoreline (Kohda et al., 1996; a particularly well-documented pattern in *Tropheus*, for example, Egger et al., 2007; Sefc, Mattersdorfer, Ziegelbecker, et al., 2017). In contrast, many incipient species in Lake Victoria exist in sympatry (Seehausen & van Alphen, 1999), and allopatric sister taxa are comparatively rare (Seehausen & Magalhaes, 2010). Sympatric diversity in Lake Victoria is at least partly the direct product of speciation with species often forming without geographic isolation and with ample opportunity for gene flow such as along depth gradients (Seehausen & Magalhaes, 2010; Seehausen et al., 2008).

Given the predominantly allopatric distributions of closely related Lake Tanganyika cichlids, it remains unclear how sympatric diversity develops in this radiation. Sympatry may simply represent instances of secondary contact following allopatric divergence. This scenario is plausible given the putative effects of periodic water level fluctuations on the dynamics of divergence and sympatry in Lake Tanganyika cichlids. Specifically, water level fluctuations, which are induced by climatic and geological events (Cohen et al., 1997; McGlue et al., 2008; Scholz et al., 2003), are thought to alter the distribution and extent of rocky habitat including the positions of dispersal barriers. The habitat changes displace associated cichlid populations, which can bring allopatric taxa into secondary contact and enable post-divergence gene flow while enforcing isolation between formerly connected populations and therefore facilitating their divergence (Koblmüller et al., 2008; Sefc, Mattersdorfer, Ziegelbecker, et al., 2017; Sturmbauer, 1998).

Alternatively, one could envision scenarios where allopatric and sympatric cichlid taxa in Lake Tanganyika form via distinct processes. For example, taxa that diverged in spatial isolation without gene flow could fail to evolve strong reproductive barriers before secondary contact and thus tend to fuse in sympatry (considered likely in cichlids by Seehausen, 2015). Taxa that successfully coexist in sympatry could instead have diverged in a process more akin to speciation documented in other cichlid radiations like Lake Victoria, where species can arise without spatial isolation and with continuous opportunity for gene flow. Unfortunately, the existing literature provides limited information about the ongoing formation of sympatric diversity in Lake Tanganyika cichlids. Most work on closely related taxa has either focused on divergence in allopatry (e.g. Koblmüller et al., 2011; Sefc, Mattersdorfer, Ziegelbecker, et al., 2017; Taylor et al., 2001; Wagner & McCune, 2009) or cases of historical interspecific gene flow from past contact (e.g. Egger et al., 2007; Gante et al., 2016; Sefc, Mattersdorfer, Hermann, & Koblmüller, 2017). Details regarding the formation of sympatric diversity in Lake Tanganyika cichlids remain inadequately explored, and clarifying the patterns and extent of gene flow among sympatric taxa along with the temporal context of these interactions may have implications for understanding how and when closely related taxa persist in sympatry.

Our work focuses on *Petrochromis* sp. 'kazumbe' and *Petrochromis* cf. *polyodon*, two tropheine cichlid taxa embedded in the Lake Tanganyika radiation. Similar to many other members of the genus *Petrochromis*, *P.* sp. 'kazumbe' and *P.* cf. *polyodon* are polygamous maternal mouthbrooders that stenotopically occupy rocky shoreline habitats and are specialized grazers of epilithic algae (Wagner et al. (2012) gives a more thorough overview of these taxa). The two taxa are phenotypically similar, and adults are visually distinguishable only by differences in orange colouration with *P.* sp. 'kazumbe' displaying more extensive orange than *P.* cf. *polyodon* (Figure 1c). Notably, these species co-occur in the Kigoma region of Tanzania, and they are found sympatrically together in suitable habitats without obvious habitat differentiation. Beyond the Kigoma region, the ranges of these species and where they co-occur remain unclear. Surveys in Lake Tanganyika south of the Kigoma region detect a *Petrochromis* taxon that is genetically similar to *P.* sp. 'kazumbe' but have not found *P.* cf. *polyodon* (Golcher-Benavides, 2021). A taxon resembling the

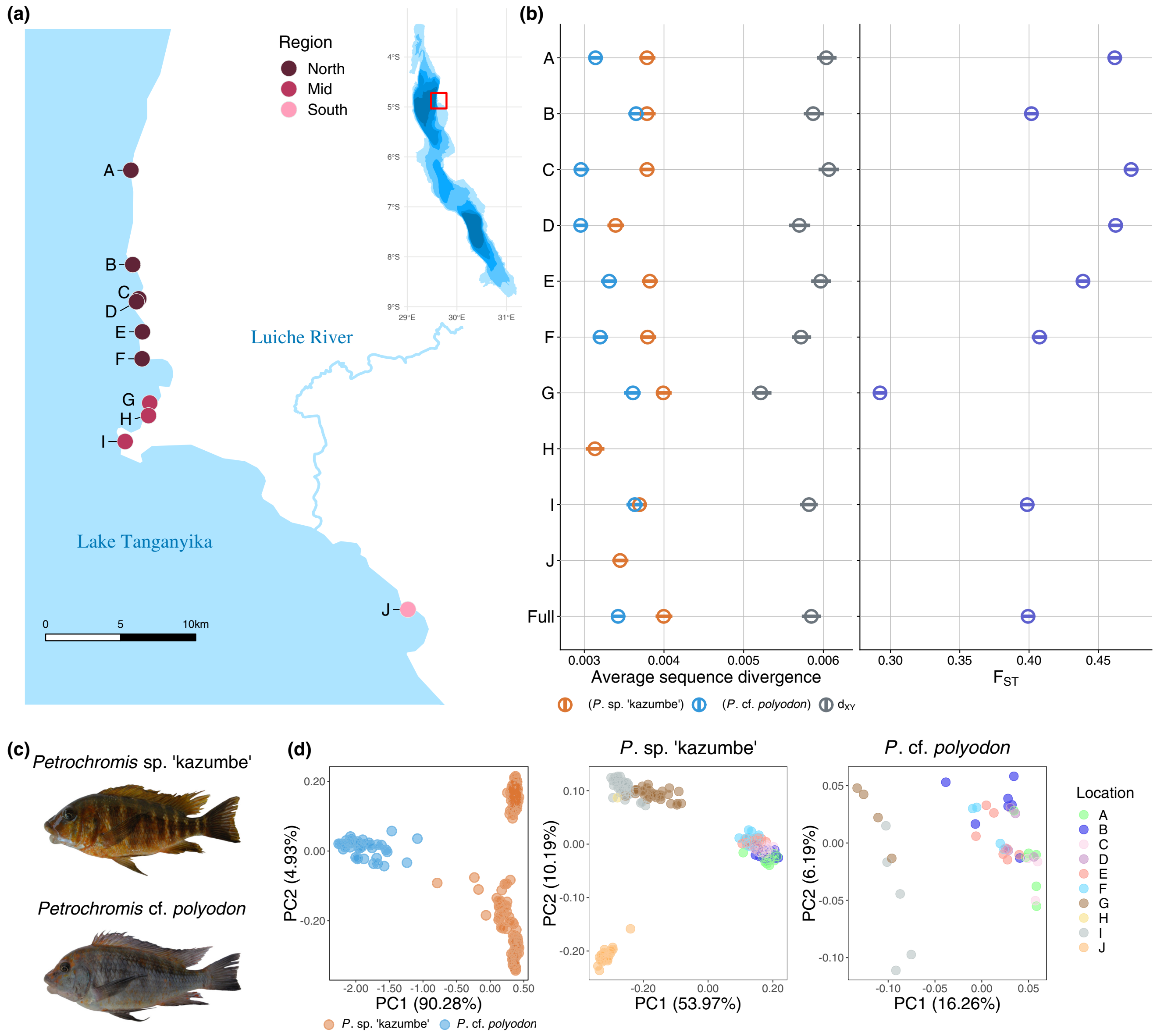

FIGURE 1 Overview of study system. (a) Map of the Lake Tanganyika study area, which is located in the Kigoma region of Tanzania. Points indicate sampling locations with colour denoting sampling region (dark red = North, medium red = Mid, pink = South). (b) Estimates of average genome-wide differentiation (F_{ST}), divergence (d_{XY}) and nucleotide diversity (π; calculated separately by species). Calculations were performed at each location with at least three individuals (i.e. three samples for a single species for π and three samples for each species for F_{ST} and d_{XY}) and also using all samples in the data set (*Full*). The hollow circle represents the point estimate of the statistic, and the horizontal bar represents the 95% confidence interval based on 500 bootstrap replicates. (c) Images of the study species: *Petrochromis* sp. 'kazumbe' (top) and *Petrochromis* cf. *polyodon* (bottom). (d) Plots showing the first two axes of principal component analyses (PCAs) performed on the following data sets: samples of both species (left), *P.* sp. 'kazumbe' (middle), and *P.* cf. *polyodon* (right). The points of the full data set PCA are coloured by species, and the points of the species-specific PCAs are coloured by sampling location [Colour figure can be viewed at wileyonlinelibrary.com]

Petrochromis studied here also exists north of the Kigoma region (Konings, 2015). However, since *P.* sp. 'kazumbe' and *P.* cf. *polyodon* have not been distinguished in other work (e.g. Konings, 2015), it is unclear which taxon or taxa occur in the north. We also note that we here use the names by which we have referred to these taxa in previous work, which differ from those used in some aquarium trade and scientific literature. Since the equivalence of the taxa we study here with those described elsewhere has not been demonstrated, we use our original nomenclature for consistency within our own detailed work on these two taxa (e.g. Wagner et al., 2012).

Previous work by Wagner et al. (2012) established *P.* sp. 'kazumbe' and *P.* cf. *polyodon* as a compelling case study of speciation in Lake Tanganyika. Wagner et al. (2012) documented considerable sharing of mitochondrial haplotypes between *P.* sp. 'kazumbe' and *P.* cf. *polyodon* but also clear differentiation based on nuclear microsatellites. The magnitude of differentiation was lower than corresponding comparisons between other *Petrochromis* taxa and equivalent to intraspecific differentiation in *P.* sp. 'kazumbe' resulting from spatial isolation. The findings confirmed the distinctiveness of *P.* sp. 'kazumbe' and *P.* cf. *polyodon* and signified the existence of barriers preventing extensive gene flow since they remain differentiated in sympatry. However, the findings also suggested that divergence occurred only recently.

In this study, we further probe the speciation history of *P.* sp. 'kazumbe' and *P.* cf. *polyodon* by applying genotyping-by-sequencing (GBS) data to describe their divergence history and contemporary patterns of interspecific gene flow. In particular, we addressed the following three questions. First, what was the timing of divergence between *P.* sp. 'kazumbe' and *P.* cf. *polyodon*? Second, how did interspecific gene flow figure into the divergence process? Specifically, we investigate whether post-divergence gene flow has occurred between the two species and characterize its amount, timing and directionality. Third, is contemporary hybridization occurring between the species? For the final question, we explore whether the species are engaging in ongoing hybridization, the prevalence of this hybridization, and to what extent hybridization varies across different co-occurring populations. *P.* sp. 'kazumbe' and *P.* cf. *polyodon* represent an unusual instance in the Lake Tanganyika cichlid species flock of naturally sympatric, closely related colour variants that appear to stably coexist in sympatry. Thus, they serve as a valuable case study for dissecting how sympatric diversity arises and interacts as well as how gene flow figures into the speciation process in this radiation.

2 | MATERIALS AND METHODS

We conducted high-performance computing on the University of Wyoming's Teton computing cluster (Advanced Research Computing Center, 2018), and we used `R` v4.1 (R Core Team, 2021) for statistical analyses and visualization.

2.1 | Sample collection

We sampled *P.* sp. 'kazumbe' and *P.* cf. *polyodon* in the Kigoma region of Tanzania in 2005, 2007, 2016, 2017 and 2018 across 10 locations covering ~60 km of shoreline. The sampling spans the Luiche River delta (between locations I and J) and Kigoma Bay (between locations F and G), which represent barriers to gene flow for rock-dwelling cichlids like *Petrochromis* (Wagner et al., 2012; Wagner & McCune, 2009) and are used here to demarcate the north, mid and south sampling regions (Figure 1a). The samples from 2005 and 2007 were previously analysed by Wagner et al. (2012) using microsatellite data, and all sampling followed the methods described therein. Briefly, we collected individuals while snorkelling in the rocky littoral zone at a 1–10 m depth. We took a fin clip from each individual (stored in either DMSO-EDTA or 95% ethanol) for genetic analyses. In total, we collected 245 *P.* sp. 'kazumbe' across 10 locations and 42 *P.* cf. *polyodon* across eight locations. From a subset of the locations used in the *P.* cf. *polyodon* and *P.* sp. 'kazumbe' sampling, we collected *Simochromis diagramma* ($n = 9$) and *Petrochromis* cf. *macrognathus* 'green' ($n = 6$). These two species are closely related to *P.* cf. *polyodon* and *P.* sp. 'kazumbe' (Ronco et al., 2021; Wagner et al., 2012) and were used as outgroups in analyses. We retained all individuals as vouchers, which are deposited at the Cornell University Museum of Vertebrates and the University of Wyoming Museum of Vertebrates.

