Received: 27 December 2017 | Revised: 30 April 2018 | Accepted: 14 May 2018

DOI: 10.1111/mec.14784

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

Speciation in sympatry with ongoing secondary gene flow and a potential olfactory trigger in a radiation of Cameroon cichlids

Jelmer W. Poelstra[1,2] | Emilie J. Richards[1] | Christopher H. Martin[1]

[1]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

[2]Department of Biology, Duke University, Durham, North Carolina

Correspondence
Christopher H. Martin, Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3280.
Email: chmartin@unc.edu

Funding information
This study was funded by a National Geographic Society Young Explorer's Grant, a Lewis and Clark Field Research grant from the American Philosophical Society, NSF CAREER Award DEB-1749764, and the University of North Carolina at Chapel Hill to CHM.

Abstract

The process of sympatric speciation in nature remains a fundamental unsolved problem. Cameroon crater lake cichlid radiations were long regarded as one of the most compelling examples; however, recent work showed that their origins were more complex than a single colonization event followed by isolation. Here, we performed a detailed investigation of the speciation history of a radiation of *Coptodon* cichlids from Lake Ejagham, Cameroon, using whole-genome sequencing data. The existence of the Lake Ejagham *Coptodon* radiation is remarkable as this 0.5 km^2 lake offers limited scope for divergence across a shallow depth gradient, disruptive selection is currently weak, and the species are sexually monochromatic. We infer that Lake Ejagham was colonized by *Coptodon* cichlids soon after its formation 9,000 years ago, yet speciation occurred only in the last 1,000–2,000 years. We show that secondary gene flow from a nearby riverine species has been ongoing, into ancestral as well as extant lineages, and we identify and date river-to-lake admixture blocks. One block contains a cluster of olfactory receptor genes that introgressed near the time of the first speciation event and coincides with a higher overall rate of admixture. Olfactory signalling is a key component of mate choice and species recognition in cichlids. A functional role for this introgression event is consistent with previous findings that sexual isolation appears much stronger than ecological isolation in Ejagham *Coptodon*. We conclude that speciation in this radiation took place in sympatry, yet may have benefited from ongoing riverine gene flow.

KEYWORDS
cichlids, coalescent, demographics, genomics, population genetics, sympatric speciation, whole-genome sequencing

1 | INTRODUCTION

Speciation in the absence of geographic barriers is a powerful demonstration that divergent selection can overcome the homogenizing effects of gene flow and recombination (Arnegard & Kondrashov, 2004; Coyne & Orr, 2004; Turelli, Barton, & Coyne, 2001). While it was long thought that sympatric speciation was very unlikely to take place in nature, the last 25 years have seen a proliferation of empirical examples as well as theoretical models that support its plausibility (Barluenga, Stölting, Salzburger, Muschick, & Meyer, 2006; Berlocher & Feder, 2002; Bolnick & Fitzpatrick, 2007; Hadid et al., 2013, 2014; Kautt, Machado-Schiaffino, & Meyer, 2016; Kautt, Machado-Schiaffino, Torres-Dowdall, & Meyer, 2016; Malinsky et al., 2015; Savolainen et al., 2006; Sorenson, Sefc, & Payne, 2003).

However, it is exceptionally hard to demonstrate that speciation has been sympatric in any given empirical case. One of the most challenging criteria is the absence of a historical phase of geographic isolation (Coyne & Orr, 2004). This can be ruled out most compellingly in

cases where multiple endemic species are found in environments that are (a) small and homogeneous, such that geographic isolation within the environment is unlikely, and (b) severely isolated, such that a single colonization likely produced the lineage that eventually diversified. Early molecular studies used a single locus or limited genomic data to establish monophyly of sympatric species in isolated environments, such as crater lakes (Barluenga et al., 2006; Schliewen, Tautz, & Pääbo, 1994) and oceanic islands (Savolainen et al., 2006). Genome-wide sequencing data can now be used to rigorously test whether or not extant species contain ancestry from secondary gene flow into the focal environment. Strikingly, evidence for such ancestry has recently been found in all seven crater lake cichlid radiations examined so far (Kautt, Machado-Schiaffino, Meyer et al., 2016; Malinsky et al., 2015; Martin, Cutler et al., 2015). However, whereas a complete lack of secondary gene flow would rule out a role for geographic isolation outside of the focal environment, the presence of secondary gene flow does not exclude the possibility of sympatric speciation.

If secondary gene flow into a pair or radiation of sympatric species has taken place, a key question is whether secondary gene flow played a role in the speciation process (Martin, Cutler et al., 2015). Speciation would still be functionally sympatric if genetic variation introduced by secondary gene flow did not contribute to speciation (Malinsky et al., 2015; Martin, Cutler et al., 2015). Secondary gene flow could also counteract speciation in the focal environment, for instance via hybridization with both incipient species during a speciation event. On the other hand, there are several ways in which secondary gene flow may be a key part of speciation (Kautt, Machado-Schiaffino, Meyer et al., 2016; Martin, Cutler et al., 2015). For instance, secondary colonization may involve a partially reproductively isolated population, in which case any resulting speciation event would have had a crucial allopatric phase. Second, the introduction of novel genetic variation and novel allelic combinations may promote speciation more generally, for example, via the formation of a hybrid swarm (Kautt, Machado-Schiaffino, Meyer et al., 2016; Meier, Marques, Wagner, Excoffier, & Seehausen, 2018; Meier et al., 2017; Seehausen, 2004), transgressive segregation (Kagawa & Takimoto, 2018), or adaptive introgression (Anderson, 1949; Feder et al., 2003; Heliconius Genome Consortium, 2012; Pardo-Diaz et al., 2012; Richards & Martin, 2017; Stankowski & Streisfeld, 2015).

Establishing or rejecting a causal role of secondary gene flow in speciation first requires identifying the timing and extent of gene flow and the donor and recipient populations. A role for secondary gene flow would be supported if divergence rapidly followed a discrete admixture event (Kautt, Machado-Schiaffino, Meyer et al., 2016); whereas if gene flow took place only after the onset of divergence, such a role would seem unlikely. Genomic data can also be used to identify segments of the genome that have experienced admixture and to examine whether these contain genes that may have been important in speciation (Lamichhaney et al., 2015; Meier et al., 2017; Richards & Martin, 2017).

Four radiations of cichlids in three isolated lakes in Cameroon (Schliewen & Klee, 2004; Schliewen et al., 1994, 2001) are one of the most widely accepted examples of sympatric speciation. Two of the lakes are crater lakes, while the third, Lake Ejagham, is now suspected to be the result of a meteor impact (Stager et al., 2017). Given their small size and uniform topology, geographic isolation within these lakes is highly unlikely (Schliewen et al., 1994). Moreover, species within the radiations were shown to be monophyletic relative to riverine outgroups based on mtDNA (for all four of the radiations, Schliewen et al., 1994) and AFLPs (for one radiation, Schliewen & Klee, 2004), which was interpreted to mean that each radiation is derived from a single colonization. However, using RADseq data, Martin, Cutler et al. (2015) recently found evidence for secondary admixture with nearby riverine populations in all four radiations.

Despite being the second smallest (0.49 km^2: Seehausen, 2006) and one of the youngest lakes (ca. 9 kya, Stager et al., 2017) containing endemic cichlids, Lake Ejagham contains two independent endemic cichlid radiations: two species of *Sarotherodon* cichlids (Neumann, 2011) and potentially four species of *Coptodon* cichlids (Dunz & Schliewen, 2010). The existence of these radiations is even more remarkable given that they are an exception to the two best predictors of endemic radiation in African cichlids: overall lake depth and sexual dichromatism (Wagner, Harmon, & Seehausen, 2012). Lake Ejagham is shallow (maximum depth of 18 m, Schliewen et al., 2001), and three *Coptodon* species plus one *Sarotherodon* species coexist within 0–2 m depth (Martin, 2013). Furthermore, neither *Coptodon* nor *Sarotherodon* lineages are sexually dichromatic within Lake Ejagham. Nonetheless, Ejagham *Coptodon* differ most strongly in sexual rather than ecological characters (Martin, 2012, 2013), and assortative mating appears to be a stronger reproductive isolating barrier than ecological disruptive selection (Martin, 2012), which is noteworthy as speciation in Cameroon lakes is generally considered to be ecologically driven (Coyne & Orr, 2004).

Some of the clearest evidence for admixture in Martin, Cutler et al. (2015) was in the Ejagham *Coptodo*n radiation. The occurrence of secondary gene flow from riverine populations could be a key piece in the puzzling occurrence of the Lake Ejagham radiations and may have initiated speciation despite limited disruptive ecological selection. Here, we use whole-genome resequencing of three species of Ejagham *Coptodon* and two riverine outgroups to provide a comprehensive picture of the history of secondary gene flow and its riverine sources and identify admixed portions of the genome.

2 | METHODS

2.1 | Sampling

Sampling efforts and procedures have been described previously in Martin, Cutler et al. (2015). Here, we sampled breeding individuals displaying reproductive coloration from three species of *Coptodon* (formerly *Tilapia*) that are endemic to Lake Ejagham in Cameroon: *Coptodon fusiforme* (n = 3), *C. deckerti* (n = 2) and *C. ejagham* (n = 2). We additionally used samples from closely related riverine species from the nearby Cross River whose ancestors likely colonized Lake

Ejagham: *C. guineensis* (*n* = 2) at Nguti, 65 km from Lake Ejagham, and an undescribed taxon, *C.* sp. "*Mamfé*" (Keijman, 2010) (*n* = 1), at Mamfé, 37 km from Lake Ejagham. Finally, we sampled a closely related outgroup species, *C. kottae*, from crater lake Barombi ba Kotto (145 km from Lake Ejagham), and a distantly related outgroup species, *Sarotherodon galilaeus* (*n* = 3), from the Cross River at Mamfé Cichlids were caught by seine or cast-net in 2010 and euthanized in an overdose of buffered MS-222 (Finquel, Inc.) following approved protocols from University of California, Davis Institutional Animal Care and Use Committee (#17455) and University of North Carolina Animal Care and Use Committee (#15-179.0), and stored in 95%–100% ethanol or RNAlater (Ambion, Inc.) in the field.

2.2 | Genome sequencing and variant calling

DNA was extracted from muscle tissue using DNeasy Blood and Tissue kits (Qiagen, Inc.) and quantified on a Qubit 3.0 fluorometer (Thermo Fisher Scientific, Inc.). Genomic libraries were prepared using the automated Apollo 324 system (WaterGen Biosystems, Inc.) at the Vincent J. Coates Genomic Sequencing Center (QB3). Samples were fragmented using Covaris sonication, barcoded with Illumina indices and quality checked using a Fragment Analyzer (Advanced Analytical Technologies, Inc.). Nine to twelve samples were pooled in four different libraries for 150PE sequencing on four lanes of an Illumina Hiseq4000.