2.2 | Genomic library preparation, sequencing and bioinformatics

We extracted DNA from the fin clips using DNeasy Blood & Tissue kits (Qiagen, Inc.). We prepared genomic libraries for high-throughput DNA sequencing following the GBS protocol described in Parchman et al. (2012). Briefly, we fragmented DNA using EcoRI and MseI restriction enzymes and then ligated a unique 8–10 base pair (bp) barcode to each individual's DNA. We used polymerase chain reaction (PCR) to amplify the restriction/ligation products (two independent replicates per individual) and then combined the PCR products to create the final libraries. Prior to sequencing, the libraries were size-selected for 250–400 bp fragments using BluePippin (Sage Science). Sequencing was completed on Illumina HiSeq 2500 and 4000 platforms (100 bp single-end) at the University of Texas Genome Sequencing and Analysis Facility (Austin, Texas, USA) and the University of Oregon Genomics and Cell Characterization Core Facility (Eugene, Oregon, USA). The individuals in this project were included in libraries containing samples of other cichlid species as part of a larger sequencing effort. Each library contained ~100 individuals and was sequenced in its own lane.

With the raw sequence data, we first matched reads to individuals and subsequently removed the barcode sequences using a custom `perl` script. We then aligned the reads to the *Pundamilia nyererei* reference genome (Brawand et al., 2014) using `bwa` v0.7.17 (Li & Durbin, 2009) with default settings.

We created five data sets for downstream analyses. For the two data sets containing only variant sites (*variant* data sets), we first identified single nucleotide variants from the individual alignment (bam) files using samtools v1.8 (Li et al., 2009) and bcftools v1.8 (Li, 2011). We further filtered the variant data with vcftools v0.1.14 (Danecek et al., 2011) to retain biallelic sites meeting the following criteria: quality value >20, 5 ≤mean read depth ≤75, minor allele frequency >0.01 and possessing data for ≥70% of individuals. We completed the variant identification and filtering steps independently for two sets of species: (1) *P.* sp. 'kazumbe' and *P.* cf. *polyodon* only (focaltaxa_variant_missing70_maf1); and (2) *P.* sp. 'kazumbe', *P.* cf. *polyodon*, *P.* cf. *macrognathus* 'green', and *S. diagramma* (alltaxa_variant_missing70_maf1). The data set focaltaxa_variant_missing70_maf1 was used for F_{ST} calculations, visualization of genetic variation with principal components analysis (PCA) and popvae (Battey et al., 2021), and ancestry estimation, while alltaxa_variant_missing70_maf1 was used for calculations of Patterson's D statistics (Green et al., 2010).

We also created three data sets that contained both variant and invariant sites (*allsites* data sets). We first called sites using bcftools for the *P.* sp. 'kazumbe' and *P.* cf. *polyodon* samples. From the raw *allsites* data set, we created three data sets by retaining different sets of samples: all samples (focaltaxa_allsites_missing70_maf0), north region samples (north_focaltaxa_allsites_missing70_maf0) and mid-region samples (mid_focaltaxa_allsites_missing70_maf0). We then applied the following site-level filters to each *allsites* data set: quality value >20, maximum of two alleles, 5 ≤mean read depth ≤75 and possessing data for ≥0.7 of individuals. With all five data sets, we removed samples with >0.7 missing data after filtering. The data set focaltaxa_allsites_missing70_maf0 was used for calculations of π and $d_{XY,}$ and the region-specific *allsites* data sets were used for demographic modelling (Table S2).

2.3 | Exploration of structure, divergence and diversity

We used two approaches to visualize genomic variation among the samples. First, we performed a PCA on the genotype covariance matrix. Second, we used popvae, which applies a pair of deep neural networks to compress and then recreate the data in low-dimensional latent space. By encoding information in just two dimensions, popvae circumvents the information loss inherent in PCA visualizations caused by information being distributed across more dimensions than can be effectively visualized (Battey et al., 2021). We ran popvae using the *search_network_sizes* setting and with a *patience* setting of 500. We conducted the PCA and popvae analyses on three sets of samples: the full data set containing both species to evaluate interspecific divergence, and data sets restricted to each species to explore intraspecific spatial structure.

We calculated several metrics to quantify genomic variation within and between *P.* cf. *polyodon* and *P.* sp. 'kazumbe'. First, we examined differentiation between species using the F_{ST} estimator proposed by Reich et al. (2009), which performs well with both small and unbalanced sample sizes (Willing et al., 2012). We also calculated nucleotide diversity (π) and divergence (d_{XY}) with pixy v0.95.0 (Korunes & Samuk, 2021). pixy avoids biases in estimates of π and d_{XY} by differentiating invariant from missing sites in its calculations. We calculated the three metrics using the full data set and the subset of samples from each sampling location (limited to locations with sampling for both species for F_{ST} and d_{XY}). Because we sampled more *P.* sp. 'kazumbe' individuals than *P.* cf. *polyodon*, differences in sample size could plausibly contribute to interspecific discrepancies in π. We evaluated potential sample size biases by recalculating π using five randomly subsampled data sets for *P.* sp. 'kazumbe' (both for the full data set and at each location with sampling for both species) of the same sample sizes as *P.* cf. *polyodon*. Location J (the south location where only *P.* sp. 'kazumbe' was found) was excluded from the pool of samples used to create the subsamples of the full data set so that the *P.* sp. 'kazumbe' and *P.* cf. *polyodon* samples were collected from similar spatial extents. To further examine intraspecific differentiation, we also calculated F_{ST} between all pairs of sampling locations for each species. For analyses that required assigning samples to species, we confirmed species identity using the K = 2 entropy model (described below). We calculated 95% confidence intervals (CIs) for all statistics using 500 bootstrap replicates.

2.4 | Ancestry estimation

We applied the hierarchical Bayesian models in entropy (Gompert et al., 2014; Shastry et al., 2021) to estimate the ancestry of each individual, which was used to evaluate intra- and interspecific divergence and identify hybridization between species. Entropy can estimate ancestry proportions under two models. The *q*-model, similar to the model in STRUCTURE (Falush et al., 2003), estimates the proportion of an individual's genome derived from K ancestry groups (denoted as q). entropy also implements the ancestry complement model, which considers the ancestry of both allele copies at all loci and can estimate the proportion of sites in the genome with inter-group ancestry (Q_{12}). Notably, entropy considers genotype likelihoods (rather than called genotypes) in its modelling, which propagates sequencing uncertainty into uncertainty in the parameter estimates.

We ran entropy models estimating q for K = 2–5. For each K value, we ran three independent Markov Chain Monte Carlo (MCMC) chains. Each chain included 80,000 steps with the first 30,000 discarded as burn-in and the chain thinned to every tenth step. We evaluated convergence and mixing of the chains by plotting traces of the MCMC iterations for a subset of the samples and parameters (Figure S5). Instead of attempting to identify an optimal K, we considered all K values because different values may reveal distinct information about intra- and interspecific genetic structure.

We also used entropy to estimate both q and Q_{12} (at K = 2) on samples restricted to each sampling region (similar to Mandeville et al., 2017). Since both species displayed clear spatial genetic differentiation structured by sampling region, this approach facilitated more accurate ancestry estimates by considering the local parental allele frequencies in each region. We restricted the region-specific analyses to the north and mid-regions since only *P.* sp. 'kazumbe' was found in the south region.

The bivariate relationship of *q* and Q_{12} enables the designation of individuals into coarse ancestry classes: complete ancestry for one of the parental species, F1 hybrids, F2 hybrids, backcrosses to parental species and later-generation recombinant hybrids. Following Mandeville et al. (2019), we used the estimates of *q* and Q_{12} from `entropy` to sort individuals into these ancestry classes. Several factors can lead to variation in ancestry within classes (e.g. true variation in ancestry, model uncertainty; discussed in Mandeville et al., 2019), which was accommodated in the ancestry classification scheme by allowing for slight deviations away from the expected values of *q* and Q_{12}. The classification scheme is thus conservative at detecting hybrids because individuals whose estimated ancestry slightly deviates from complete ancestry of one of the parental species may still be classified as unadmixed even if the deviations represent true admixture. See Appendix S1 (Section 1.1) for further information on ancestry class designation and running `entropy`.

2.5 | Quantifying variation in allele sharing with D statistics

We calculated a series of Patterson's D statistics (Green et al., 2010) to explore how interspecific allele sharing varied across populations within each species, which may indicate variation in admixture. We included D statistics in our exploration of gene flow as a complement to the `entropy` analyses for two reasons. First, the D statistic is a highly flexible method for detecting post-divergence gene flow because it is generally robust to the timing of gene flow (Hibbins & Hahn, 2021) and the degree of divergence between taxa (Zheng & Janke, 2018). Additionally, the D statistic is sensitive to admixture in some circumstances under which model-based clustering methods may misperform including when there is considerable incomplete lineage sorting or asymmetric parental contributions to ancestry (Kong & Kubatko, 2021).

D statistics involve four taxa with the topology [((P1, P2), P3), Outgroup] and are calculated based on biallelic sites where taxa either possess the ancestral (A) or derived (B) allele. The D statistic focuses on two site patterns that are discordant with the species tree: BABA (P2 is ancestral, while P1 and P3 are derived) and ABBA (P1 is ancestral, while P2 and P3 are derived). In the absence of gene flow between P3 and P1/P2, incomplete lineage sorting should result in equal frequencies of the ABBA and BABA patterns (producing a D statistic of 0). However, if gene flow has occurred between P3 and either P1 or P2, then we expect an excess of one of the patterns and a non-zero D statistic, with gene flow between P1 and P3 generating more BABAs (positive D) and gene flow between P2 and P3 generating more ABBAs (negative D). With gene flow between P3 and both P1 and P2, we would expect equal gene flow to produce balanced allele sharing and a D statistic near zero while unequal gene flow would result in a D statistic that deviates in the direction of the taxa with higher levels of gene flow.

To determine whether the degree of allele sharing between the two species varied across populations (i.e. sampling locations), we calculated two complementary sets of D statistics with the following topologies (Figure 2a): (1a) [((P1: *P.* sp. 'kazumbe'$_{\text{pop. 1}}$, P2: *P.* sp. 'kazumbe'$_{\text{pop. 2}}$), P3: *P.* cf. *polyodon*$_{\text{full}}$), Outgroup] and (2a) [((P1: *P.* cf. *polyodon*$_{\text{pop. 1}}$, P2: *P.* cf. *polyodon*$_{\text{pop. 2}}$), P3: *P.* sp. 'kazumbe'$_{\text{full}}$), Outgroup], where *pop. 1* represents samples from a single population, *pop. 2* represents samples from a separate population, and *full* represents all the samples from the species. Topology 1a tests for variation in allele sharing with *P.* cf. *polyodon* between populations of *P.* sp. 'kazumbe'. That is, it determines whether one of the *P.* sp. 'kazumbe' populations displays greater allele sharing with *P.* cf. *polyodon* than the other *P.* sp. 'kazumbe' population. Conversely, topology 2a tests for variation in allele sharing with *P.* sp. 'kazumbe' between populations of *P.* cf. *polyodon*. We calculated D statistics for all possible combinations of populations with at least three samples for each species (the samples for P3 were limited to the collective set of locations used for P1 and P2). All D statistics used *S. diagramma* as the outgroup and were calculated with `AdmixTools` v5.0 (Patterson et al., 2012).

We used two approaches to identify outlier D statistics that we consider as evidence for differences in interspecific allele sharing among populations. Our first approach addresses the uncertainty in D statistic values caused by random sampling of individuals. We performed 1000 random permutations of the samples within each species and then re-calculated D statistics for each randomized data set to create a D statistic null distribution associated with each observed value. Since this procedure maintained sample sizes for each calculation, each distribution reflects the variability in values possible based on a calculation's specific sample sizes. We considered observed D statistic values that fell in the tails of their respective distributions as those representing substantial evidence for differences in interspecific allele sharing: values in the 2.5–5 or 95–97.5 percentile ranges (strong evidence) and values that were below the 2.5 or above the 97.5 percentiles (strongest evidence). Second, to address the uncertainty in D statistic values caused by the random sampling of SNPs, we calculated z-scores using the block jackknife procedure implemented in `AdmixTools`. Based on the block jackknife results, we considered a D statistic to significantly deviate from zero if it possessed an absolute z-score>3 (Durand et al., 2011). We completed several auxiliary analyses to evaluate the robustness of the D statistics to the outgroup identity and specification of the P3 taxa (Appendix S1, Section 1.2).