We mapped raw sequencing reads in FASTQ format to the *Oreochromis niloticus* genome assembly (version 1.1, https://www.ncbi.nlm.nih.gov/assembly/GCF_000188235.2/, Brawand et al., 2014) with BWA-MEM (version 0.7.15, Li, 2013). Using PICARD Tools (version 2.10.3, http://broadinstitute.github.io/picard), the resulting .sam files were sorted (SORTSAM tool), and the resulting .bam files were marked for duplicate reads (MARKDUPLICATES tool) and indexed (BUILDBAMINDEX tool). SNPs were called using the HaplotypeCaller program in the Genome Analysis Toolkit (GATK; DePristo et al., 2011), following the GATK Best Practices guidelines (Van der Auwera et al., 2013, https://software.broadinstitute.org/gatk/best-practices/). Since no high-quality known variants are available to recalibrate base quality and variant scores, SNPs were called using hard filtering in accordance with the GATK guidelines (DePristo et al., 2011; Van der Auwera et al., 2013: QD < 2.0, MQ < 40.0, FS > 60.0, SOR > 3.0, MQRankSum < −12.5, ReadPosRankSum < −8.0). SNPs that did not pass these filters were removed from the resulting VCF files using VCFTOOLS (version 0.1.14, Danecek et al., 2011; using "–remove-filtered-all" flag) as were SNPs that differed from the reference but not among focal samples (using "max-non-ref-af 0.99" in VCFTOOLS) and SNPs with more than two alleles (using "-m2 -M2" flags in BCFTOOLS, version 1.5 (Li, 2011)). Genotypes with a genotype quality below 20 and depth below 5 were set to missing (using "–minGQ" and "–minDP" flags in VCFTOOLS, respectively), and sites with more than 50% missing data were removed (using "–max-missing" flag in VCFTOOLS). Our final dataset consisted of 15,523,738 SNPs with a mean sequencing depth of 11.82 (range: 7.20–16.83) per individual.

2.3 | Phylogenetic trees, networks and genetic structure

We employed several approaches to estimate relationships among the three species in the *Coptodon* Ejagham radiation and the two riverine *Coptodon* taxa. These analyses were repeated for four outgroup configurations: (i) no outgroup (unrooted trees), using (ii) only *C. kottae*, (iii) only *S. galilaeus* and (iv) both *C. kottae* and *S. galilaeus* as outgroups. Only sites with less than 10% missing data were used for phylogenetic reconstruction.

Using the GTR-CAT maximum-likelihood model without rate heterogeneity, as implemented in RAXML (version 8.2.10, Stamatakis, 2014), we inferred phylogenies for all SNPs concatenated, as well as separately for each 100-kb window with at least 250 variable sites ("gene trees"). This resulted in sets of 1,532–2,559 trees, depending on the outgroup configuration. Next, rooted gene trees were used, to compute Internode Confidence All (ICA) scores (Salichos , Stamatakis, & Rokas, 2014, using the "-L MR" flag in RAXML) for each of the nodes of the whole-genome trees. Rooted gene trees were also used to construct species trees in PHYLONET (version 3.6.1, Than, Ruths, & Nakhleh, 2008; Wen, Yu, Zhu, & Nakhleh, 2018) using the minimize deep coalescence criterion (Than & Nakhleh, 2009; "Infer_ST_MDC" command) and maximum likelihood ("Infer_Network_ML" command with zero reticulations), and using a maximum pseudolikelihood method implemented in MP-EST (version 1.5, Liu, Yu, & Edwards, 2010). Finally, we used ASTRAL (version 2.5.5, Mirarab et al., 2014) to infer species trees from unrooted gene trees.

To visualize patterns of genealogical concordance and discordance, we computed a phylogenetic network using the NeighborNet method (Bryant & Moulton, 2004) implemented in SPLITSTREE (version 4.14.4, Huson & Bryant, 2006), using all SNPs.

We used the machine-learning program SAGUARO (Zamani et al., 2013) to determine the dominant topology across the genome and calculate the percentages of the genome that supported specific relationships, such as monophyly of the Ejagham *Coptodon* radiation. SAGUARO combines a hidden Markov model with a self-organizing map to characterize local phylogenetic relationships among individuals without requiring a priori hypotheses about the relationships. This method infers local relationships among individuals in the form of genetic distance matrices and assigns segments across the genomes to these topologies. These genetic distance matrices can then be transformed into neighbourhood joining trees to visualize patterns of evolutionary relatedness across the genome. To be comprehensive in our search, we allowed SAGUARO to propose 31 topologies for the genome, but otherwise applied default parameters. We investigated the effect of the number of proposed topologies on the proportion of genomes assigned to our two categories and found that the percentages were robust after 20 proposed topologies with increasingly smaller percentages of the genome being assigned to new additional topologies.

2.4 | Genomewide tests for admixture

We tested for admixture between the two riverine species and the three Lake Ejagham species using several statistics based on patterns of derived-allele sharing among these species. We used the ADMIXTOOLS (version 4.1, Patterson et al., 2012) suite of programs to compute four-taxon *D*-statistics ("ABBA-BABA tests," *qpDstat* program) and a five-taxon f_4-ratio test (*qpF4ratio* program), and the software DFOIL (release 2017-06-14, http://www.github.com/jbpease/dfoil, Pease & Hahn, 2015) to compute five-taxon D_{FOIL} statistics. For all analyses, we used *S. galilaeus* as the outgroup species.

Given a topology (((P1, P2), P3), O), *D* can identify admixture between either P1 or P2 on one hand, and P3 on the other based on the relative occurrence of ABBA and BABA patterns. First, we computed *D*-statistics to test for admixture between *C. guineensis* (P1) or *C.* sp. *"Mamfé"* (P2) and any Lake Ejagham species (P3). Given that all three of these comparisons indicated admixture between *C.* sp. *"Mamfé"* and Lake Ejagham species (Figure 3a), we next tested whether there was evidence for differential admixture from *C.* sp. *"Mamfé"* among the three Ejagham *Coptodon* species, using the three possible pairs of Lake Ejagham species as P1 and P2, and *C.* sp. *"Mamfé" as P3*.

Another way to test for differential *C.* sp. *"Mamfé"* admixture among Ejagham *Coptodon* species is using f_4-ratio tests, wherein taxon "*X*" is considered putatively admixed, containing ancestry proportion α from the branch leading to P2 (after its divergence from taxon P1), and ancestry proportion $\alpha - 1$ from the branch leading to taxon P3. Given the constraints imposed by the topology of our phylogeny, we could only test for admixed ancestry of either *C. deckerti* or *C. ejagham* with *C.* sp. *"Mamfé,"* after divergence of the *C. deckerti–C. ejagham* ancestor from *C. fusiforme*. Testing for admixed ancestry of *C. fusiforme* using an f_4-ratio test would merely produce a lower bound of α (see Mailund, 2014), while we were instead interested in an estimate or upper bound on α, as our null hypothesis was $\alpha = 1$; that is, *C. fusiforme* has ancestry only from the *C. deckerti–C. ejagham* ancestor. Furthermore, the two possible f_4-ratio tests (one with *C. deckerti* and the other with *C. ejagham* as the possibly admixed population) necessarily produce mirrored results, and we therefore present the results for the test that resulted in a significant contribution by *C.* sp. *"Mamfé"* (i.e., an estimate of α lower than 1).

The five-taxon D_{FOIL} statistics enable testing of the timing, and in some cases, direction of introgression in a symmetric phylogeny with two pairs of taxa with a sister relationship within the provided phylogeny, and an outgroup. Given our six-taxon phylogeny, we performed this test for three sets of five species, each with a unique combination of two of the three Ejagham *Coptodon* species as one species pair (P1 and P2), and *C. guineensis* and *C.* sp. *"Mamfé"* as the second species pair (P3 and P4; the outgroup again being *S. galilaeus*). The test involves the computation of four D_{FOIL} statistics (D_{FO}, D_{IL}, D_{FI} and D_{OL}), each essentially performing a three-taxon comparison. The combination of results for these statistics can inform whether introgression predominantly occurred among any of the four ingroup extant taxa, in which case the direction of introgression can also be inferred (e.g., P1→P3), or among an extant taxon and the ancestor of the other species pair, in which case the direction of introgression cannot be inferred (e.g., P1↔P3, P4). Unlike *D* and f_d statistics, D_{FOIL} statistics by default also include counts of patterns where only a single taxon has the derived allele (e.g., BAAAA), under the assumption of similar branch lengths across taxa. When this assumption is violated, DFOIL can be run in "dfoilalt" mode, thereby excluding single derived-allele counts (Pease & Hahn, 2015). As we observed significantly fewer single derived-allele sites for *C.* sp. *"Mamfé"* than for *C. guineensis*, we ran DFOIL in "dfoilalt" mode at a significance level of 0.001.

2.5 | Inference of demographic history with G-PHOCS

For a detailed reconstruction of the demographic history of Ejagham *Coptodon* and the two closely related riverine species, we used the program Generalized Phylogenetic Coalescent Sampler (G-PHOCS, version 1.3, Gronau, Hubisz, Gulko, Danko, & Siepel, 2011). G-PHOCS implements a coalescent-based approach using Markov chain Monte Carlo (MCMC) to jointly infer population sizes, divergence times, and optionally migration rates among extant as well as ancestral populations, given a predefined population phylogeny. To infer migration rates, one or more unidirectional migration bands can be added to the model, each between a pair of populations that overlap in time. G-PHOCS can thus infer the timing of migration within the bounds presented by the population splits in the phylogeny.

As input, G-PHOCS expects full sequence data for any number of loci. As G-PHOCS models the coalescent process without incorporating recombination, it assumes no recombination within loci, and free recombination between loci. Following several other studies (Choi et al., 2017; Gronau et al., 2011; Hung et al., 2014; McManus et al., 2015), we picked 1 kb loci separated by at least 50 kb. Following (Gronau et al., 2011), loci were selected not to contain the following classes of sites within the *O. niloticus* reference genome—that is, rather than being simply masked, these sites were not allowed to occur in input loci: (a) hard-masked (N) or soft-masked (lowercase bases) sites in the publicly available genome assembly; (b) sites that were identified to be prone to ambiguous read mapping using the program SNPable (Li, 2009, using $k = 50$ and $r = 0.5$ and excluding rankings 0 and 1); and (c) any site within an exon or <500 bp from an exon boundary. Furthermore, loci were chosen to contain no more than 25% missing data (uncalled and masked genotypes). Using these selection procedures, a total of 2,618 loci were chosen using custom scripts (available at https://github.com/jelmerp/EjaghamCoptodon/gphocs) and a VCF to Fasta conversion tool (Bergey, 2012).

Prior distributions for demographic parameters are specified in G-PHOCS using α and β parameters of a gamma distribution. We determined the mean of the prior distribution (α/β) for each parameter using a number of preliminary runs, while keeping the variance (α/β^2) large following (Gronau et al., 2011) to minimize the impact of the prior on the posterior (see Supporting information Table S10 for all G-PHOCS settings). Preliminary runs confirmed that regardless of the

choice of the prior mean, MCMC runs converged on similar posterior distributions.

For each combination of migration bands (see below), we performed four replicate runs. Each G-PHOCS run was allowed to continue for a week on 8–12 cores on a single 2.93 GHz compute node of the UNC Killdevil computing cluster, resulting in runs with 1–1.5 million iterations. The first 250,000 iterations were discarded as burn-in, and the remaining iterations were sampled 1 in every 50 iterations. Convergence, stationarity and mixing of MCMC chains were assessed using TRACER (version 1.6.0, Rambaut, Suchard, Xie, & Drummond, 2014).

Because the total number of possible migration bands in a six-taxon phylogeny is prohibitively high for effective parameter inference and because comparing model fit across different G-PHOCS runs (e.g., with different migration bands) is currently not possible (I. Gronau, personal communication), we took the following strategy. Our primary focus was on testing migration bands from *C.* sp. *"Mamfé"* ("Mam") and *C. guineensis* ("Gui") to the Lake Ejagham *Coptodon* species and their ancestors: *C. deckerti* ("Dec"), *C. ejagham* ("Eja"), *C. fusiforme* ("Fus"), "DE" (the ancestor to Dec and Eja) and "DEF" (the ancestor to DE and Fus). We first performed runs each with a single one of these migration bands. As all migration bands from *C.* sp. *"Mamfé"* had nonzero migration rates, we next performed runs with all of these migration bands at once. However, in those runs we observed failures to converge, higher variance in parameter estimates, and the dropping to zero of rates of migration to the ancestral Lake Ejagham lineage (see Figure 4). The latter is surprising given that for single-band runs, this migration rate was the highest inferred, and is also in sharp contrast to other analyses that show much stronger support for migration to the ancestral lineage than to extant species. While we suspect that runs with all migration bands have poor performance due to the number of parameters, runs with single migration bands may be prone to overestimation of the migration rate. We therefore also performed runs with migration bands either to all three extant species or to both ancestral lineages (see Supporting information Figure S5), and report results for all of these run types separately. Finally, we performed runs with no migration bands. We did not examine models with migration from the Ejagham radiation to neighbouring rivers because this is not relevant to sympatric speciation scenarios in this lake.

To convert the θ ($4 \times N_e \times \mu$) and τ ($T \times \mu$) parameters reported by G-PHOCS, which are scaled by the mutation rate, to population sizes N_e and divergence times T, we used a per year mutation rate μ of 7.1×10^{-9} as used in Kautt et al. (2016), based on a per-generation mutation rate of 7.5×10^{-9} estimated in stickleback (Guo, Chain, Bornberg-Bauer, Leder, & Merilä, 2013). We used a generation time of 1 year similar to East African cichlids and corresponding to observations of laboratory growth rates (although note that these species have rarely been bred in captivity). We converted the migration rate parameter m for a given migration band to several more readily interpretable statistics. First, the population migration rate ($2Nm$) is twice the number of migrants in the source population that arrived by migration from the target population, per generation. It is calculated using the value of θ for the target population ($2Nm_{s\to t} = m_{s\to t} \times \theta_t/4$), and as such it does not depend on an estimate of the mutation rate. Second, the proportion of migrants per generation is calculated by multiplying m by the mutation rate. Third, the "total migration rate" M (Gronau et al., 2011) can be interpreted as the probability that a locus in the target population has experienced migration from the source population, and is calculated by multiplying m by the time span of the migration band, which is the time window during which both focal populations existed ($M_{s\to t} = m_{s\to t} \times \tau_{s,t}$). Parameter estimates are presented as point estimates (mean across all retained iterations) with association 95% highest posterior density (HPD) values, calculated using the hpd() function in the R package TEACHINGDEMOS (version 2.10, Snow, 2016).

2.6 | Local admixture tests

To identify genomic regions with evidence for admixture between one of the riverine species and one or more of the Lake Ejagham species, we first computed the f_d statistic (Martin, Davey, & Jiggins, 2015) along sliding windows of 50 kb with a step size of 5 kb, using ABBABABA.py (Martin, 2015). The f_d statistic is a modified version of the Green et al. (2010) estimator of the proportion of introgression (f), and has been shown to outperform D for the detection of introgression in small genomic windows (Martin, Davey et al., 2015).

In the topology ((P1, P2), P3), O), f_d tests for introgression between P2 and P3. For each window, f_d was calculated for two types of configurations. First, those that can identify the source of any riverine admixture, using the two riverine species as P1 and P2 and a Lake Ejagham species as P3 (e.g., P1 = *C. guineensis*, P2 = *C.* sp. *"Mamfé"*, P3 = *C. ejagham*). Second, those that can identify differential admixture from a riverine species among two Lake Ejagham species (e.g., P1 = *C. deckerti*, P2 = *C. ejagham*, P3 = *C.* sp. *"Mamfé"*). As f_d only detects introgression between P2 and P3, f_d was also computed for every triplet with P1 and P2 swapped (see Supporting information Table S11 for a list of all triplets for which f_d was computed).

p-values for f_d were estimated by *Z*-transforming single-window f_d values based on a standard normal distribution, followed by multiple testing correction using the false discovery rate method (FDR, Benjamni & Hochberg, 1995), using a significance level of 0.05. Next, putative admixture blocks were defined by combining runs of significant f_d values that were consecutive or separated by at most three nonsignificant (FDR > 0.05) windows. Because any secondary admixture must have occurred within the last ~10 k years, after colonization of Lake Ejagham, true admixture blocks are expected to be large, and blocks of less than five total windows or with maximum f_d values below 0.5 were excluded from consideration. Therefore, detection of putative admixture blocks was limited to genomic scaffolds of at least 70 kb (i.e., 557 scaffolds or 97.40% of the assembled genome). Blocks indicating differential admixture with a riverine species among two Lake Ejagham species (in ingroup triplets with a pair of Lake Ejagham species as P1 and P2, and a riverine species as P3) were retained only when the riverine source of admixture could be distinguished in a direct comparison, by intersection with blocks

indicating differential admixture with a Lake Ejagham species among the two riverine species. For instance, a block indicating admixture between *C. deckerti* (P2) and *C.* sp. *"Mamfé"* (P3) in an ingroup triplet with *C. ejagham* as P1 (i.e., identifying differential admixture among two lake species) was only retained if it overlapped with an admixture block with *C. guineensis* as P1, *C.* sp. *"Mamfé"* as P2, and *C. deckerti* as P3 (i.e., identifying differential admixture among the riverine sources with the same lake species).

Putative admixture blocks as defined by f_d values were validated and aged using HYBRIDCHECK (Ward & van Oosterhout, 2016), using the same mutation rate as for our G-PHOCS analysis. HYBRIDCHECK identifies blocks that may have admixed between two sequences by comparing sequence similarity between triplets of individuals along sliding windows, and next estimates, for each block, the coalescent time between the two potentially admixed sequences. While HybridCheck can also discover admixture blocks ab initio, we employed it to test user-defined blocks with the "addUserBlock" method. Given that HYBRIDCHECK accepts triplets of individuals, f_d blocks detected in a given species triplet were tested twice in HYBRIDCHECK for that species triplet, each using a different individual of the admixed Lake Ejagham species. Blocks were retained when HYBRIDCHECK reported admixture between the same pair of individuals as the f_d statistic, and with a *p*-value smaller than 0.001 for both triplets of individuals. Our final set of "likely blocks" consisted of those with an estimated age smaller than the G-PHOCS point estimate (in runs with all possible migration bands from *C.* sp. *"Mamfé"*) of the divergence time between the Lake Ejagham ancestor ("DEF") and the riverine ancestor ("AU"), while "high-confidence blocks" were defined as those with the *upper* bound of the 95% confidence interval of the age estimate smaller than the *lower bound* of the 95% HPD of the divergence time estimate between DEF and AU (for whichever set of G-PHOCS runs, either with no, some or all migration bands from *C.* sp. *"Mamfé,"* had the lowest value for this parameter).

To characterize the patterns of admixture for these pairwise admixture blocks further, we calculated localized D_{FOIL} statistics for each. As these statistics depend on the occurrence of sufficient numbers of all possible four-taxon derived-allele frequency occurrence patterns among five taxa, these only produced results for a subset of blocks (for the same reason, we were not able to use these statistics for ab initio admixture block discovery along sliding windows). As we already established the presence of admixture for these blocks and performed these analyses to determine the pattern of admixture, we did not require significance for each D_{FOIL} statistic, but also considered it to be positive or negative if the statistic was more than half its maximum value and had at least 10 informative sites.

We also calculated F_{ST} and d_{xy} between each species pair along sliding windows of 50 kb with a step size of 5 kb, using popgenWindows.py (Martin, Davey et al., 2015).

2.7 | Gene ontology for admixture blocks

We assessed whether "high-confidence" admixture blocks were enriched for specific gene categories using Gene Ontology (GO) analyses. Entrez Gene gene identifiers were extracted by intersecting the genomic coordinates of admixture blocks with a GFF file containing the genome annotation for *O. niloticus* (Annotation Release 102, available at https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Oreochromis_niloticus/102/), and GO annotations for each gene were collected using the R/Bioconductor package biomaRt (Durinck, Spellman, Birney, & Huber, 2009). Next, GO enrichment analysis was carried out with the R/Bioconductor package GOSEQ (Young, Wakefield, Smyth, & Oshlack, 2010), using a flat probability weighting function, the Wallenius method for calculating enrichment scores, and correcting *p*-values for multiple testing using the false discovery rate method (FDR, Benjamni & Hochberg, 1995). GO terms were considered enriched for FDRs below 0.05.

3 | RESULTS

3.1 | Phylogeny of the Lake Ejagham *Coptodon* radiation

As a first step in revealing the speciation history of the Lake Ejagham *Coptodon* radiation (hereafter: Ejagham radiation), we took several approaches to infer the phylogenetic relationships among the three Lake Ejagham species *C. deckerti*, *C. ejagham* and *C. fusiforme*, as well as two closely related riverine species from the neighbouring Cross River drainage, *C. guineensis* and *C.* sp. *"Mamfé"* (Figure 1), *C. kottae*, a Cameroon crater lake endemic that did not diversify in situ, and the much more distantly related *Sarotherodon galilaeus*.

Maximum-likelihood (ML) trees based on concatenated genome-wide SNPs using RAXML with any of three outgroup configurations (only *C. kottae* / only *S. galilaeus* / both species) resulted in monophyly of Lake Ejagham species and a sister relationship between *C. deckerti* and *C. ejagham* with 100% bootstrap support (Figure 2a). However, inferences on whether one of the two riverine species is more closely related to Ejagham *Coptodon*, or the two are sister species, differed among outgroup configurations (Supporting information Figure S1). To further investigate the relationships among the two riverine species relative to Ejagham *Coptodon*, we estimated species trees from 100-kb gene trees using two different approaches. Species trees based on rooted gene trees using ML and the minimize deep coalescence (MDC) criterion in PHYLONET, as well as a species tree based on unrooted gene trees using ASTRAL, all indicated monophyly of the Ejagham radiation, and a sister relationship between *C. deckerti* and *C. ejagham* (Supporting information Figure S2).