2.6 | Relationships between abundance and admixture

We assessed the relationship between abundance and the degree of admixture at five locations with both genetic sampling and

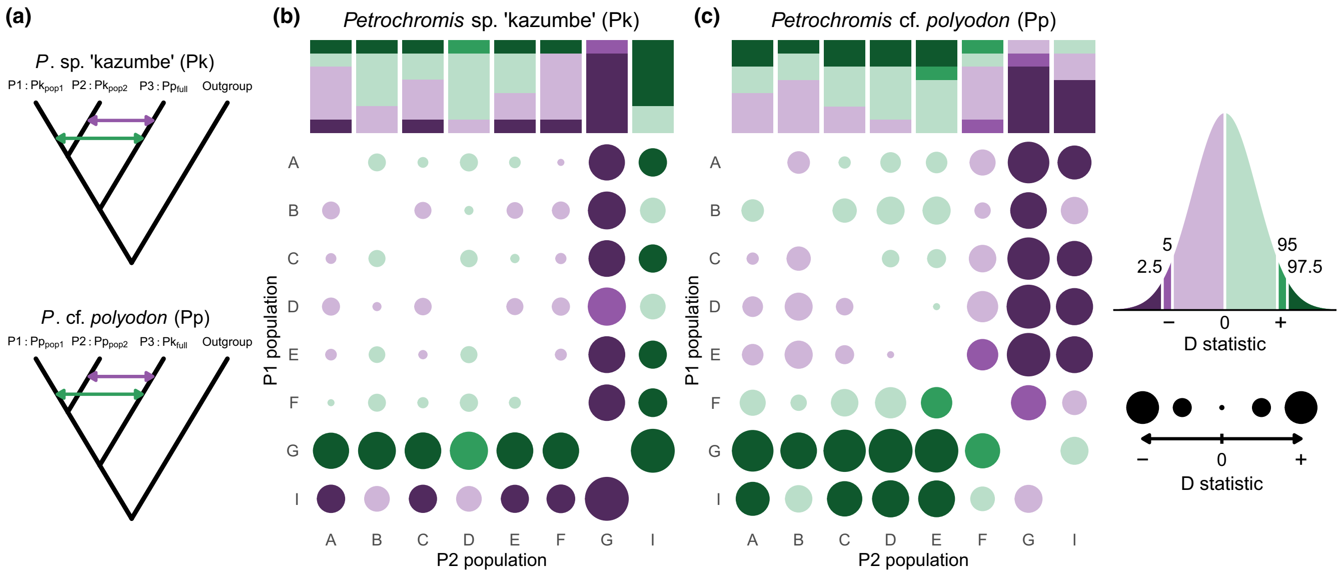

FIGURE 2 D statistics comparing two populations of one species (either *Petrochromis* sp. ‘kazumbe’ or *Petrochromis* cf. *polyodon*) and all samples of the other species reveal variation in interspecific allele sharing across populations. (a) The top topology (Topology 1a) illustrates the D statistic set-up to examine variation in interspecific allele sharing between populations of *P.* sp. ‘kazumbe’: [((P1: *P.* sp. ‘kazumbe’$_{\text{pop. 1}}$, P2: *P.* sp. ‘kazumbe’$_{\text{pop. 2}}$), P3: *P.* cf. *polyodon*$_{\text{full}}$), Outgroup]. The Topology 1a D statistics correspond to (b). The bottom topology (Topology 2a) illustrates the D statistic set-up to examine variation in interspecific allele sharing between populations of *Petrochromis* cf. *polyodon*: [((P1: *P.* cf. *polyodon*$_{\text{pop. 1}}$, P2: *P.* cf. *polyodon*$_{\text{pop. 2}}$), P3: *P.* sp. ‘kazumbe’$_{\text{full}}$), Outgroup]. The Topology 2a D statistics correspond to (c). In (b) and (c), D statistics are arranged in a matrix with locations ordered from north to south on both the horizontal and vertical axes. The size of the point represents the absolute magnitude of the D statistic (i.e. point size increases with larger deviations from zero). Positive D statistics are represented by green colours, and negative values are represented by purple colours. Positive values indicate that P1 displays greater interspecific allele sharing while negative values indicate greater interspecific allele sharing in P2. We identified outlier D statistics by comparing each observed value to a distribution of 1000 D statistics calculated from data sets where samples were randomly permuted within species. The shade of the colour indicates the percentile of the observed value in its respective randomized distribution: >5% and <95% (light green/purple); between 2.5% and 5% or between 95% and 97.5% (medium green/purple); <2.5% or >97.5% (dark green/purple). The barplot above each matrix summarizes the results from the P2 population (summarizing each column of the matrix). All D statistics shown in this figure were calculated using *Simochromis diagramma* as the outgroup [Colour figure can be viewed at wileyonlinelibrary.com]

abundance information to explore how variation in abundance corresponded to admixture. We extracted abundance information for *P.* sp. ‘kazumbe’ and *P.* cf. *polyodon* from visual surveys of cichlid communities by Golcher-Benavides (2021) (Table S6). Briefly, surveys of adults were conducted twice at each location by two divers who recorded species abundances at point counts arranged in a transect orthogonal to the shoreline. The point counts were located every 4 m, and the transect spanned a depth range of 0–15 m. Since a survey's total area varied based on the steepness of the depth gradient, we standardized species counts by the total area of the survey (obtaining an average density of individuals). We first standardized surveys within each year and then averaged the densities for each location's surveys to minimize the potential effects of an anomalous survey. We quantified admixture using the proportion of ancestry of the minor species (i.e. species with <50% ancestry) of each sample estimated with q from the $K = 2$ `entropy` model.

We described the relationship between standardized abundance (predictor) and admixture (response) separately for each species using generalized additive models (GAMs), which permitted flexibility in the shape of the relationship. We fit each GAM using a beta error and logit link function, a thin plate regression spline smooth function, and a basis dimension of four for the smooth term. We built GAMs with the `mgcv` package (Wood, 2017) using restricted maximum-likelihood estimation, and we assessed model fit using QQ-plots and plots of residuals versus fitted values created with the `gratia` package (Simpson, 2021; Figures S6 and S7).

2.7 | Demographic modelling

To examine the timing of divergence between *P.* cf. *polyodon* and *P.* sp. ‘kazumbe’ and how gene flow figured into the divergence process, we modelled the demographic history of the species using the composite likelihood approach implemented in `fastsimcoal2` v2.6 (Excoffier et al., 2013). We performed identical but separate analyses on the samples from the mid and north regions (the south region samples were excluded because we only found *P.* sp. ‘kazumbe’), which offers two strengths compared to analysing the regions together. First, analysing each region separately allows us to simplify the models by obviating the need to model intraspecific divergence between the regions, which is evidenced in both species by several analyses (e.g. PCAs, intraspecific F_{ST} calculations). Second, the separate analyses enable us to perform independent evaluations of the divergence history of *P.* sp. ‘kazumbe’ and *P.* cf. *polyodon*. The replicate analyses will help us gauge our confidence in how the species diverged depending on the degree to which the data sets yield congruent inferences.

We constructed the folded site frequency spectra (SFSs) separately for the mid and north regions using the `easySFS` script (https://github.com/isaacovercast/easySFS). The inclusion of both variant and invariant sites enabled estimation of absolute parameter values. During SFS creation, we addressed missing data by down-projecting each species to the number of individuals that maximized segregating sites in *P.* cf. *polyodon*. Projection down to a smaller sample size involves averaging over all possible resamples of the original data and can be used to form a complete SFS in the presence of missing data because the procedure can include sites with incomplete calls for all samples (see Marth et al. (2004) and Gutenkunst et al. (2009) for further details). This approach resulted in identical sample sizes and similar numbers of segregating sites in the two species for both regions' SFSs.

Population growth can bias parameter estimates and model selection when unaccounted for in demographic modelling (Momigliano et al., 2021). Thus, prior to modelling with `fastsimcoal2`, we used the folded SFSs to reconstruct the population sizes through time for each species/region subset (e.g. *P.* cf. *polyodon* from the north region) with `Stairway Plot` v2.1.1 (Liu & Fu, 2020), which is most accurate at inferring recent population histories (Liu & Fu, 2015; Patton et al., 2019). We used 1000 input files for pseudo-CI estimation, and, following the program's defaults, we used 67% of sites for training and four breakpoints calculated based on the number of sequences (nseq): (nseq - 2)/4, (nseq - 2)/2, (nseq - 2)*3/4, and nseq - 2. `Stairway Plot` inferred population size changes in both species (Figure S9). Thus, we included exponential population size change parameters for both species in all models.

We built 13 models (models involving gene flow are visualized in Figure 3) categorized into four classes based on the timing of gene flow during divergence: divergence in isolation (no gene flow), divergence with continuous gene flow (continuous gene flow), divergence with gene flow followed by isolation (early gene flow) and divergence in isolation followed by gene flow (recent gene flow). We built four different model variants within each model class involving gene flow: asymmetric gene flow, symmetric gene flow, unidirectional gene flow from *P.* sp. 'kazumbe' to *P.* cf. *polyodon* and unidirectional gene flow from *P.* cf. *polyodon* to *P.* sp. 'kazumbe'.

For each model, we conducted 500 independent runs to identify the parameter values leading to the maximum likelihood (the best run). Each run involved 150,000 coalescent simulations, a maximum of 40 expectation–maximization cycles, a minimum of 10 for the observed SFS entry count and a minimum relative difference in parameter values of 0.001 for the stopping criterion. We used broad search ranges with uniform and log-uniform distributions for all parameters (see Table S11). All models used a mutation rate of 3.5×10^{-9} per bp per generation, which is based on a mutation rate estimate for Lake Malawi cichlids (Malinsky et al., 2018).

Similar to Meier, Sousa, et al. (2017), we used two approaches to compare relative fits of models. First, we compared each model's best run using AIC. Second, based on the parameter values of each model's best run, we computed likelihood values from 100 SFS approximations each using 1000,000 coalescent simulations. Substantial overlaps in the likelihood distributions of models would indicate that disparity in fit was likely an artefact of variance in `fastsimcoal2`'s likelihood estimates. Additionally, with the best fit model, we compared the predicted data from the model to the observed data to evaluate whether the model provided satisfactory fit. Following Bagley et al. (2017), we visually compared the observed versus expected two-dimensional SFSs and the observed versus expected marginal one-dimensional SFSs.

To obtain estimates and CIs for the parameters of the best fit model, we implemented the non-parametric block bootstrapping approach used by Meier, Sousa, et al. (2017), which helps account for linkage disequilibrium. First, we created 100 bootstrap data sets by splitting the original data set into 100 non-overlapping blocks and then sampling 100 blocks with replacement to form data sets of equivalent size to the original. From each bootstrap data set, we constructed the SFS and then completed 300 runs of the best fit model (150,000 simulations per run). We report the median value and 95% CIs for the parameters based on the parameter estimates from the best run of each bootstrap data set.

3 | RESULTS

Sequencing generated a total of 7.74×10^{8} reads with 7.20×10^{8} reads (~93%) aligned to the *P. nyererei* reference genome and an average ± SD of $2.51\times10^{6}\pm3.10\times10^{3}$ aligned reads per individual. Across all data sets, average site-level mean depth ranged from 10.63 to 16.73. The number of sites in the *variant* and *allsites* data sets ranged from 40,400 to 69,409 and 2,620,316 to 3,102,141, respectively. Further details on each data set are shown in Table S3. The final data sets included 231 *P.* sp. 'kazumbe', 41 *P.* cf. *polyodon*, nine *S. diagramma* and six *P.* cf. *macrognathus* 'green'.

3.1 | Structure, divergence and diversity

The PCAs supported the distinctiveness of *P.* sp. 'kazumbe' and *P.* cf. *polyodon* and revealed fine-scale spatial differentiation in each species. The species in the full data set PCA separated into exclusive clusters along PC1 (Figure 1d). In the single-species PCAs, individuals sorted based on sampling region on PC1 and PC2 (Figure 1d), which reflects the isolation between regions imposed by the Luiche River delta (between the south and mid regions) and the Kigoma Bay (between the mid and north regions). The `popvae` analyses yielded comparable patterns to the PCAs (Figure S2). *P.* sp. 'kazumbe' clustered cleanly by sampling region while in *P.* cf. *polyodon*, the clustering largely reflected geography but was less clean than in *P.* sp.