We used two methods to more explicitly examine the prevalence of discordant phylogenetic patterns. In keeping with the results from phylogenetic trees, a phylogenetic network based on genomewide SNPs produced by SPLITSTREE showed limited discordance along the branch to the Ejagham *Coptodon* ancestor, with higher levels of discordance along the branch to the *C. deckerti*–*C. ejagham* ancestor and especially near the divergence of the riverine species (Figure 2b). Second, phylogenetic relationships along local segments of the genome grouped by the machine-learning approach *Saguaro* into 30 unrooted trees ("cacti") indicate that in 90.02% of the genome,

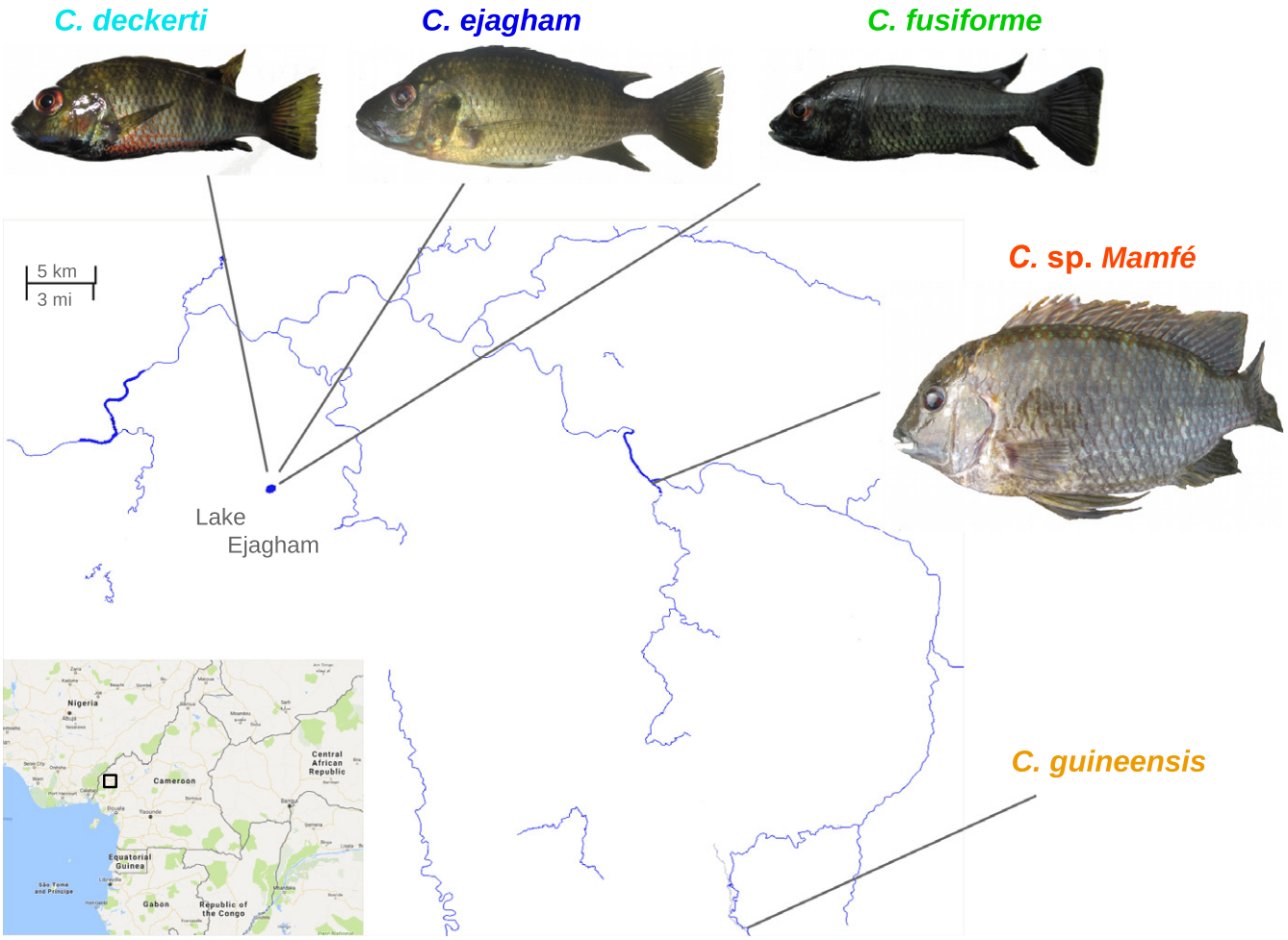

FIGURE 1 Lake Ejagham and its surrounding rivers in southwestern Cameroon. The focal species in this study are shown: three species of Lake Ejagham *Coptodon* and two closely related riverine species. As outgroups, we used *C. kottae*, a crater lake endemic that did not diversify, and *Sarotherodon galilaeus* [Colour figure can be viewed at wileyonlinelibrary.com]

Ejagham *Coptodon* and the two riverine species each form exclusive clades (Figure 2c, Supporting information Figure S3, Supporting information Table S9). Similarly, in 87.61% of the genome, individuals in each of the three Ejagham species grouped monophyletically (Figure 2c, Supporting information Table S9).

3.2 | Genomewide tests of admixture suggest ongoing gene flow from *C.* sp. "Mamfé"

To further investigate admixture between riverine and Lake Ejagham taxa, we first used genome-wide formal tests of admixture. Genomewide *D*-statistics in configurations that test for admixture between one of the two riverine species and an Ejagham *Coptodon* species, repeated for each Ejagham species, all indicate admixture between *C.* sp. *"Mamfé"* and Ejagham *Coptodon* (Figure 3a, top three bars). Values of *D* were very similar (0.1578–0.1594) across the three Ejagham species, indicating similar levels of admixture from *C.* sp. *"Mamfé."* This suggests that admixture may have predominantly taken place prior to diversification within Lake Ejagham.

We tested this interpretation using five-taxon D_{FOIL} statistics (Figure 3b). D_{FOIL} statistics take advantage of derived-allele frequency patterns in a phylogeny that contains an outgroup and two pairs of sister populations (i.e., the phylogeny is symmetric, see Pease & Hahn, 2015) that differ in coalescence time (i.e., one population pair diverged before the other pair did). The combination of signs (significantly positive, significantly negative, or not significantly different from zero) across four D_{FOIL} statistics, D_{FO}, D_{IL}, D_{FI} and D_{OL}, can distinguish (a) admixture along terminal branches between a population in each of the two population pairs from (b) admixture between the ancestral population of the most recently diverged population pair and a population in the other pair. In the case of admixture along terminal branches, the direction of gene flow can also be inferred, whereas it cannot for ancestral gene flow. The four statistics are not affected by gene flow within each population pair. Here, we repeated the test with each of three possible pairs of Lake Ejagham species as P1 and P2, and with P3 and P4 for the pair of riverine species, which diverged prior to the Ejagham species (see next section). D_{FOIL} statistics using both pairs of Lake Ejagham taxa that involve *C. fusiforme* indicated a pattern of admixture between *C.* sp.

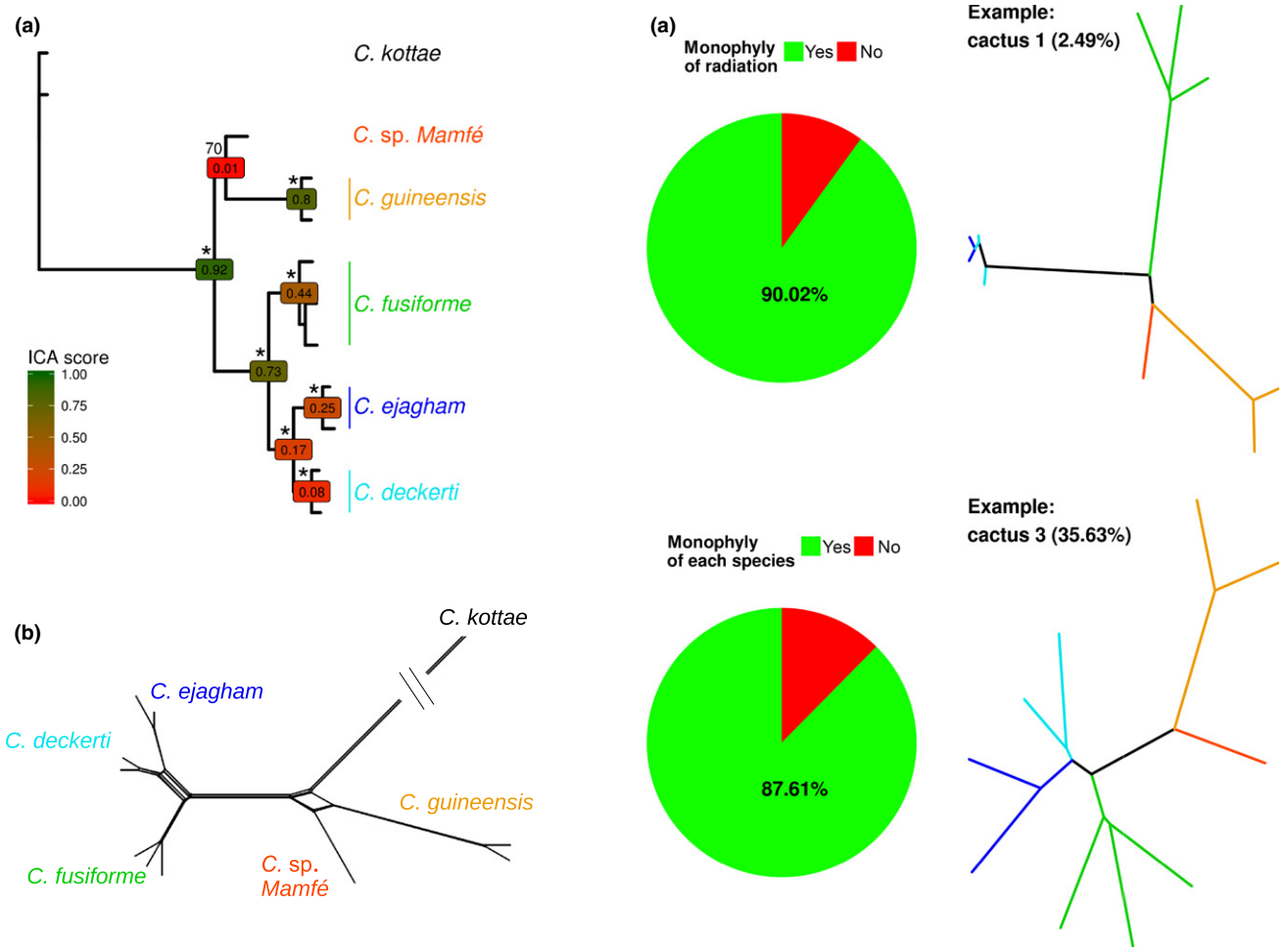

FIGURE 2 Support for monophyly of the Lake Ejagham *Coptodon* radiation across the genome. (a) Maximum-likelihood tree based on concatenated SNPs across the genome, with bootstrap support (* = 100% support), and ICA (Internode Confidence All) values based on ML gene trees for 100-kb windows. Support for the sister relationship between the riverine species *C.* sp. *"Mamfé"* and *C. guineensis* is much lower than that for the monophyly of the three lake Ejagham species, *C. fusiforme*, *C. ejagham* and *C. deckerti*. (b) A phylogenetic network shows limited conflict along the branch leading to lake Ejagham species and a rather clearly resolved topology within the radiation. In line with results from panel A, more conflict is observed around the divergence of *C.* sp. *"Mamfé"* and *C. guineensis*. (c) Local phylogenies (Saguaro "cacti") indicate that along most of the genome, the Ejagham *Coptodon* clade (top) is monophyletic and that individuals within the clade cluster by species (bottom) [Colour figure can be viewed at wileyonlinelibrary.com]

"Mamfé" and the Lake Ejagham ancestor (Figure 3b, left). D_{FOIL} statistics are designed to uncover a single admixture pattern, such that multiple instances of gene flow may lead to a combination of signs across D_{FOIL} statistics without a straightforward interpretation, which may explain the pattern observed for the comparison with *C. deckerti* and *C. ejagham* as P1 and P2 (Figure 3b, right).

Consistent with more complex patterns of admixture, *D*-statistics for comparisons that explicitly test for differential admixture between Ejagham species with *C.* sp. *"Mamfé"* indicate that *C. ejagham* and *C. deckerti* experienced slightly higher levels of admixture than *C. fusiforme* after their divergence (Figure 3a, bottom bars). Furthermore, an f_4-ratio test suggests that 4.7% of *C. ejagham* ancestry derives from admixture with *C.* sp. *"Mamfé"* during or after its divergence from *C. deckerti* (Figure 3c), but it should be noted that *D*-statistics did not indicate differential admixture for this comparison (Figure 3a, bottom bar). Overall, we infer that differential gene flow from *C.* sp. *"Mamfé"* into the three Ejagham species has been relatively minor in comparison with gene flow shared among the species. The difference in magnitude can be seen in Figure 3a, in which the upper three bars represent shared gene flow and the lower three bars differential gene flow to Ejagham species. In line with the results from D_{FOIL} statistics, this in turn suggests that gene flow to the ancestral Lake Ejagham population was more pronounced than to extant species, an interpretation that we tested further using G-PHOCS.

3.3 | Estimation of the demographic speciation history of the Ejagham radiation

To infer postdivergence rates of gene flow, divergence times and population sizes among the extant and ancestral Lake Ejagham lineages and the two riverine species, we used the generalized phylogenetic coalescent sampler (G-PHOCS), providing the species tree

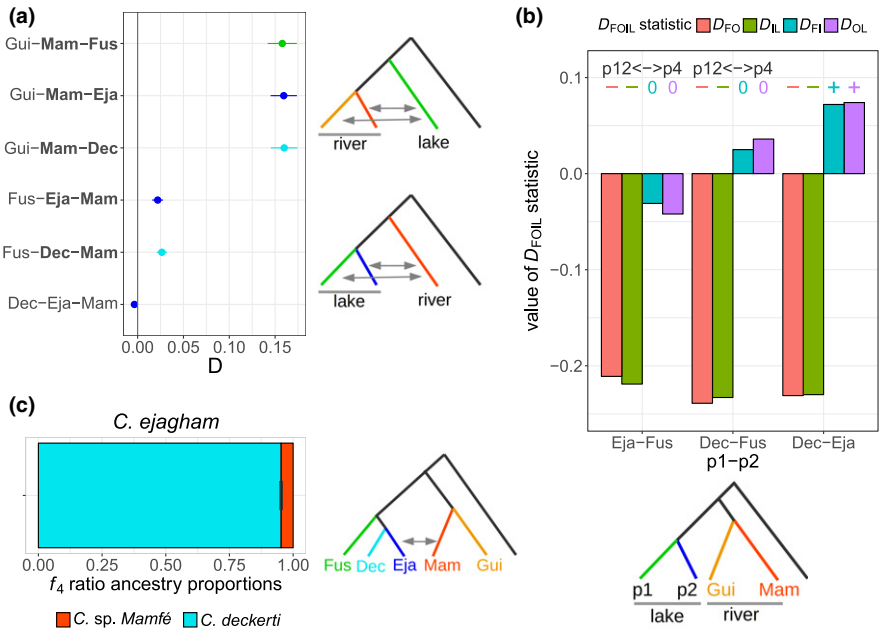

FIGURE 3 Genomewide admixture statistics suggest secondary riverine gene flow from *C.* sp. *"Mamfé."* (a) *D*-statistics for several ingroup triplets indicate that all three Ejagham *Coptodon* species ("Fus": *C. fusiforme*, "Eja": *C. ejagham*, "Dec": *C. deckerti*) experienced admixture with *C.* sp. *"Mamfé"* ("Mam"), at similar levels relative to *C. guineensis* ("Gui"), as shown by the top three bars. The lower three bars show the much weaker evidence for differential *C.* sp. *"Mamfé"* admixture among Ejagham *Coptodon* species. Species between which admixture is inferred (significant *D*-statistics) are denoted in bold. (b) D_{FOIL} statistics for the three combinations of two Ejagham *Coptodon* species show a preponderance of ancestral gene flow with *C.* sp. *"Mamfé."* Negative D_{FO} and D_{IL} in combination with nonsignificant D_{FI} and D_{OL} statistics, as for the first two comparisons, indicate ancestral gene flow, while the pattern for the third combination does not have a straightforward interpretation, although it is qualitatively similar to the first two comparisons. (c) An f_4-ratio test for differential *C.* sp. *"Mamfé"* admixture between *C. ejagham* and *C. deckerti* indicates that *C. ejagham* has experienced 4.7% additional admixture from *C.* sp. *"Mamfé."* [Colour figure can be viewed at wileyonlinelibrary.com]

topology inferred above. Gene flow rates in G-PHOCS can be estimated using specific "migration bands" between any two lineages that overlap in time. We focused on migration bands that had a riverine lineage as the source population and an extant or ancestral Lake Ejagham lineage as the target population. We first inferred rates in models with single migration bands and then combined significant migration bands in models with multiple migration bands. While models with all migration bands performed more poorly due to the high number of parameters (see Methods), models with single migration bands may be prone to overestimation of that specific migration rate. We therefore also ran models with an intermediate number of migration bands (either to all three extant Ejagham species or to both ancestral lineages) and present results for all these different models in Figure 4 and Table 1. Divergence times and population sizes mentioned below represent only those from models with all significant migration bands.

Divergence between the ancestral riverine and Lake Ejagham lineages was estimated to have occurred around 9.76 kya (95% Highest Posterior Density (HPD): 8.27–11.23, Figure 4a), which we consider an estimate of the timing of the colonization of Lake Ejagham. Encouragingly this coincides with the age of the lake estimated from core samples (9 kya: Stager et al., 2017). In contrast to rapid colonization of the new lake, we estimated that the first speciation event in Lake Ejagham only occurred 1.20 [0.81–1.62] ka ago, rapidly followed by the second 0.69 [0.29–1.10] ka ago. These divergence dates remained relatively similar even in models with no gene flow (point estimates 8.80, 2.15 and 1.05 ka ago, Figure 4b).

Inferred effective population sizes among Ejagham *Coptodon* varied about fourfold. We inferred a smaller effective population size for *C. ejagham* (N_e = 933 [406–1,524]) compared to the other two crater lake species (*C. deckerti*: 3,680 [1,249–6,539], *C. fusiforme*: 2,864 [1,514–4,743], Table 1, Figure 4e–f), which is in line with field observations of its low abundance (Martin, 2013) and piscivorous ecology (Dunz & Schliewen, 2010).

In agreement with the results from genome-wide admixture statistics, we infer that secondary gene flow from riverine species has taken place mostly or only from *C.* sp. *"Mamfé"* relative to *C. guineensis*. In models with single migration bands, significant gene flow was inferred from *C.* sp. *"Mamfé"* into all Ejagham lineages (Figure 4d–f). Rates of gene flow to ancestral populations dropped relative to extant lineages in models with all migration bands, in particular for gene flow to the lineage ancestral to all three species (Figure 4d–f).

Overall, G-PHOCS inferred similar rates of gene flow from *C.* sp. *"Mamfé"* to extant species (Figure 4d–f). Nevertheless, due to a higher inferred rate to the *C. deckerti–C. ejagham* ancestor than to *C. fusiforme*, we infer that since its divergence, *C. fusiforme* experienced less gene flow than *C. deckerti* and *C. ejagham* (40.6% and 43.2% less, respectively, in terms of the "total migration rate" estimated in single