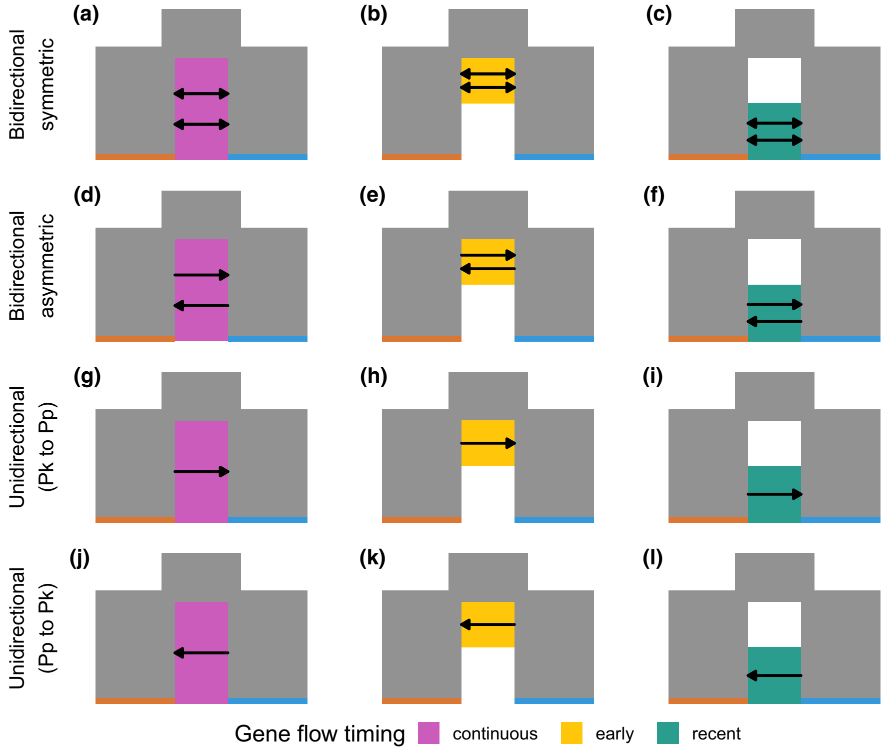

FIGURE 3 Visualizations of the post-divergence gene flow scenarios of speciation between *Petrochromis* sp. 'kazumbe' and *Petrochromis*. cf. *polyodon* tested with `fastsimcoal2`. These represent all of the tested scenarios of speciation except for divergence without post-divergence gene flow. Models are categorized into classes (each column of the figure) based on the timing of gene flow in the divergence process: continuous gene flow (gene flow throughout divergence; [a, d, g, j]), early gene flow (gene flow followed by isolation; [b, e, h, k]), and recent gene flow (isolation followed by gene flow; [c, f, i, l]). Each of the model classes involving gene flow included four model variants that differ based on the parameterization of gene flow: bidirectional symmetric gene flow (single parameter estimated for migration in both directions; [a–c]), bidirectional symmetric gene flow (separate parameters estimated for each migration direction; [d–f]), unidirectional gene flow from *P.* sp. 'kazumbe' to *P.* cf. *polyodon* (Pk to Pp; [g–i]) and unidirectional gene flow from *P.* cf. *polyodon* to *P.* sp. 'kazumbe' (Pp to Pk; [j–l]) [Colour figure can be viewed at wileyonlinelibrary.com]

'kazumbe' with several north region samples clustering with the mid-region samples.

The population genetic summary statistics further supported the distinctiveness of each species and fine-scale spatial differentiation. Interspecific F_{ST} estimates indicated strong differentiation between the species with a full data set F_{ST} (calculated with all individuals) of 0.399 (95% CI: 0.395–0.403) and the location-specific estimates ranging from 0.293 to 0.474. Estimates of d_{XY} were greater than π in both species. *P.* sp. 'kazumbe' was more diverse than *P.* cf. *polyodon*, with π estimates for *P.* sp. 'kazumbe' greater than *P.* cf. *polyodon* in the full data set and at all locations with sampling for both species, although the location I estimates were nearly identical (Figure 1b, Table S4). The inferences of higher π in *P.* sp. 'kazumbe' appeared generally robust to interspecific differences in sample size because π calculated from the subsampled data sets of *P.* sp. 'kazumbe' were higher than *P.* cf. *polyodon* at the data set level and at all locations except I (Figure S1). The intraspecific F_{ST} estimates ranged from 0.004 to 0.150 in *P.* sp. 'kazumbe' and 0.001 to 0.081 in *P.* cf. *polyodon* with values generally scaling with distance between the sampling locations and marked jumps in F_{ST} when the locations were from different sampling regions (Table S5, Figure S3).

3.2 | Ancestry estimation

The `entropy` analyses further confirmed the distinctiveness of each species and the presence of intraspecific spatial divergence in *P.* sp. 'kazumbe' (Figure 4). The individuals sorted cleanly by species based on *q* (proportion of ancestry) in the $K = 2$ model. Models with higher values of *K* (i.e. 3 and 4) distinguished both interspecific divergence and spatial divergence in *P.* sp. 'kazumbe', with *P.* sp. 'kazumbe' and *P.* cf. *polyodon* sorting into distinct groups and *P.* sp. 'kazumbe' further splitting into groups based on sampling region (Figure S4).

The `entropy` analyses provided evidence for small but appreciable amounts of ongoing hybridization that varied in extent across sampling locations. Based on our hybrid classification scheme using *q* (proportion of ancestry) and Q_{12} (interspecific ancestry), the majority of individuals (91.46%) from the mid and north regions were pure *P.* cf. *polyodon* or *P.* sp. 'kazumbe'. Hybrids were detected at four of the nine sampling locations in the mid and north regions (locations B, E, G, and I), and nearly all were consistent with being backcrosses. One location G individual had estimated *q* and Q_{12} values that placed it just outside of the F1 classification (slightly below the apex of the triangle in Figure 4c). This individual

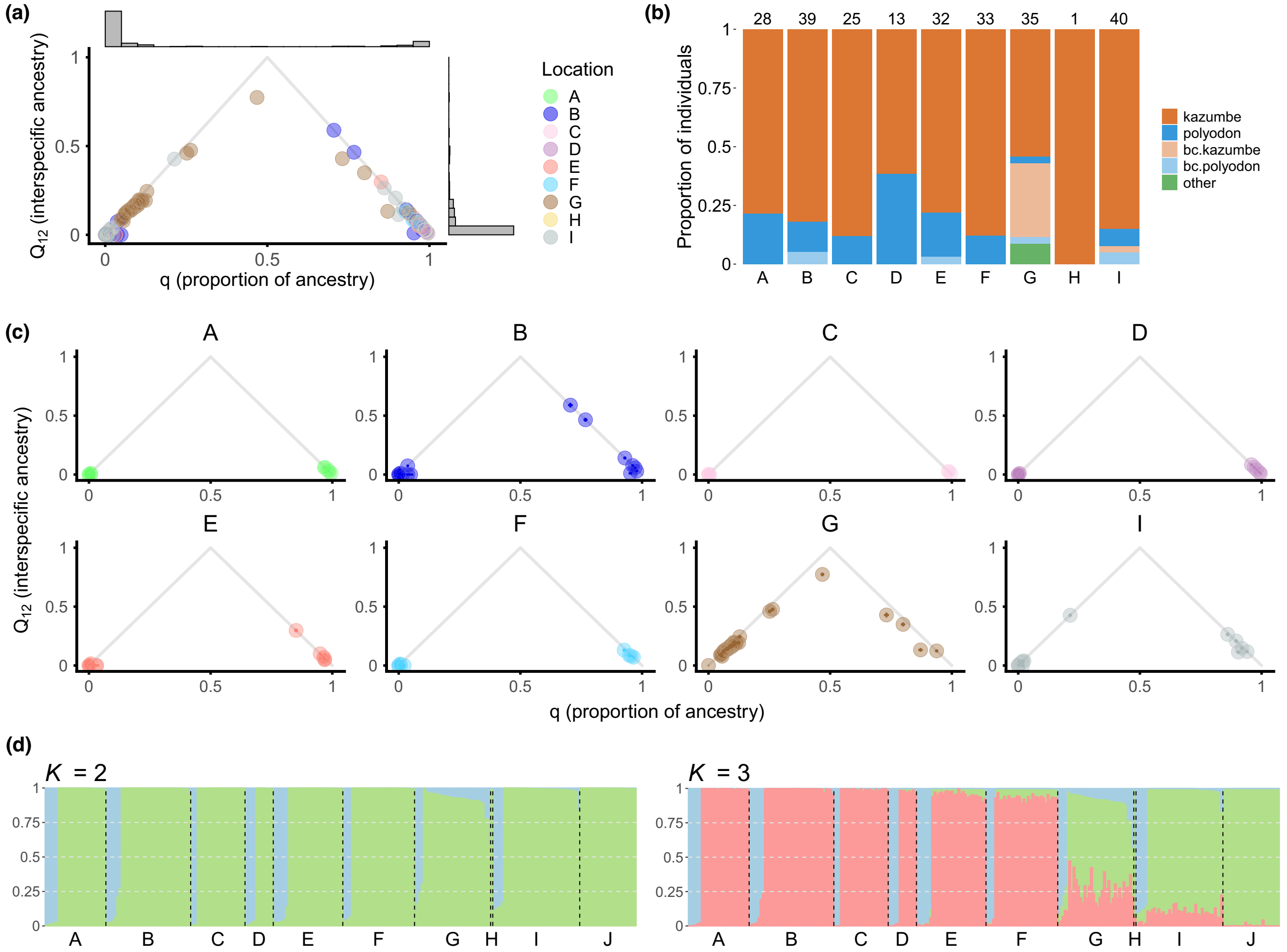

FIGURE 4 Estimation of ancestry with `entropy` supports inter- and intraspecific divergence and also admixed ancestry in some of the samples. (a) Estimates of *q* (proportion of ancestry; horizontal axis) and Q_{12} (interspecific ancestry; vertical axis) for all samples from the mid and north regions. Both *q* and Q_{12} are constrained between 0 (*q* = 0 indicates complete *Petrochromis* sp. 'kazumbe' ancestry; $Q_{12} = 0$ indicates no interspecific ancestry) and 1 (*q* = 1 indicates complete *Petrochromis* sp. *polyodon* ancestry; $Q_{12} = 1$ indicates complete interspecific ancestry). The marginal histograms show the densities of *q* and Q_{12} values. (b) Relative frequencies of different hybrid classes at each sampling location based on estimates of *q* and Q_{12}. The classes include full *P.* sp. 'kazumbe' (*kazumbe*), full *P.* cf. *polyodon* (*polyodon*), backcross to a parental species (*bc.kazumbe* and *bc.polyodon*) and admixed individuals that do not fall into a defined category (*other*). The number above each bar indicates the total number of individuals sampled at that location. (c) Plots of *q* and Q_{12} estimates split by sampling location (locations with sampling for both species are shown). Each sample includes the 95% credible intervals for *q* and Q_{12}, which are visualized as horizontal and vertical bars, respectively (note that all credible intervals are smaller than the point representing the mean ancestry estimate). (d) Barplots of *q* estimates support the existence of two species, which cluster separately in both the *K* = 2 (green: *P.* sp. 'kazumbe', blue: *P.* cf. *polyodon*) and *K* = 3 models (green/red: *P.* sp. 'kazumbe', blue: *P.* cf. *polyodon*). In addition, the *K* = 3 model reveals spatial structure in *P.* sp. 'kazumbe' with the north region samples (red; locations A, B, C, D, E and F) clustering separately from the mid and south region samples of *P.* sp. 'kazumbe' (green; locations G, H, I and J) [Colour figure can be viewed at wileyonlinelibrary.com]

may have represented an F1 and the disparity between its estimated ancestry and the expected F1 ancestry ($q = 0.5$, $Q_{12} = 1$) could have been caused by uncertainty in the allele frequencies of the parental species. Location G showed the most extensive admixture with hybrids constituting 42.9% of the sampled individuals. Hybridization appeared biased towards *P.* cf. *polyodon* with 19.5% of individuals with majority *P.* cf. *polyodon* ancestry belonging to a hybrid class versus 6.34% of individuals with majority *P.* sp. 'kazumbe' ancestry. Estimates of *q* from the *K* = 2 `entropy` model supported a *P.* sp. 'kazumbe' identity for all individuals from location J (the only south region location) and found no evidence of admixture since all individuals had >99% estimated *P.* sp. 'kazumbe' ancestry.