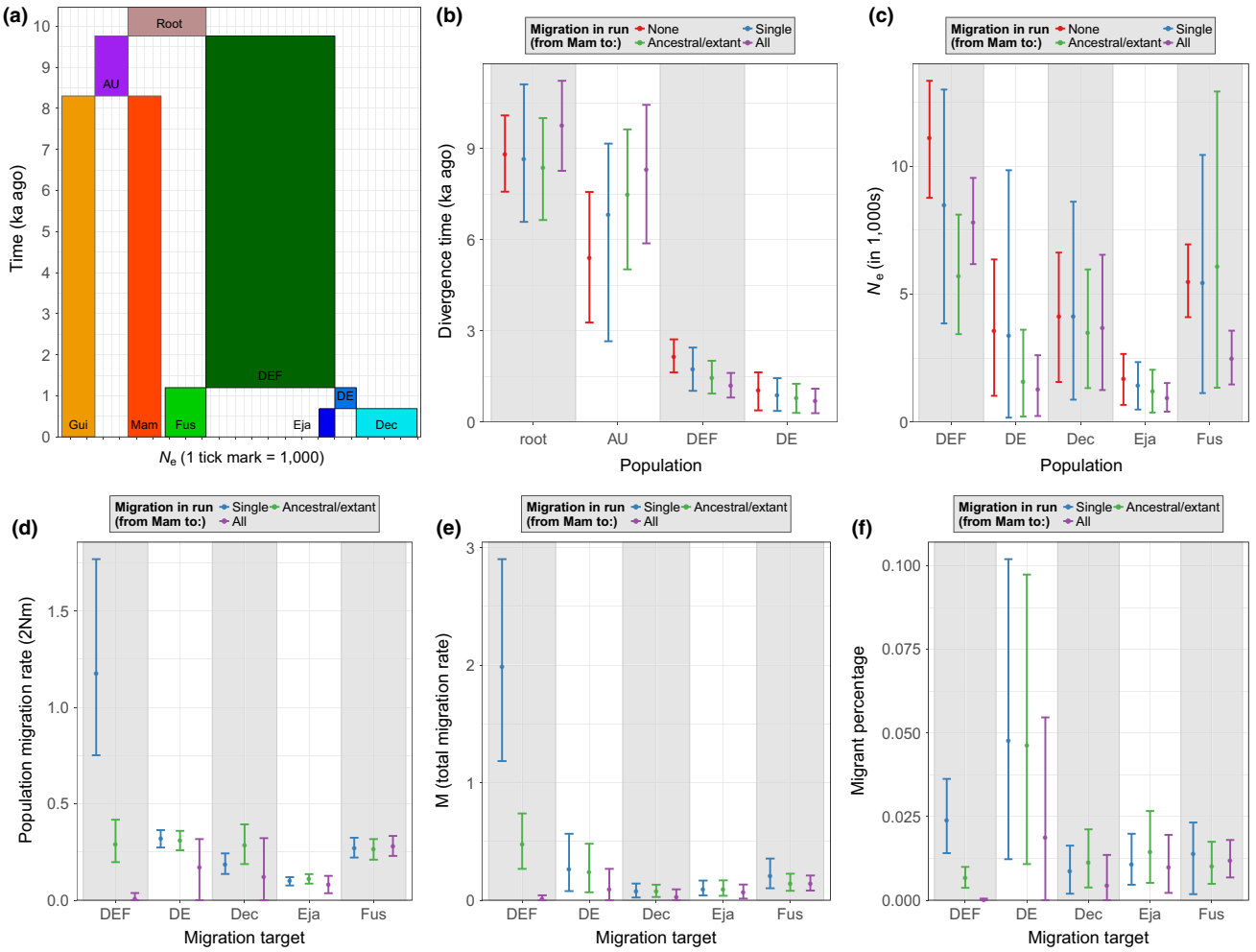

FIGURE 4 A comprehensive picture of the demographic speciation history of Ejagham *Coptodon*. (a) Overview of the divergence times and population sizes inferred by G-PhoCS under the scenario of migration bands to all Lake Ejagham lineages. Box widths (*x*-axis) correspond to population sizes only for Lake Ejagham lineages: *C. deckerti* ("Dec"), C. *ejagham* ("Eja"), *C. fusiforme* ("Fus"), the ancestor of Dec and Eja ("DE") and the ancestral Ejagham lineage ("DEF"). (b–f) Estimates of divergence times (b), population sizes (c) and migration rates (d–f) across runs with varying migration bands from *C.* sp. *"Mamfé"* to lake lineages: "none," "single," "ancestral/current" and "all" indicate that individual runs estimated zero, one, several (either to the two ancestral lineages, DE and DEF, or to the three extant species) or all possible migration bands (to both ancestral and all three ancestral lineages), respectively (see Supporting information Figure S4) [Colour figure can be viewed at wileyonlinelibrary.com]

migration band models), which agrees with the results from *D*-statistics (Figure 3a). However, due to the higher rate inferred in the band between *C.* sp. *"Mamfé"* and the Ejagham ancestor, and the longer time span of this band, the estimated total migration rate since the split of the ancestral Ejagham lineage differs only by 6.63% between *C. fusiforme* and *C. ejagham*, 6.39% between *C. fusiforme* and *C. deckerti*, and 0.67% between *C. deckerti* and *C. ejagham* (Table 1, Figure 4d–f).

We did not find clear evidence for gene flow into Ejagham *Coptodon* from other sources besides *C.* sp. *"Mamfé"* using G-PHOCS. All rates of gene flow into Lake Ejagham lineages from *C. guineensis* or from the riverine ancestor (prior to the split between *C.* sp. *"Mamfé"* and *C. guineensis*) had 95% HPD intervals that overlapped with zero, and all except two had means very close to zero (Table 1, Supporting information Figure S4A). Only the estimates of gene flow from *C. guineensis* into the two ancestral Ejagham lineages had mean population migration rates above 0.01 (0.18 and 0.47) and high variance (Supporting information Figure S4A), suggesting either the possibility of low levels of ancestral gene flow from *C. guineensis*, or that gene flow from *C. guineensis* at that period may be conflated with gene flow from *C.* sp. *"Mamfé"* (as this is closer to the coalescence time of *C. guineensis* and *C.* sp. *"Mamfé"*). In support of the latter idea, in models that combined gene flow to ancestral Ejagham lineages from *C.* sp. *"Mamfé"* and *C. guineensis*, gene flow from *C. guineensis* was again not different from zero, while the variance was much smaller, and gene flow from *C.* sp. *"Mamfé"* remained significant (Supporting information Figure S4B).

We also did not find clear evidence for gene flow among Ejagham *Coptodon* lineages using G-PHOCS. We evaluated models with each one of all possible migration bands in both directions, and 95% HPD for all migration rates overlapped with zero (Supporting information Figure S4C). The mean inferred population migration rate was higher than

TABLE 1 Summary of G-PHOCS parameter estimates

Parameter	Lineage	Mean Single	Mean Anc/ext	Mean All	Mean None	95% HPD Single	95% HPD Anc/ext	95% HPD All	95% HPD None
τ	Root	8,649	8,369	9,760	8,803	6,587–11,112	6,647–10,001	8,267–11,229	7,579–10,090
τ	AU	6,823	7,498	8,298	5,393	2,658–9,163	5,024–9,631	5,880–10,438	3,275–7,571
τ	DEF	1,740	1,454	1,205	2,150	1,027–2,451	936–2,015	806–1,616	1,633–2,721
τ	DE	892	778	689	1,049	365–1,443	300–1,259	291–1,096	3,83–1,635
N_e	DEF	8,482	5,714	7,794	11,121	3,857–13,001	3,435–8,109	6,175–9,545	8,768–13,343
N_e	DE	3,373	1,589	1,288	3,574	171–9,846	216–3,608	235–2,613	1,025–6,358
N_e	Dec	4,133	3,500	3,681	4,128	874–8,615	1,328–5,967	1,250–6,539	1,566–6,625
N_e	Eja	1,425	1,180	933	1,684	489–2,343	371–2,044	406–1,525	670–2,662
N_e	Fus	5,432	6,069	2,474	5,488	1,131–10,444	1,342–12,925	1,469–3,572	4,100–6,946
2*Nm*	DEF	1.18	0.29	0.01	NA	0.75–1.77	0.2–0.42	0–0.04	NA
2*Nm*	DE	0.32	0.31	0.17	NA	0.27–0.36	0.26–0.36	0–0.32	NA
2*Nm*	Dec	0.19	0.28	0.12	NA	0.14–0.24	0.19–0.39	0–0.32	NA
2*Nm*	Eja	0.10	0.11	0.08	NA	0.08–0.12	0.09–0.14	0.04–0.13	NA
2*Nm*	Fus	0.27	0.26	0.28	NA	0.22–0.32	0.21–0.32	0.23–0.33	NA
M (total)	DEF	1.98	0.48	0.01	NA	1.18–2.9	0.27–0.74	0–0.04	NA
M (total)	DE	0.27	0.24	0.09	NA	0.08–0.57	0.07–0.48	0–0.27	NA
M (total)	Dec	0.08	0.07	0.03	NA	0.02–0.14	0.03–0.13	0–0.09	NA
M (total)	Eja	0.09	0.09	0.07	NA	0.04–0.17	0.04–0.17	0.01–0.13	NA
M (total)	Fus	0.20	0.14	0.14	NA	0.1–0.35	0.08–0.23	0.08–0.21	NA
% Migrants	DEF	1.39e-4	5.07e-5	7.03 e-7	NA	8.9 e-5–2.1 e-4	3.5 e-5–7.3 e-5	0–4.8 e-6	NA
% Migrants	DE	9.47 e-5	1.96 e-4	1.33 e-4	NA	8.1 e-5–1.1 e-4	1.6 e-4–2.3 e-4	0–2.5 e-4	NA
% Migrants	Dec	4.52 e-5	8.11 e-5	3.33 e-5	NA	3.3 e-5–5.9 e-5	5.4 e-5–1.1 e-4	0–8.8 e-5	NA
% Migrants	Eja	6.91 e-5	9.45 e-5	8.61 e-5	NA	5.4 e-5–8.4 e-5	7.3 e-5–1.2 e-4	3.9 e-5–1.4 e-4	NA
% Migrants	Fus	4.94 e-5	4.34 e-5	1.14 e-4	NA	4.1 e-5–6.0 e-5	3.5 e-5–5.2 e-5	9.3 e-5–1.4 e-4	NA

Notes. "τ": divergence time; "2*Nm*": population migration rate; "*M* (total)": total migration rate; "% Migrants": percentage of migrants received in each generation; "HPD": Highest Posterior Density; "AU": ancestor of *C.* sp. *"Mamfé"* and *C. guineensis*; "DEF": ancestor of all three lake Ejagham species; "DE": ancestor of *C. deckerti* and *C. ejagham*; "Dec": *C. deckerti*; "Eja": *C. Ejagham*; "Fus": *C. fusiforme*. Divergence time τ represents the estimated time that the named lineage split into its daughter lineage (see Figure 4a). All migration rates are from migration from *C.* sp. *"Mamfé"* to Lake Ejagham lineages. Parameter estimates are given separately for runs with no migration ("None"), with a single migration band ("Single"), with migration bands to either both ancestral or all three extant lineages ("Anc/ext"), or to all Lake Ejagham lineages ("All").