3.3 | D statistics

The D statistics provided clear support for variation in interspecific allele sharing between populations in both *P.* sp. 'kazumbe' and *P.* cf. *polyodon* (Figure 2). Based on the random permutations of samples, 39.3% of comparisons between *P.* sp. 'kazumbe' populations (Topology 1a) and 39.3% of comparisons between *P.* cf. *polyodon* populations (Topology 2a) were outside of the inner 90% percentile of their associated randomized D statistic distributions, and 90.9% (Topology 1a) and 81.8% (Topology 2a) of these outlier values were also outside of the inner 95% percentile. These comparisons represent strong evidence of imbalanced interspecific allele sharing among locations. For the *P.* sp. 'kazumbe' population comparisons, the sample randomization implicated two mid-region locations (G and I) as the primary outliers with location G showing evidence of having higher allele sharing with *P.* cf. *polyodon* than all other populations while location I showed less allele sharing in five out of seven comparisons. For the *P.* cf. *polyodon* population comparisons, the mid-region locations emerged as having evidence for excess allele sharing with *P.* sp. 'kazumbe' relative to most other populations (six out of seven for location G; four out of seven for location I), while location F also demonstrated evidence for more *P.* sp. 'kazumbe' allele sharing than location E.

The block jackknife results supported more extensive and complicated variation in interspecific allele sharing compared with the sample randomization approach (Figure S8). The majority of comparisons deviated significantly from zero (absolute *z*-score >3) both between *P.* sp. 'kazumbe' populations (Topology 1a; 53.6%) and between *P.* cf. *polyodon* populations (Topology 2a; 85.7%). Despite discrepancies in the prevalence of outliers identified by the sample randomization versus block jackknife approaches, they yielded partially congruent results since the set of comparisons identified as outliers from sample randomization were all outliers from the block jackknife. In summary, the block jackknife approach for identifying outlier D statistics suggested that many of the observed values deviate substantially from expectations under the random sampling of SNPs. However, based on the random sampling of individuals, a smaller set of values (primarily involving the mid region locations) show notable deviations.

The D statistics were highly consistent over the different variations of the calculations used to test for robustness. D statistic values were not sensitive to using all samples of the species for P3 (a topologies) versus limiting the P3 samples to the same location as P2 (b topologies), as evidenced by highly correlated D statistic matrices (all correlations >0.95). The D statistic results were also robust to the identity of the outgroup (*S. diagramma* versus *P.* cf. *macrognathus* 'green') with corresponding sets of D statistics using the different outgroups showing high correlations (all correlations >0.92). All supporting D statistic analyses are provided in the Appendix S1 (Tables S8–S10; Figure S8).

3.4 | Relationships between abundance and admixture

The visual surveys found higher standardized abundances of *P.* sp. 'kazumbe' than corresponding *P.* cf. *polyodon* populations at all locations (Figure 5c). The GAM results (Table S7) indicate that *P.* cf. *polyodon* displayed a clear, nearly linear negative relationship between standardized abundance and admixture (edf = 1.00, chi.sq = 17.42, p < 0.001; Figure 5b), suggesting that admixture in *P.* cf. *polyodon* was generally less prevalent at locations with high conspecific abundance. *P.* sp. 'kazumbe' showed elevated admixture at the location where it was least abundant but then showed minimal admixture with no clear relationship with abundance across the remaining locations where it was more common (edf = 2.97, chi.sq = 258.22, *p* < .001; Figure 5a).

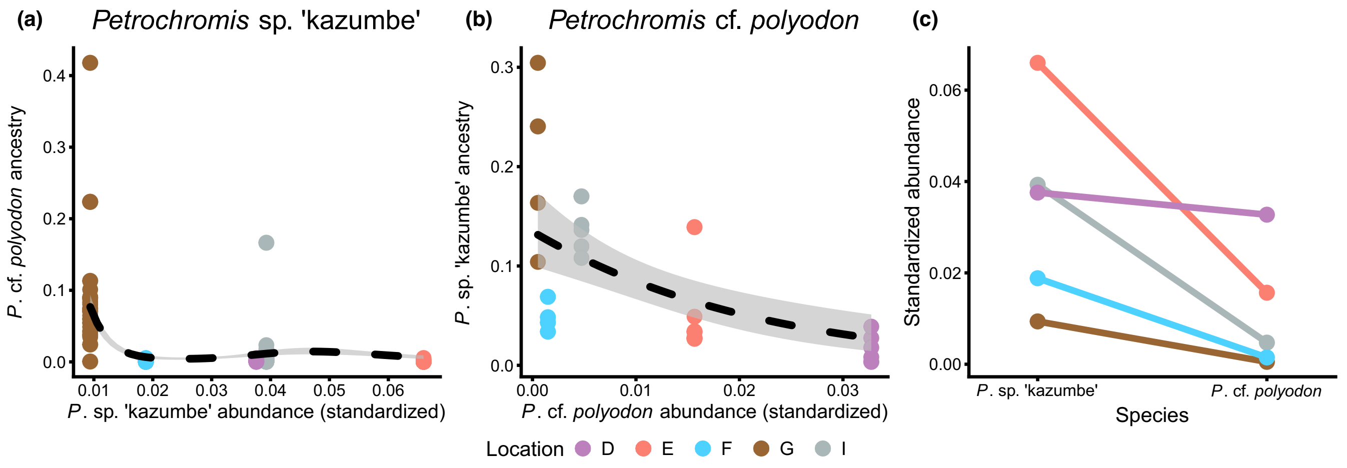

FIGURE 5 Generalized additive models revealed (a) elevated admixture in *Petrochromis* sp. 'kazumbe' (quantified as ancestry of *Petrochromis* cf. *polyodon*) at its rarest location but no clear trend in admixture across the remaining locations where it was more common. (b) *P.* cf. *polyodon* showed a clear negative relationship between abundance and ancestry of *P.* sp. 'kazumbe'. (c) *P.* cf. *polyodon* was the less abundant species at all locations used to evaluate the relationships between abundance and minor parental species ancestry. Together, these results suggest that abundance may influence the prevalence of admixture. We derived all ancestry estimates from the *K* = 2 `entropy` model [Colour figure can be viewed at wileyonlinelibrary.com]

3.5 | Demographic modelling

After down-projection, the data sets used in creating the SFSs contained 16 (mid) and 50 (north) haploid sequences of each species and included the following numbers of segregating sites: 42,563 (*P.* sp. 'kazumbe') and 41,721 (*P.* cf. *polyodon*) in the north region; 38,352 (*P.* sp. 'kazumbe') and 32,576 (*P.* cf. *polyodon*) in the mid region.

Modelling using the mid and north region data sets yielded similar inferences (Figure 6). With both data sets, models with no gene flow performed the worst, with the inclusion of gene flow substantially improving fit. Within each model class involving gene flow, the model with asymmetric, bidirectional gene flow outperformed the alternative models in both data sets based on AIC. The best performing models (based on AIC) slightly differed between the regions. The top three for the north region were the recent, early and continuous asymmetric gene flow models, while the top three for the mid region were the models involving asymmetric and symmetric recent gene flow and also asymmetric continuous gene flow. Details on model fit including $\log_{10}$ (likelihood), ΔLikelihood (difference between maximum possible and obtained likelihood on the $\log_{10}$ scale), AIC, ΔAIC and relative likelihood (Akaike's weight of evidence; Excoffier et al., 2013) are included in Table S12.

The recent, asymmetric gene flow scenario was the best supported of the 13 tested models in both regions (i.e. highest likelihood and lowest AIC; relative likelihoods of 1.00 for both mid and north; Table S12). The likelihood distributions of the best fit models were non-overlapping with the distributions of the other models (Figure 6b,d), indicating that their superior fits were not caused by variance in `fastsimcoal`'s likelihood approximation. Our comparison of the observed two-dimensional and marginal one-dimensional SFSs to the expected SFSs generated under each of the best fit models indicated that the models provided fairly good fits to the observed data (Figures S10–S13).

The best fit model from each region yielded similar but not fully congruent parameter estimates (95% CIs provided in parentheses). Divergence time estimates were similar between the models: 176,858 (133,475–221,999) generations in the mid region; 188,900 (148,672–255,114) generations in the north region. Both regions' models estimated that the onset of gene flow occurred only recently relative to the divergence times (11,118 (5416–20,241) generations and 25,669 (15,318–50,860) generations in the mid and north regions, respectively). Additionally, both models estimated a higher gene flow probability from *P.* cf. *polyodon* into *P.* sp. 'kazumbe' (going forward in time) than in the inverse direction. We provide the estimates (median bootstrap value) and 95% CIs for all parameters of the best fit models in Table S13.

4 | DISCUSSION

Studying how gene flow figures into the speciation process represents a critical endeavour in speciation research. Identifying the timing, source and duration of gene flow during speciation facilitates a better understanding of the factors influencing divergence and the maintenance of distinctiveness. Here, we examined recent divergence between *P.* sp. 'kazumbe' and *P.* cf. *polyodon*, two members of the Lake Tanganyika cichlid radiation. First, we corroborated and extended a previous evaluation of these species (Wagner et al., 2012) by showing that *P.* sp. 'kazumbe' and *P.* cf. *polyodon* represent genetically distinct taxa, both species exhibit fine-scale spatial differentiation structured by dispersal barriers, and the species maintain their distinctiveness in sympatry. Second, we provide evidence of ongoing hybridization that varies across populations and find that hybridization in some cases shows correspondence to variation in abundance. Third, demographic modelling provides evidence that the species diverged in initial isolation followed by gene flow. Below, we elaborate on these findings and discuss their implications for understanding how sympatric diversity develops in adaptive radiation.

4.1 | Speciation in *P.* sp. 'kazumbe' and *P.* cf. *polyodon*

Given the current sympatric distributions of *P.* sp. 'kazumbe' and *P.* cf. *polyodon* in the Kigoma region of Tanzania, it was unclear whether these taxa diverged in sympatry or whether gene flow occurred during the speciation process. The demographic models we tested were designed to distinguish between a scenario consistent with sympatric divergence with gene flow and divergence in isolation followed by gene flow upon secondary contact. We found that the recent gene flow model, wherein the species diverged in isolation with subsequent gene flow, was the best fit model in both regions' data sets (Figure 6; Table S12). The recent gene flow model is most consistent with an allopatric speciation scenario where the species diverged in spatial isolation and then recommenced gene flow upon secondary contact.

Beyond supporting the same speciation scenario, the models from the two regions' data sets yielded similar but not completely congruent inferences. Both data sets yielded similarly recent divergence times (~189,000 versus ~177,000 generations) with substantially overlapping confidence intervals. These estimates indicate that *P.* sp. 'kazumbe' and *P.* cf. *polyodon* represent a recent speciation event of comparable timing to some estimates of spatial divergence among cichlid populations in Lake Tanganyika (e.g. Koblmüller et al., 2011; Sefc, Mattersdorfer, Ziegelbecker, et al., 2017; Winkelmann et al., 2017).

Additionally, both data sets supported very recent estimates for the onset of gene flow (north: ~26,000 generations; mid: ~11,000 generations), suggesting that the species have experienced secondary contact only briefly relative to their divergence time. However, the estimates for each region differ, with the north region data set possessing a later estimate. Such discrepancies between regions could indicate differences in the timing of secondary contact across the current zone of sympatry between the species. Historical lake level fluctuations are thought to regularly redistribute rocky littoral habitats and associated cichlids (Koblmüller et al., 2008; Sefc,

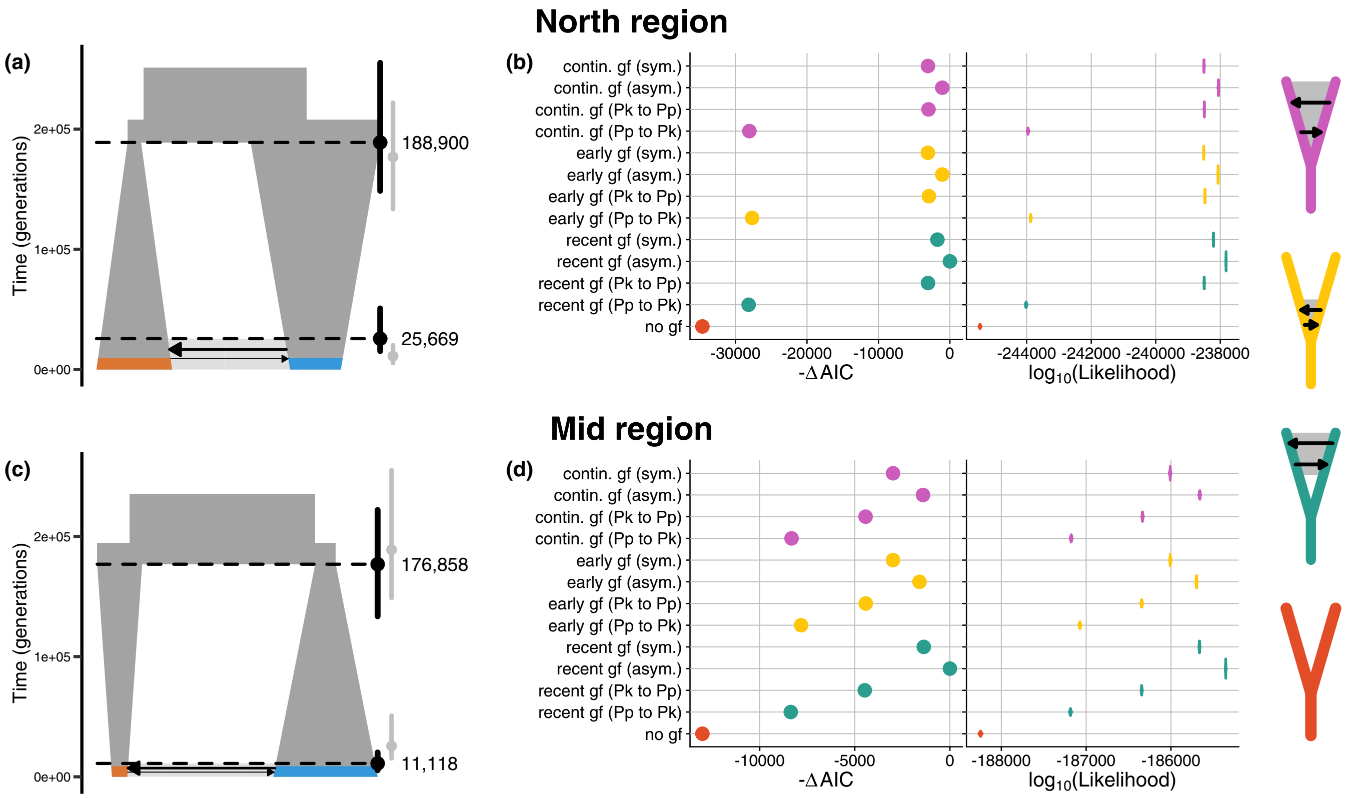

FIGURE 6 Comparison of 13 models of divergence with `fastsimcoal2` provides support for a secondary contact scenario where *Petrochromis* sp. 'kazumbe' and *Petrochromis* cf. *polyodon* diverged in isolation and recently initiated gene flow. Separate but identical modelling was performed on the north (top row; (a) and (b)) and mid region (bottom row; (c) and (d)) data sets. The two left figures ((a) and (c)) visualize the best fit models for the north and mid regions based on the median parameter estimates derived from block bootstrapping. The left, orange polygon represents *P.* sp. 'kazumbe', and the right, blue polygon represents *P.* cf. *polyodon*. The polygons in (a) and (c) are scaled to estimated effective population sizes, and the arrow sizes are scaled to estimated migration rates (visualized forward in time). The higher dashed line in (a) and (c) denotes the divergence time, and the lower dashed line denotes the estimated timing of migration onset (point estimates for these values are shown to the right of the lines). The black vertical bars at the right end of the dashed lines represent the 95% confidence intervals for divergence time and timing of gene flow recommencement. To facilitate comparison between each region's results, the models in (a) and (c) are plotted on the same time scale, and the estimates for divergence time and timing of gene flow recommencement from the alternative region are plotted in grey. The plots in (b) and (d) display metrics of model fit for all tested models. The left plots show -ΔAIC, and the right plots show the distribution of likelihood values from 100 expected site frequency spectra approximated based on the parameter values of the best fit model. For both the -ΔAIC and likelihood plots, model fit improves along the horizontal axis. Colours in (b) and (d) correspond to the timing of gene flow (*gf*) in the model: continuous (pink), early (yellow), recent (green) and none (red). Models are labelled in (b) and (d) by the timing and specific parameterization of gene flow: symmetric (*sym.*), asymmetric (*asym.*), migration only from *P.* sp. 'kazumbe' to *P.* cf. *polyodon* back in time (*Pk to Pp*), migration only from *P.* cf. *polyodon* to *P.* sp. 'kazumbe' back in time (*Pp to Pk*). Detailed information on model fit can be found in Table S12 [Colour figure can be viewed at wileyonlinelibrary.com]

Mattersdorfer, Ziegelbecker, et al., 2017; Sturmbauer, 1998), and such fluctuations could plausibly facilitate the movement or expansion of one or both species' ranges. In this scenario, we would expect the populations at the margins of each species' range closest to the initial point of overlap to have experienced the longest time in sympatry. Alternatively, given that the mid region has the more recent estimates for both gene flow recommencement and divergence, the disparity in the estimated timings of both events could reflect model inaccuracy in one or both regions, such as biases introduced by unmodeled demographic phenomena.

Several caveats are important to recognize related to the demographic modelling results. First, although we consider the allopatric scenario most likely, demographic modelling alone cannot demonstrate initial allopatry, and thus this result could be consistent with alternative scenarios such as the rapid development of reproductive barriers without spatial isolation followed by a weakening of barriers that facilitated post-divergence gene flow. For example, gene flow can rapidly attenuate in sympatry via the development of prezygotic barriers (e.g. mate choice) prompted by disruptive selection and then a shift of the selection regime could weaken the barriers and favour gene flow following a period of isolation (e.g. Seehausen et al., 1997). Second, the histories of cichlid taxa in Lake Tanganyika likely involve substantial complexity including temporally fluctuating population sizes and variance in the history of genetic isolation both spatially and through time. Our models represent a balance between this complexity and model tractability, which means that our modelling cannot comprehensively characterize the histories of *P.* sp. 'kazumbe' and *P.* cf. *polyodon*. For example, the recent gene flow in our models could actually represent several cycles of isolation followed by gene flow, which would be undetectable based on our current modelling and data limitations. Future work could consider more of this complexity, which could be facilitated by more extensive genomic data sets (e.g. whole genome data and sampling across more species).

4.2 | Ongoing hybridization

Demographic modelling and `entropy` analyses support ongoing, bidirectional gene flow between *P.* sp. 'kazumbe' and *P.* cf. *polyodon* (Figures 4 and 6). These analyses further suggest that interspecific gene flow is limited. The best fit demographic models estimate low probabilities of gene flow in both directions (Table S11). Concordantly, most individuals showed no admixed ancestry based on the `entropy` analyses, and hybrids were largely limited to backcrosses and later-stage hybrids. Evidence of admixture was restricted to the north and mid sampling regions where the species co-occur. Despite ongoing hybridization, *P.* sp. 'kazumbe' and *P.* cf. *polyodon* appear to be maintaining their distinctiveness given that all co-occurring populations showed substantial differentiation and divergence (Figure 1).

Our analyses also provided clear evidence of spatially varying admixture in both species. In particular, `entropy` identified admixed individuals at only a subset of sampling locations (Figure 4), and D statistics provided evidence of differences in interspecific allele sharing between some populations in both species (Figure 2). Despite taking different approaches to the inference of admixture, D statistics and `entropy` showed reassuring congruence. The strongest evidence of extensive admixture from both approaches involved mid region locations (G and I). `entropy` inferred admixed ancestry for a substantial proportion of individuals in both species at location G and in *P.* cf. *polyodon* at location I. Similarly, the D statistics provided strong evidence that P. sp. 'kazumbe' at location G and *P.* cf. *polyodon* at both locations G and I displayed higher interspecific allele sharing than conspecifics from most other locations.

Both `entropy` and D statistics generally supported minimal variation in admixture between locations in the north region (locations A–F), although the results from this region revealed some minor discrepancies between the inferences from each method. The `entropy`-based hybrid classification identified several *P.* cf. *polyodon* individuals with admixed ancestry at two north region locations (B and E), while D statistics did not support an excess of interspecific allele sharing in *P.* cf. *polyodon* at these locations when compared to other locations. Additionally, the D statistics revealed evidence of excess allele sharing in *P.* cf. *polyodon* at location F relative to E despite the `entropy`-based hybrid classification detecting no admixed individuals at F. The apparent discrepancies between results from these methods highlight the value of using multiple distinct approaches to study hybridization. D statistics are sensitive to gene flow under a broad set of conditions (e.g. large asymmetries in parental contributions, various timings of gene flow; Hibbins & Hahn, 2021; Kong & Kubatko, 2021). In contrast, our `entropy`-based hybrid classification is based on expected ancestry without precise bounds and is conservative at detecting hybrids with minimal ancestry from the minor parental species. Individuals whose estimated ancestry slightly deviated from the parental species ancestries (Q_{12} = 0 and q = 0 or 1) would thus be classified as unadmixed, whereas D statistics may be identifying small amounts of admixture in species' histories.

P. sp. 'kazumbe' and *P.* cf. *polyodon* serve as a detailed case study of admixture between closely related Lake Tanganyika cichlids. Hybridization between Tanganyikan cichlids has been documented in several other taxa mostly using mitochondrial and microsatellite genetic data. Prior work in *Tropheus* provides evidence for historical admixture from past contact in multiple populations (e.g. Egger et al., 2007; Sefc, Mattersdorfer, Hermann, & Koblmüller, 2017; Sefc, Mattersdorfer, Ziegelbecker, et al., 2017). Recent/ongoing gene flow has also been detected in *Ophthalmotilapia* species, which is restricted to regions of sympatry (Nevado et al., 2011). By using genomic data and several modelling approaches, our study reveals additional details about how closely related taxa in sympatry can interact via gene flow. We demonstrate that the extent of admixture can vary both based on where taxa occur in sympatry and across different co-occurring populations. It remains unclear whether the patterns and prevalence of gene flow demonstrated here are unusual or common among closely related Lake Tanganyika cichlid taxa in sympatry. However, the inference of minimal gene flow following divergence between sympatric *Altolamprologus* species (Koblmüller et al., 2017) and the common pattern of reciprocal monophyly in mitochondrial haplotypes among sympatric taxa (e.g. Sturmbauer et al., 2003; Wagner et al., 2012) suggests that gene flow may not always be pervasive among Lake Tanganyikan cichlids.

Hybridization outcomes among species can differ because of variation in the ecological or evolutionary contexts in which the species interact, and we discuss several potential explanations for the variation in hybridization below. First, abundances of the interacting species could lead to variation in hybridization by altering opportunities for conspecific versus heterospecific matings. One scenario, proposed by Hubbs (1955), posits that with an imbalance in interspecific abundances, the rarer species will more readily hybridize due to reduced opportunities for conspecific mating. Hubbs' principle could explain the relationships between admixture and abundance in the two *Petrochromis* species. *P.* cf. *polyodon* was rarer than corresponding *P.* sp. 'kazumbe' populations at all locations. Accordingly, we found that admixture in *P.* cf. *polyodon* declined with abundance (Figure 5b), which aligns with the idea that *P.* cf. *polyodon* more frequently interbreeds with individuals with *P.* sp. 'kazumbe' ancestry when conspecifics are scarcer. Besides elevated admixture at its rarest location, admixture in *P.* sp. 'kazumbe' was largely invariable (and minimal) despite variation in abundance (Figure 5a), which is expected under Hubbs' principle for the more common species. It may also be true that both species tend to hybridize more frequently when rare or that there are threshold effects with rarity that decrease admixture nonlinearly with abundance. Such phenomena could explain the increase in admixture with decreased abundance in the ubiquitously rare *P.* cf. *polyodon* as well as the elevated admixture in *P.* sp. 'kazumbe' at only its rarest location.

Several caveats are important to consider when evaluating abundance as a potential driver of hybridization. First, our explanation for the relationships between abundance and admixture presupposes that the contemporary abundances are representative of abundances over the timespan that the detected admixture was generated. This idea is reasonable given that we expect that the admixed ancestry estimated with `entropy` largely reflects hybridization from the recent past, during which the species abundances could have conceivably remained stable. Admixed individuals were largely composed of backcrosses, and repeated backcrossing rapidly attenuates minor parental species ancestry to where it is difficult (and in some cases, impossible) to estimate after tens of generations or less (McFarlane & Pemberton, 2019). Additionally, there are inherent limitations in studying these relationships at only five locations, such as spurious relationships that could arise because of factors that influence hybridization but were unaccounted for in our models. Future studies should more rigorously assess this hypothesis to explain spatial variation in hybridization.

There are several other possible (but not mutually exclusive) explanations for the observed patterns in admixture. Anthropogenic disturbance may facilitate hybridization by promoting close contact of previously isolated species via habitat homogenization, impairing individuals' abilities to discern conspecifics from heterospecifics, or reducing selection against hybrids (reviewed in Grabenstein & Taylor, 2018). Anthropogenic disturbance could lead to variation in hybridization across locations if the locations experience differing degrees of disturbance, as has been observed in Lake Victoria cichlids where a gradient of water turbidity from eutrophication led to spatially varying hybridization and species collapse by eroding visually-mediated species boundaries (Seehausen et al., 1997, 2008). Although we lack complete data on water clarity or disturbance for the study sites included here, available evidence provides little support for a dominant role of disturbance. Specifically, we detected admixed individuals at both a high (G; Hilltop) and low (I; Jakobsen's Beach) disturbance location based on levels of anthropogenic sedimentation observed by McIntyre et al. (2005). Additionally, McIntyre et al. (2005) documented high levels of sedimentation at locations G (Hilltop) and E (Kalalangabo), the latter of which had minimal admixture.