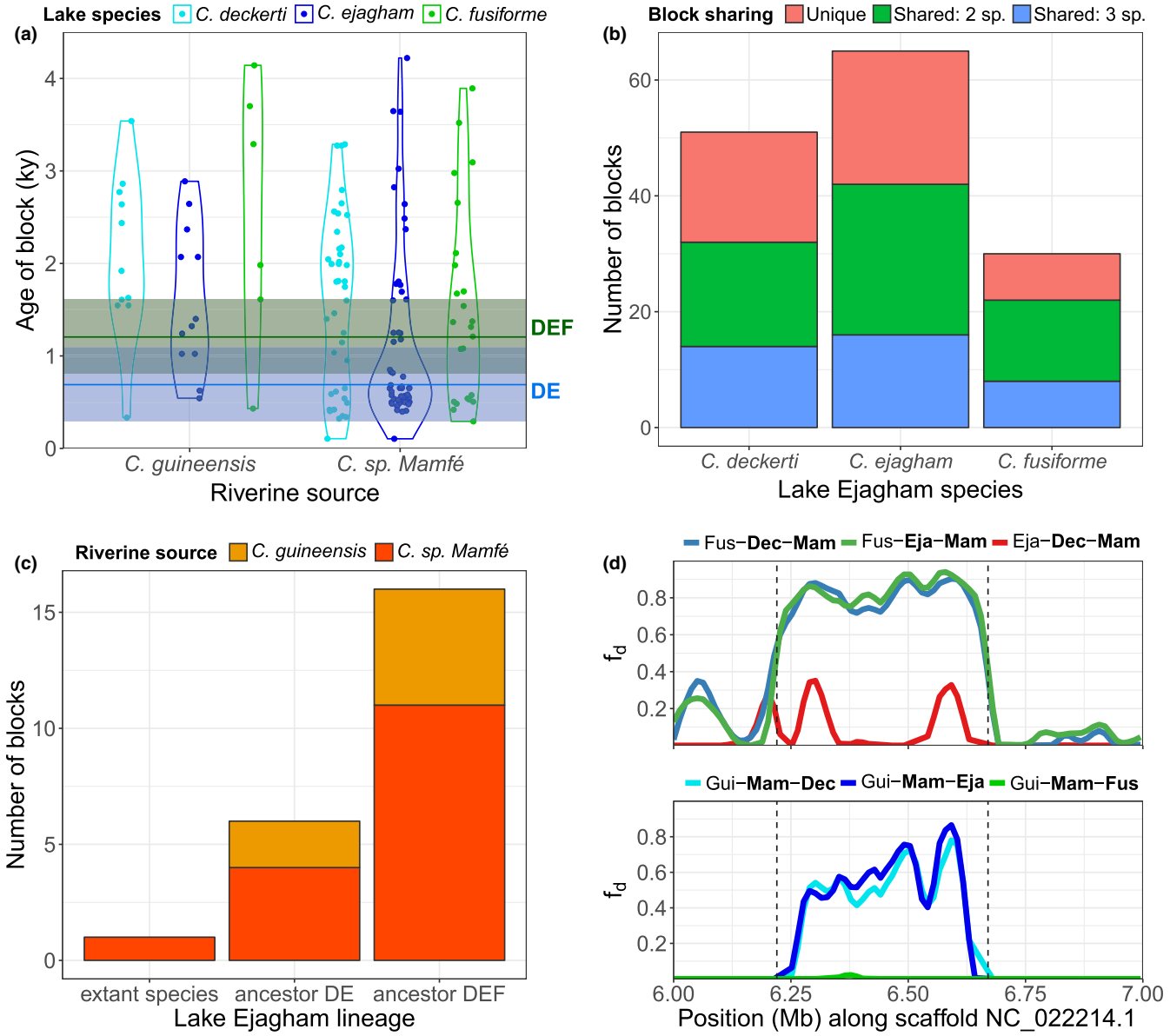

FIGURE 5 Evidence for introgression from admixture blocks. Only "high-confidence" admixture blocks, that is, with a maximum estimated age younger than the minimum estimated divergence time of Ejagham *Coptodon* are shown. (a) Age estimates of admixture blocks show ongoing introgression. Estimated divergence times of *C. deckerti* and *C. ejagham* (blue line DE), and of *C. fusiforme* and the DE ancestor (green line DEF), and the corresponding 95% HPD intervals, are also shown. (b) Both unique and shared (either among two or three species) admixture blocks were detected, and the fewest blocks were detected in *C. fusiforme*. (c) A subset of blocks could be categorized using D_{FOIL} statistics, the large majority of which introgressed into the ancestral Ejagham lineage ("ancestor DEF"). (d) An example of an admixture block, which is shared between *C. deckerti* and *C. ejagham*, and estimated by HYBRIDCHECK to have been introgressed 2,486 (1,651–3,554) years ago [Colour figure can be viewed at wileyonlinelibrary.com]

0.01 only for *C. fusiforme* to *C. deckerti* (0.27) and to *C. ejagham* (0.02). Such limited evidence for secondary gene flow within the radiation is surprising, given that these species are in the earliest stages of speciation (Martin, 2013). However, due to the very recent divergence of these lineages few informative coalescence events are likely to be present, in turn resulting in low power to identify ongoing gene flow. Furthermore, representative breeding pairs at the tail ends of the unimodal phenotype distribution for *C. fusiforme/deckerti* were selectively chosen for sequencing (Martin, 2012), while excluding ambiguous individuals that could not be assigned to a particular species.

3.4 | Admixture blocks support ongoing gene flow from *C.* sp. "Mamfé"

To identify genomic blocks of admixture between riverine and Lake Ejagham species, we first defined putative blocks as contiguous sliding windows that were outliers for f_d, a four-population introgression statistic related to *D* that is suitable for application to small genomic regions, and subsequently used HYBRIDCHECK (Ward & van Oosterhout, 2016) to validate and age these blocks. We used all combinations of ingroup triplets that could differentiate between admixture from *C. guineensis* and

C. sp. "*Mamfé*," as well as those that could identify differential admixture among Lake Ejagham species (from either riverine species) (Supporting information Table S10). Of 1,138 putative blocks identified as f_d outliers, 340 were also identified by HYBRIDCHECK (93 from *C. guineensis*, and 247 from *C.* sp. "*Mamfé*"). While such blocks represent areas with ancestry patterns consistent with admixture, these patterns can also be produced by incomplete lineage sorting (ILS). To distinguish between ILS and admixture, we took advantage of our estimates of block age (coalescence time between the focal species pair) from HYBRIDCHECK and our estimates of divergence times from G-PHOCS. While nearly a quarter of blocks were estimated to be older than the Lake Ejagham lineage, and therefore likely represent ILS (Supporting information Figure S6), we identified 259 "likely" candidate regions (with a point estimate of age younger than that of the Lake Ejagham lineage), including a subset of 146 "high-confidence" regions (with nonoverlapping confidence intervals of age estimates), resulting from secondary gene flow into Ejagham. In total, high-confidence admixture blocks comprised only 0.64% (5.7 Mb) of the queried part of the genome.

In accordance with the much stronger evidence for Lake Ejagham admixture with *C.* sp. "*Mamfé*" than with *C. guineensis*, the majority of likely (68.3%) and high-confidence (80.1%) admixture blocks involved *C.* sp. "*Mamfé*" as the riverine species, and likely and high-confidence admixture blocks with *C.* sp. "*Mamfé*" were, on average, younger (2.94 and 1.37 ka, respectively) than those with *C. guineensis* (4.55 and 1.97 ka, respectively, Figure 5a).

Because f_d and HYBRIDCHECK detect admixture only between species pairs, we took two approaches to investigate at which point along the Lake Ejagham phylogeny admixture took place for likely admixture blocks. First, we intersected admixture blocks involving different Lake Ejagham species, but the same riverine species, and detected 76 likely (and 38 high-confidence) blocks involving a single Lake Ejagham species, 88 (50) blocks shared among two Lake Ejagham species and 95 (87) blocks shared among all three Lake Ejagham species (Figure 5b). Thus, 29.3% of likely blocks (and 26.0% of high-confidence blocks) were unique to a single lake species, but this may be an overestimate, as such blocks may have been present but escaped statistical detection in other species, for instance due to recombination within the block. This possibility is underscored by the age distribution of admixture blocks: Admixture blocks detected in one species were not younger than those detected in multiple species (Supporting information Figure S6). In line with results from genomewide admixture statistics and G-PHOCS, we found more admixture blocks into *C. deckerti*, *C. ejagham* and their ancestor, compared to *C. fusiforme* (Figure 5b).

Second, we used D_{FOIL} statistics to distinguish between admixture involving the ancestral Lake Ejagham lineage ("DEF"), the *C. deckerti*–*C. ejagham* ancestor ("DE"), and the terminal branches. We were able to categorize 23 likely (and 13 high-confidence) admixture blocks with D_{FOIL} statistics, showing a pattern of decreasing occurrence of admixture blocks through time, with only a single likely (and 0 high-confidence) block involving a terminal Lake Ejagham branch (Figure 5c). For cases where admixture is with an ancestral (lake) clade, D_{FOIL} statistics cannot infer the direction of introgression, but the single classified admixture block with an extant lake taxon is, as expected, inferred to have been into the lake.

3.5 | Admixture of olfactory genes into *C. deckerti* and *C. ejagham*

Among all high-confidence blocks, 11 gene ontology terms were enriched (Table 2). Eight genes in a single admixture block on scaffold

TABLE 2 Gene Ontology term enrichment among genes in admixture blocks

Ontology	Category	Term	FDR	Nr. of genes	*C. deckerti*	*C. ejagham*	*C. fusiforme*	Unique	Shared: 2 species	Shared: 3 species
BP	GO:0007608	Sensory perception of smell	2.08e-09	8	1	1	0	0	1	0
MF	GO:0004984	Olfactory receptor activity	2.08e-09	8	1	1	0	0	1	0
BP	GO:0050911	Detection of chemical stimulus involved in sensory perception of smell	2.08e-09	8	1	1	0	0	1	0
BP	GO:0050896	Response to stimulus	6.69e-08	8	1	1	0	0	1	0
MF	GO:0004871	Signal transducer activity	1.51e-07	14	1	1	0	0	1	0
BP	GO:0007186	G-protein coupled receptor signaling pathway	1.47e-05	13	1	1	0	0	1	0
MF	GO:0004930	G-protein coupled receptor activity	4.79e-05	12	1	1	0	0	1	0
BP	GO:0007165	Signal transduction	6.42e-05	14	1	1	0	0	1	0
CC	GO:0005886	Plasma membrane	2.55e-04	11	1	1	0	0	1	0
MF	GO:0004336	Galactosylceramidase activity	4.49e-03	2	1	1	1	0	0	1
BP	GO:0006683	Galactosylceramide catabolic process	4.49e-03	2	1	1	1	0	0	1

Notes. FDR and number of genes are given for genes in all "high-confidence" admixture blocks. The last six columns indicate whether (1) or not (0) each term was also enriched (FDR < 0.05) for subsets of admixture blocks involving each species and each block sharing category ("unique": blocks unique to one Lake Ejagham species; "shared: 2/3 species": blocks shared among two/three Lake Ejagham species. No additional GO terms were enriched for admixture blocks subsets only. Ontologies: BP, Biological Process; CC, Cellular Component; MF, Molecular Function.

NC_022214.1 were responsible for the three most enriched categories; seven of these genes are characterized as olfactory receptors and the eighth as "olfactory receptor-like protein" (none have a gene name, and only one has 1-to-1 orthologs in other species on Ensembl Release 90 (Supporting information Table S12)). The admixture block containing this cluster of genes, which is shown in Figure 5d, was estimated to have introgressed from *C.* sp. *"Mamfé"* into both *C. deckerti* and *C. ejagham* 2,486 (1,651–3,554) years ago, shortly prior to the divergence of the *C. deckerti*/*C. ejagham* ancestor from *C. fusiforme*, 1,205 (806–1,616) years ago. Among all high-confidence admixture blocks, this block was the largest, had the highest summed f_d score and had the second lowest HYBRIDCHECK *p*-value. Across the entire block, *C. deckerti*, *C. ejagham* and *C.* sp. *"Mamfé"* show uniformly low genetic differentiation (Supporting information Figure S7).

When performing GO analyses separately for blocks involving each Lake Ejagham species, no additional terms were found to be enriched. With respect to admixture blocks involving each Lake Ejagham species, the same 11 terms were enriched for *C. ejagham*, nine of these terms were enriched for *C. deckerti*, and none were enriched for *C. fusiforme* (Table 2). Blocks unique to one Lake Ejagham species (either taken together, or separately by species) were not enriched for any terms, while blocks shared between two species were enriched for nine terms and blocks shared between all three species for two terms (Table 2).

4 | DISCUSSION

Here, we showed that the young Lake Ejagham was rapidly colonized by the ancestors of the endemic *Coptodon* radiation and that no major secondary colonizations have taken place. Yet in contrast to the classic paradigm of a highly isolated lake colonized only once by a single cichlid pair (Schliewen et al., 1994), we found low levels of gene flow from one of the riverine species into all three species in the lake throughout their speciation histories. Interestingly, one of the clearest signals of introgression came from a cluster of olfactory receptor genes that introgressed into the ancestral population around 2.5 kya, just prior to the first speciation event, suggesting that gene flow may have facilitated speciation.

4.1 | Rapid initial colonization of Lake Ejagham

Our estimate of the timing of colonization of Lake Ejagham by the *Coptodon* lineage (9.76 ka ago, Figure 4a) was similar to the estimated age of the lake itself (9 ka years ago, Stager et al., 2017), suggesting that the lake was rapidly colonized by the ancestral lineage. It should be noted that this estimate in turn relies on an estimate of the mutation rate. We here use an estimate from stickleback (Guo et al., 2013), following previous studies on cichlids (Kautt, Machado-Schiaffino, Meyer et al., 2016; Kautt, Machado-Schiaffino, Torres-Dowdall et al., 2016), but it cannot be excluded that our focal species may have substantially different spontaneous mutation rates (Martin & Höhna, 2018; Martin et al., 2017; Recknagel, Elmer, & Meyer, 2013).