Differences in the nature of species boundaries across locations could also contribute to the observed patterns in hybridization. Variation in the genomic architecture underlying reproductive isolation could lead to disparate hybridization outcomes if different co-occurring populations developed distinct genomic architectures. Both species are non-vagile and likely experience minimal intraspecific gene flow between populations separated by unsuitable non-rocky shoreline. If dispersal barriers have been sufficiently stable to impede gene flow for an extended period of time, it is possible that unique genetic bases of reproductive isolation could arise from processes that promote population differentiation.

Additionally, variation in the strength or duration of processes that reduce hybridization, including reinforcement and ecological character displacement (which can strengthen reproductive isolation as a by-product), could lead to differences in the extent of hybridization (Hopkins, 2013). Since our demographic modelling results suggest that *P.* sp. 'kazumbe' and *P.* cf. *polyodon* have recently achieved secondary contact and the timings of gene flow recommencement slightly vary between sampling regions (Figure 6), different populations of the two species may have been co-occurring for different amounts of time. Reinforcement or ecological character displacement may thus have been operating for different durations in different populations across the region of co-occurrence. Although speculative, our demographic inferences regarding the timing of gene flow recommencement and the prevalence of admixture (i.e. based on ancestry estimation and interspecific allele sharing) in the mid and north regions follow the expected pattern under reinforcement and/or ecological character displacement since the region with the more recent estimated onset of gene flow also has the higher prevalence of hybridization (mid region).

4.3 | Sympatry in Lake Tanganyika cichlids

Despite an abundance of incipient diversity (i.e. distinctive colour variants), intriguingly few recently formed taxa (i.e. diverged in the last several hundred thousand years) exist in sympatry in Lake Tanganyika cichlids. Our findings together with previous investigations of intra- and interspecific divergence in Lake Tanganyika cichlids are informative for considering the limits of sympatric diversity in this radiation. First, since many Lake Tanganyika rock-dwelling cichlids can readily diverge across even minor dispersal barriers, the generation of divergent populations may not be limiting to sympatric diversity compared with the challenges to species boundaries posed by secondary contact. Depending on the stability of dispersal barriers, incipient forms may not persist long enough to achieve secondary contact. Because incipient forms frequently show an ostensible lack of ecological differentiation, it is also possible that they may fail to ecologically coexist even if sufficient reproductive barriers exist that would prevent species collapse.

Several lines of evidence indicate that incipient forms may need substantial time to evolve reproductive barriers, which suggests that populations may tend to fuse in secondary contact even with substantial divergence in allopatry. First, closely related colour variants in Lake Tanganyika regularly demonstrate weak mate preferences for conspecifics (Egger et al., 2010; Sefc et al., 2015; Sefc, Mattersdorfer, Ziegelbecker, et al., 2017; reviewed in Rometsch et al., 2020), colour variants have failed to maintain their distinctiveness in artificial sympatry (Egger et al., 2012), and admixture in colour variants has been documented in other putative cases of natural secondary contact (e.g. Sefc, Mattersdorfer, Hermann, & Koblmüller, 2017; Sefc, Mattersdorfer, Ziegelbecker, et al., 2017). Additionally, experimental crosses of haplochromine cichlids have produced viable hybrids across millions of years of divergence (Stelkens et al., 2010, 2015), which suggests that many East African cichlids may require substantial time to evolve complete reproductive isolation. The time interval between divergence and secondary contact in *P.* sp. 'kazumbe' and *P.* cf. *polyodon* estimated with demographic models (~163,000 (north) and ~166,000 (mid) generations based on point estimates; Figure 6; Table S13) also aligns with the idea that incipient forms may experience considerable time in isolation before successfully coexisting in sympatry. It will be interesting for future work to evaluate isolation intervals for other sister taxa across the Lake Tanganyika cichlid radiation, which will help establish if sympatry among closely related taxa in this radiation is generally preceded by an extended phase of isolation.

5 | CONCLUSIONS

In this study, we demonstrate that applying demographic modelling approaches in concert with methods for detecting admixture represents a powerful strategy for characterizing gene flow during speciation. In particular, the `entropy` and D statistic analyses document current patterns of admixture while the demographic modelling provided important context by clarifying the timing of interspecific gene flow during the divergence process. For *P.* sp. 'kazumbe' and *P.* cf. *polyodon*, these approaches provided evidence that divergence occurred in isolation with recent secondary contact and that the degree of admixture varies in extent between different co-occurring populations.

Our work provides an illustration of how closely related species in secondary contact can show variable reproductive isolation. Situations involving substantial but porous gene flow barriers between species may prove especially valuable for unravelling how and when speciation proceeds at its latter stages, which remain poorly understood (Kulmuni et al., 2020) including in cichlids (Rometsch et al., 2020). Additionally, the factors that dictate the possibility and prevalence of sympatry in incipient species remain unresolved, and these dynamics may be a key constraint on both the persistence of new species and the build-up of diversity (Gillespie et al., 2020; Price, 2008).

AUTHOR CONTRIBUTIONS

ALL and CEW designed the project. CEW and JGB conducted fieldwork (sample collection and visual surveys). ALL, JGB and CEW conducted lab work. ALL performed analyses with contributions from JAR. ALL drafted the manuscript with contributions from all authors.

CONFLICT OF INTERESTS

The authors declare no conflicts of interest.

ACKNOWLEDGEMENTS

We thank past and present members of the Wagner laboratory for feedback and assistance at various stages of the project. We particularly want to acknowledge Liz Mandeville for assistance with lab work and bioinformatics and Chad Brock for discussions and assistance with analyses. We also thank George Kazumbe for many years of help during fieldwork and Amy McCune and Irby Lovette for their important contributions to the initial genetic work on this system. The manuscript was substantially improved by comments from Alex Buerkle, Matt Carling, Hannah Augustijnen, Kay Lucek and two anonymous reviewers. Computing was performed with an allocation of the Teton Intel x86_64 Computing Cluster from the University of Wyoming's Advanced Research Computing Center. This project was supported by NSF grant DEB-1556963.

DATA AVAILABILITY STATEMENT

Raw sequence data (fastq files) are deposited in the NCBI Sequence Read Archive (BioProject ID PRJNA783341). Scripts used in processing and analysis and input files for demographic modelling are archived on Zenodo (DOI: 10.5281/zenodo.5703115) and can also be found at the following github repository: https://github.com/AlexLewanski/petrochromis_project.

ORCID

Alexander L. Lewanski https://orcid.org/0000-0001-5843-0837
Jessica A. Rick https://orcid.org/0000-0002-8927-220X
Catherine E. Wagner https://orcid.org/0000-0001-8585-6120

REFERENCES

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C. A., Buggs, R., Butlin, R. K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S. H., Hermansen, J. S., Hewitt, G., Hudson, A. G., Jiggins, C., ... Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, *26*, 229–246.

Advanced Research Computing Center (2018). Teton Computing Environment: Intel x86_64 cluster.

Bagley, R. K., Sousa, V. C., Niemiller, M. L., & Linnen, C. R. (2017). History, geography and host use shape genomewide patterns of genetic variation in the redheaded pine sawfly (Neodiprion lecontei). *Molecular Ecology*, *26*, 1022–1044.

Battey, C. J., Coffing, G. C., & Kern, A. D. (2021). Visualizing population structure with variational autoencoders. *G3*, *11*, 1–11.

Bolnick, D. I., & Fitzpatrick, B. M. (2007). Sympatric speciation: Models and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics*, *38*, 459–487.

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A. Y., Lim, Z. W., Bezault, E., Turner-Maier, J., Johnson, J., Alcazar, R., Noh, H. J., Russell, P., Aken, B., Alföldi, J., Amemiya, C., Azzouzi, N., ... Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, *513*, 375–381.

Cohen, A. S., Lezzar, K. E., Tiercelin, A. J. J., & Soreghan, M. (1997). New palaeogeographic and lake-level reconstructions of Lake Tanganyika: Implications for tectonic, climatic and biological evolution in a rift lake. *Basin Research*, *9*, 107–132.

Coyne, J. A., & Orr, H. A. (1989). Patterns of speciation in drosophila. *Evolution*, *43*, 362–381.

Coyne, J. A., & Orr, H. A. (1997). "Patterns of speciation in drosophila" revisited. *Evolution*, *51*, 295–303.

Cutter, A. D. (2012). The polymorphic prelude to Bateson-Dobzhansky-muller incompatibilities. *Trends in Ecology and Evolution*, *27*, 209–218.

Cutter, A. D., & Gray, J. C. (2016). Ephemeral ecological speciation and the latitudinal biodiversity gradient. *Evolution*, *70*, 2171–2185.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158.

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*, 2239–2252.

Egger, B., Koblmüller, S., Sturmbauer, C., & Sefc, K. M. (2007). Nuclear and mitochondrial data reveal different evolutionary processes in the Lake Tanganyika cichlid genus Tropheus. *BMC Evolutionary Biology*, *7*, 1–14.

Egger, B., Mattersdorfer, K., & Sefc, K. M. (2010). Variable discrimination and asymmetric preferences in laboratory tests of reproductive isolation between cichlid colour morphs. *Journal of Evolutionary Biology*, *23*, 433–439.

Egger, B., Sefc, K. M., Makasa, L., Sturmbauer, C., & Salzburger, W. (2012). Introgressive hybridization between color morphs in a population of cichlid fishes twelve years after human-induced secondary admixis. *Journal of Heredity*, *103*, 515–522.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*, e1003905.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*, 1567–1587.

Gante, H. F., Matschiner, M., Malmstrøm, M., Jakobsen, K. S., Jentoft, S., & Salzburger, W. (2016). Genomics of speciation and introgression in princess cichlid fishes from Lake Tanganyika. *Molecular Ecology*, *25*, 6143–6161.

Gillespie, R. G., Bennett, G. M., De Meester, L., Feder, J. L., Fleischer, R. C., Harmon, L. J., Hendry, A. P., Knope, M. L., Mallet, J., Martin, C., Parent, C. E., Patton, A. H., Pfennig, K. S., Rubinoff, D., Schluter, D., Seehausen, O., Shaw, K. L., Stacy, E., Stervander, M., … Wogan, G. O. (2020). Comparing adaptive radiations across space, time, and taxa. *Journal of Heredity*, *111*, 1–20.

Golcher-Benavides, J. (2021). *Ecological and evolutionary drivers of fish community diversity in Lake Tanganyika*, Ph.D. thesis. University of Wyoming.

Gompert, Z., Lucas, L. K., Buerkle, C. A., Forister, M. L., Fordyce, J. A., & Nice, C. C. (2014). Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology*, *23*, 4555–4573.

Good, J. M., Handel, M. A., & Nachman, M. W. (2008). Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. *Evolution*, *62*, 50–65.

Grabenstein, K. C., & Taylor, S. A. (2018). Breaking barriers: Causes, consequences, and experimental utility of human-mediated hybridization. *Trends in Ecology and Evolution*, *33*, 198–212.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H. Y., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Pääbo, S. (2010). A draft sequence of the neandertal genome. *Science*, *328*, 710–722.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*, e1000695.

Hibbins, M. S., & Hahn, M. W. (2021). Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, *220*, iyab173.

Hopkins, R. (2013). Reinforcement in plants. *New Phytologist*, *197*, 1095–1103.

Hubbs, C. L. (1955). Hybridization between fish species in nature. *Systematic Zoology*, *4*, 1–20.

Irisarri, I., Singh, P., Koblmüller, S., Torres-Dowdall, J., Henning, F., Franchini, P., Fischer, C., Lemmon, A. R., Lemmon, E. M., Thallinger, G. G., Sturmbauer, C., & Meyer, A. (2018). Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. *Nature Communications*, *9*, 3159.

Koblmüller, S., Eigner, E., Salzburger, W., Obermüller, B., Sefc, K. M., & Sturmbauer, C. (2011). Separated by sand, fused by dropping water: Habitat barriers and fluctuating water levels steer the evolution of rock-dwelling cichlid populations in Lake Tanganyika. *Molecular Ecology*, *20*, 2272–2290.

Koblmüller, S., Nevado, B., Makasa, L., Van Steenberge, M., Vanhove, M. P., Verheyen, E., Sturmbauer, C., & Sefc, K. M. (2017). Phylogeny and phylogeography of Altolamprologus: Ancient introgression and recent divergence in a rock-dwelling Lake Tanganyika cichlid genus. *Hydrobiologia*, *791*, 35–50.

Koblmüller, S., Sefc, K. M., & Sturmbauer, C. (2008). The Lake Tanganyika cichlid species assemblage: Recent advances in molecular phylogenetics. *Hydrobiologia*, *615*, 5–20.

Kocher, T. D. (2004). Adaptive evolution and explosive speciation: The cichlid fish model. *Nature Reviews Genetics*, *5*, 288–298.

Kohda, M., Yanagisawa, Y., Sato, T., Nakaya, K., Niimura, Y., Matsumoto, K., & Ochi, H. (1996). Geographical colour variation in cichlid fishes at the southern end of Lake Tanganyika. *Environmental Biology of Fishes*, *45*, 237–248.

Köhler, C., Scheid, O. M., & Erilova, A. (2010). The impact of the triploid block on the origin and evolution of polyploid plants. *Trends in Genetics*, *26*, 142–148.

Kong, S., & Kubatko, L. S. (2021). Comparative performance of popular methods for hybrid detection using genomic data. *Systematic Biology*, *70*, 891–907.

Konings, A. (2015). *Tanganyika cichlids in their natural habitat* (3rd ed.). Cichlid Press.

Kornfield, I., & Smith, P. F. (2000). African cichlid fishes: Model systems for evolutionary biology. *Annual Review of Ecology, Evolution, and Systematics*, *31*, 163–196.

Korunes, K. L., & Samuk, K. (2021). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*, 1359–1368.

Kulmuni, J., Butlin, R. K., Lucek, K., Savolainen, V., & Westram, A. M. (2020). Towards the completion of speciation: The evolution of reproductive isolation beyond the first barriers. *Philosophical Transactions of the Royal Society B*, *375*, 20190528.

Larson, E. L., Vanderpool, D., Sarver, B. A., Callahan, C., Keeble, S., Provencio, L. L., Kessler, M. D., Stewart, V., Nordquist, E., Dean, M. D., & Good, J. M. (2018). The evolution of polymorphic hybrid incompatibilities in house mice. *Genetics*, *209*, 845–859.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*, 2987–2993.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Liu, X., & Fu, Y. X. (2015). Exploring population size changes using SNP frequency spectra. *Nature Genetics*, *47*, 555–559.

Liu, X., & Fu, Y. X. (2020). Stairway plot 2: Demographic history inference with folded SNP frequency spectra. *Genome Biology*, *21*, 13059–14020.

Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., & Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, *2*, 1940–1955.

Mandeville, E. G., Parchman, T. L., McDonald, D. B., & Buerkle, C. A. (2015). Highly variable reproductive isolation among pairs of Catostomus species. *Molecular Ecology*, *24*, 1856–1872.

Mandeville, E. G., Parchman, T. L., Thompson, K. G., Compton, R. I., Gelwicks, K. R., Song, S. J., & Buerkle, C. A. (2017). Inconsistent reproductive isolation revealed by interactions between Catostomus fish species. *Evolution Letters*, *1*, 255–268.

Mandeville, E. G., Walters, A. W., Nordberg, B. J., Higgins, K. H., Burckhardt, J. C., & Wagner, C. E. (2019). Variable hybridization outcomes in trout are predicted by historical fish stocking and environmental context. *Molecular Ecology*, *28*, 3738–3755.

Marth, G. T., Czabarka, E., Murvai, J., & Sherry, S. T. (2004). The allele frequency Spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, *166*, 351–372.

Matute, D. R., & Cooper, B. S. (2021). Comparative studies on speciation: 30 years since Coyne and Orr. *Evolution*, *75*, 764–778.

McCune, A. R. (1997). How fast is speciation: Molecular, Geological and phylogenetic evidence from adaptive radiations of fishes. In T. J. Givnish & K. J. Sytsma (Eds.), *Molecular evolution and adaptive radiation* (pp. 585–610). Cambridge Univeresity Press.

McFarlane, S. E., & Pemberton, J. M. (2019). Detecting the true extent of introgression during anthropogenic hybridization. *Trends in Ecology and Evolution*, *34*, 315–326.

McGlue, M. M., Lezzar, K. E., Cohen, A. S., Russell, J. M., Tiercelin, J. J., Felton, A. A., Mbede, E., & Nkotagu, H. H. (2008). Seismic records of late Pleistocene aridity in Lake Tanganyika, tropical East Africa. *Journal of Paleolimnology*, *40*, 635–653.

McIntyre, P. B., Michel, E., France, K., Rivers, A., Hakizimana, P., & Cohen, A. S. (2005). Individual- and assemblage-level effects of anthropogenic sedimentation on snails in Lake Tanganyika. *Conservation Biology*, *19*, 171–181.

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, *8*, 14363.

Meier, J. I., Sousa, V. C., Marques, D. A., Selz, O. M., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Demographic modelling with whole-genome data reveals parallel origin of similar Pundamilia cichlid species after hybridization. *Molecular Ecology*, *26*, 123–141.

Momigliano, P., Florin, A. B., & Merilä, J. (2021). Biases in demographic modeling affect our understanding of recent divergence. *Molecular Biology and Evolution*, *38*, 2967–2985.

Nevado, B., Fazalova, V., Backeljau, T., Hanssens, M., & Verheyen, E. (2011). Repeated unidirectional introgression of nuclear and mitochondrial DNA between four congeneric Tanganyikan cichlids. *Molecular Biology and Evolution*, *28*, 2253–2267.

Nosil, P., Harmon, L. J., & Seehausen, O. (2009). Ecological explanations for (incomplete) speciation. *Trends in Ecology and Evolution*, *24*, 145–156.

Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., & Buerkle, C. A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, *21*, 2991–3005.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*, 1065–1093.

Patton, A. H., Margres, M. J., Stahlke, A. R., Hendricks, S., Lewallen, K., Hamede, R. K., Ruiz-Aravena, M., Ryder, O., McCallum, H. I., Jones, M. E., Hohenlohe, P. A., & Storfer, A. (2019). Contemporary demographic reconstruction methods are robust to genome assembly quality: A case study in Tasmanian devils. *Molecular Biology and Evolution*, *36*, 2906–2921.

Pinho, C., & Hey, J. (2010). Divergence with gene flow: Models and data. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 215–230.

Price, T. (2008). *Speciation in birds*. Roberts and Company.

R Core Team (2021) R: A language and environment for statistical computing.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*, 489–494.

Rometsch, S. J., Torres-Dowdall, J., & Meyer, A. (2020). Evolutionary dynamics of pre- and postzygotic reproductive isolation in cichlid fishes. *Philosophical Transactions of the Royal Society B*, *375*, 20190535.

Ronco, F., Matschiner, M., Böhne, A., Boila, A., Büscher, H. H., El Taher, A., Indermaur, A., Malinsky, M., Ricci, V., Kahmen, A., Jentoft, S., & Salzburger, W. (2021). Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature*, *589*, 76–81.

Scholz, C. A., King, J. W., Ellis, G. S., Swart, P. K., Stager, J. C., & Colman, S. M. (2003). Paleolimnology of Lake Tanganyika, East Africa, over the past 100 kyr. *Journal of Paleolimnology*, *30*, 139–150.

Seehausen, O. (2015). Process and pattern in cichlid radiations - inferences for understanding unusually high rates of evolutionary diversification. *New Phytologist*, *207*, 304–312.

Seehausen O, Magalhaes IS (2010) Geographical mode and evolutionary mechanism of ecological speciation in cichlid fish. In: Grant PR, Grant BR, (eds.), *In search of the causes of evolution: From field observations to mechanisms*, January 2010, (pp. 282–308), Princeton University Press.

Seehausen, O., Terai, Y., Magalhaes, I. S., Carleton, K. L., Mrosso, H. D., Miyagi, R., Van Der Sluijs, I., Schneider, M. V., Maan, M. E., Tachida, H., Imai, H., & Okada, N. (2008). Speciation through sensory drive in cichlid fish. *Nature*, *455*, 620–626.

Seehausen, O., & van Alphen, J. J. (1999). Can sympatric speciation by disruptive sexual selection explain rapid evolution of cichlid diversity in Lake Victoria? *Ecology Letters*, *2*, 262–271.

Seehausen, O., van Alphen, J. J. M., & Witte, F. (1997). Cichlid fish diversity threatened by eutrophication that curbs sexual selection. *Science*, *277*, 1808–1811.

Sefc, K. M., Hermann, C. M., Steinwender, B., Brindl, H., Zimmermann, H., Mattersdorfer, K., Postl, L., Makasa, L., Sturmbauer, C., & Koblmüller, S. (2015). Asymmetric dominance and asymmetric mate choice oppose premating isolation after allopatric divergence. *Ecology and Evolution*, *5*, 1549–1562.

Sefc, K. M., Mattersdorfer, K., Hermann, C. M., & Koblmüller, S. (2017). Past lake shore dynamics explain present pattern of unidirectional introgression across a habitat barrier. *Hydrobiologia*, *791*, 69–82.

Sefc, K. M., Mattersdorfer, K., Ziegelbecker, A., Neuhüttler, N., Steiner, O., Goessler, W., & Koblmüller, S. (2017). Shifting barriers and phenotypic diversification by hybridisation. *Ecology Letters*, *20*, 651–662.

Shastry, V., Adams, P. E., Lindtke, D., Mandeville, E. G., Parchman, T. L., Gompert, Z., & Buerkle, C. A. (2021). Model-based genotype and ancestry estimation for potential hybrids with mixed-ploidy. *Molecular Ecology Resources*, *21*, 1434–1451.

Simpson GL (2021) Gratia: Graceful 'ggplot'-based graphics and other functions for GAMs fitted using 'mgcv'.

Stelkens, R. B., Schmid, C., & Seehausen, O. (2015). Hybrid breakdown in cichlid fish. *PLoS One*, *10*, e0127207.

Stelkens, R. B., Young, K. A., & Seehausen, O. (2010). The accumulation of reproductive incompatibilities in African cichlid fish. *Evolution*, *64*, 617–633.

Sturmbauer, C. (1998). Explosive speciation in cichlid fishes of the African Great Lakes: A dynamic model of adaptive radiation. *Journal of Fish Biology*, *53*, 18–36.

Sturmbauer, C., Hainz, U., Baric, S., Verheyen, E., & Salzburger, W. (2003). Evolution of the tribe Tropheini from Lake Tanganyika: Synchronized explosive speciation producing multiple evolutionary parallelism. *Hydrobiologia*, *500*, 51–64.

Svardal, H., Quah, F. X., Malinsky, M., Ngatunga, B. P., Miska, E. A., Salzburger, W., Genner, M. J., Turner, G. F., & Durbin, R. (2020). Ancestral hybridization facilitated species diversification in the

Lake Malawi cichlid fish adaptive radiation. *Molecular Biology and Evolution*, *37*, 1100–1113.

Svardal, H., Salzburger, W., & Malinsky, M. (2021). Genetic variation and hybridization in evolutionary radiations of cichlid fishes. *Annual Review of Animal Biosciences*, *9*, 55–79.

Taylor, M. I., Rüber, L., & Verheyen, E. (2001). Microsatellites reveal high levels of population substructuring in the speciespoor Eretmodine cichlid lineage from Lake Tanganyika. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*, 803–808.

Wagner, C. E., & McCune, A. R. (2009). Contrasting patterns of spatial genetic structure in sympatric rock-dwelling cichlid fishes. *Evolution*, *63*, 1312–1326.

Wagner, C. E., McCune, A. R., & Lovette, I. J. (2012). Recent speciation between sympatric Tanganyikan cichlid colour morphs. *Molecular Ecology*, *21*, 3283–3292.

Willing, E. M., Dreyer, C., & van Oosterhout, C. (2012). Estimates of genetic differentiation measured by Fst do not necessarily require large sample sizes when using many SNP markers. *PLoS One*, *7*, 1–7.

Winkelmann, K., Rüber, L., & Genner, M. J. (2017). Lake level fluctuations and divergence of cichlid fish ecomorphs in Lake Tanganyika. *Hydrobiologia*, *791*, 21–34.

Wood, S. N. (2017). *Generalized additive models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC.

Zheng, Y., & Janke, A. (2018). Gene flow analysis method, the D-statistic, is robust in a wide parameter space. *BMC Bioinformatics*, *19*, 1–19.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lewanski, A. L., Golcher-Benavides, J., Rick, J. A., & Wagner, C. E. (2022). Variable hybridization between two Lake Tanganyikan cichlid species in recent secondary contact. *Molecular Ecology*, *31*, 5041–5059. https://doi.org/10.1111/mec.16636