Martin, Cutler et al. (2015) argued that the Cameroon lakes containing cichlid radiations may not be as isolated as has previously been suggested, based on the inference of secondary gene flow into all four radiations and the fact that each lake has been colonized by several different fish lineages (five in the case of Lake Ejagham). Our inference of a rapid, successful colonization process and evidence for ongoing gene flow are both in support of this view. In this light, it is worth pointing out that lake Ejagham (a) has an outflow in the wet season which may be connected to the Munaya River (a tributary of the Cross River system), (b) does not have a waterfall that could prevent fish from entering the lake as in crater lakes Barombi Mbo and Bermin (C. H. Martin, personal observation) and (c) is at an elevation of only 141 m, about 60 m higher than the closest river drainage (Barombi Mbo and Bermin crater lakes are at altitudes of 314 and 472 m, respectively).

4.2 | No major secondary colonizations

Our data suggest that the initial colonization of the lake established the population that has since diversified within Lake Ejagham and we found no evidence for major secondary colonizations that either gave rise to a new lineage or resulted in a hybrid swarm. Several lines of evidence indicate that such events are unlikely to have taken place. First, considerable phylogenetic conflict would be expected if diversification happened rapidly after a secondary colonization event, while we found widespread monophyly across the genome (89.34%, Supporting information Table S9). Second, we inferred a long time lag between colonization and the first speciation event within the lake (9.76 ka and 1.20 ka ago, respectively, Figure 4a, Table 1). Third, we estimated gene flow into the ancestral lake lineage to be relatively low (Figure 4b). Similarly, models with and without postdivergence gene flow between riverine and lake lineages resulted in similar (9.76 and 8.80 ka ago, respectively, Table 1) estimates of the divergence time of the ancestral lake lineage.

4.3 | Ongoing low levels of gene flow from one of two Cross River *Coptodon* species

Even though we found that Ejagham *Coptodon* was established by a single major colonization, our results are not consistent with subsequent isolation of the lake population. We found evidence for secondary gene flow from the riverine source population that was ongoing, that is, into ancestral as well as extant Ejagham lineages. The riverine source population diverged into *C. guineensis* and *C.* sp. *"Mamfé"* after the split with the Ejagham lineage. Results from all three types of approaches that we used to identify secondary gene flow (demographic analysis with G-PHOCS, genome-wide admixture statistics, and the identification of admixture blocks) show that gene flow originated predominantly from one of these riverine lineages, *C.* sp. "Mamfé" (Figures 3a, 4b and 5). Little is known about the precise geographic distribution of *C.* sp. "Mamfé," yet this asymmetry is consistent with the closer sampling location of this species (37 km from Lake Ejagham to the Cross River at Mamfé) relative to that of *C. guineensis* (65 km from Lake Ejagham to a tributary of the

Cross River at Nguti; see also Figure 1 that depicts all major rivers). Both *Coptodon* lineages are known to coexist within the Cross River drainage. Our data suggest that *C.* sp. "Mamfé" is most likely a new species.

Evidence for gene flow from *C. guineensis* was much weaker compared to *C.* sp. "*Mamfé*" and was mostly restricted to ancestral Lake Ejagham lineages (admixture blocks: Figure 5, G-PHOCS: Supporting information Figure S4A–B). It should furthermore be noted that the assignment of the riverine source lineage is likely to be more error-prone further back in time, given the recent divergence between *C. guineensis* and *C.* sp. "*Mamfé.*" However, the clearest evidence of gene flow from *C. guineensis* comes from admixture blocks, where an inference of differential ancestry from the two riverine species was required. As we were only able to include a single *C.* sp. "*Mamfé*" individual, it is nevertheless possible that we missed substantial genetic variation in that species connecting it to *C. guineensis*.

4.4 | Differential gene flow into the Ejagham radiation and contemporary hybridization

We found some evidence for differential riverine admixture from *C.* sp. "*Mamfé*" into the three Ejagham species. While the admixture proportion of *C. ejagham* may be slightly higher than that of *C. deckerti* (f_4-ratio test: Figure 3b, but see *D*-statistics, Figure 3a, and G-PHOCS: Figure 4a–c), the evidence was stronger for elevated riverine admixture with sister species *C. deckerti* and *C. ejagham* relative to *C. fusiforme* (*D*-statistics: Figure 3a, admixture blocks: Figure 5), which specifically appears to originate from higher admixture into the *C. deckerti* / *C. ejagham* ancestor (G-PHOCS: Figure 4b). In accordance with this, Martin, Cutler et al. (2015) identified riverine admixture with the *C. deckerti*/*C. ejagham* ancestor using Treemix. Martin, Cutler et al. (2015) found that a proportion of *C. fusiforme* individuals appeared more admixed than any other Ejagham *Coptodon*. The magnitude of the effect in their PCA plot (Figure 3c in Martin, Cutler et al., 2015), as well as the fact that only some of the *C. fusiforme* individuals were involved, suggests contemporary hybridization; however, this was not supported by their STRUCTURE analysis of the same data. Contemporary hybridization may have resulted from the purposeful introduction of riverine fishes into this lake by an Eyumojock town council member in 2000–2001 (Martin, 2012). This resulted in the establishment of a *Parauchenoglanis* catfish species within the lake, still abundant in 2016 (CHM personal observation). However, no riverine *Coptodon* have been confirmed beyond a posted sign reporting introduced river fishes. In this study, we found no evidence that any of our individuals were recent hybrids (Supporting information Figure S4), but our limited sample size precludes us from ruling out their presence in the lake.

4.5 | Introgression of a cluster of olfactory receptor genes shortly prior to speciation

Complex patterns of secondary gene flow such as those observed here are not easily interpreted in terms of their contribution to speciation. The formation of hybrid swarms has been suggested to promote speciation (Kautt, Machado-Schiaffino, Torres-Dowdall et al., 2016; Meier et al., 2017; Seehausen, 2004), yet we did not find evidence for major secondary colonizations that could be linked to the timing of speciation. The inferred pattern of ongoing gene flow to ancestral as well as extant lineages could theoretically inhibit speciation, by counteracting incipient divergence within the lake, or promote speciation, by introducing novel genetic variation or co-adapted gene complexes.

Interestingly, one admixture block contained a cluster of eight olfactory receptor genes (Supporting information Table S12), causing a highly significant overrepresentation of several gene ontology terms containing these genes (Table 2). While in mammals, the olfactory receptor (OR) gene family is the largest gene family with around 1,000 genes, mostly due to the expansion of a single group of genes, fish species examined so far have far fewer (69–158 complete genes) in a more diverse set of OR genes (Azzouzi, Barloy-Hubler, & Galibert, 2014; Niimura & Nei, 2005). Unfortunately, little additional information is known about the eight admixed olfactory receptor genes.

This cluster of OR genes was contained in the largest and arguably most striking of all high-confidence admixture blocks (Figure 5d), which is estimated to have introgressed from *C.* sp. "*Mamfé*" into *C. deckerti* and *C. ejagham*, but not *C. fusiforme*, just prior to the estimated divergence time of *C. fusiforme* and the ancestor of *C. deckerti* and *C. ejagham*. Thus, the timing, source, and target of introgression all correspond with the inference of elevated levels of gene flow from *C.* sp. "*Mamfé*" to the *C. deckerti/ejagham* ancestor relative to *C. fusiforme* (Figures 3a, 4b and 5). These patterns may suggest a role for the introgression of this block in initiating speciation in Ejagham *Coptodon*. Patterns of f_d (Figure 5d) as well as of other population genetic statistics (F_{ST}, d_{xy}: Supporting information Figure S7) were very similar between *C. deckerti* and *C ejagham* and were also all strikingly uniform across the entire block (Figure 5d, Supporting information Figure S7), suggesting that the introgressed block is present in similar form in both species and has not undergone recombination with nonintrogressed haplotypes.

Chemosensory communication, in general, and olfactory receptors, specifically, have often been linked to speciation, especially with respect to sexual isolation (Smadja & Butlin, 2008). A host of studies has shown the importance of olfactory signalling in conspecific mate recognition in fishes (Crapon de Caprona & Ryan, 1990; Kodric-Brown & Strecker, 2001; McLennan, 2004; McLennan & Ryan, 1999), and in a pair of closely related Lake Malawi cichlids, female preference for conspecific males was shown to rely predominantly if not exclusively on olfactory cues (Plenderleith, van Oosterhout, Robinson, & Turner, 2005). Moreover, in a comparative genomic study, evidence for repeated bouts of positive selection on V1Rs, a family of olfactory receptor genes, was found among East African cichlids (Nikaido et al., 2014). Not surprisingly, it has repeatedly been suggested that olfactory signals and their perception may help explain explosive speciation in cichlids (Azzouzi et al., 2014; Blais et al., 2009; Keller-Costa, Canário, & Hubbard, 2015; Nikaido et al., 2013, 2014).

Olfactory signalling seems particularly relevant to mate choice and speciation in Ejagham *Coptodon*, as three species occur

syntopically, assortative mating among species appears to represent the strongest isolating barrier, and sexual dichromatism is absent (Martin, 2012, 2013). Important next steps will be to examine the importance of olfactory cues in mate recognition in Lake Ejagham *Coptodon*, specifically between *C. fusiforme* and the other two species, and to characterize these genes and their patterns of divergence and admixture in more detail.

4.6 | Waiting time for sympatric speciation

While we inferred that colonization of Lake Ejagham took place more than 9 kya, the first branching event among Ejagham *Coptodon* was estimated to have occurred as recently as 1.20 kya (Figure 4a, Table 1). We did not include the fourth nominal *Coptodon* species in the lake, *C. nigrans*, but extreme phenotypic similarity to *C. deckerti* (Dunz & Schliewen, 2010) and our inability to identify or distinguish these individuals in field collections and observations (Martin, 2012, 2013) suggest a close relationship between *C. deckerti* and this nominal species, which would not change this inference. It thus appears that during the large majority of time that the *Coptodon* lineage was present in Lake Ejagham, no diversification occurred. One possibility is that earlier speciation events did occur, but were followed by extinction. While we cannot exclude this scenario, there are no indications for environmental disruptions such as major changes in water chemistry or depth during the history of Lake Ejagham (Stager et al., 2017).

Assuming that the divergence of *C. fusiforme* was the first within this radiation, a striking difference emerges between the waiting time to the first (7.74 kya) and the next two speciation events, which both occurred within 1.20 kya. The opposite pattern, a slowing speciation rate, would be expected if speciation followed a niche-filling model of ecological opportunity in the lake (Martin 2016). At least two nonmutually exclusive explanations may account for this counterintuitive result.

First, an initial lack of ecological opportunity in young Lake Ejagham may have prevented a rapid first speciation event. A similar pattern is seen in the sympatric radiation of Tristan da Cunha buntings (Ryan, Bloomer, Moloney, Grant, & Delport, 2007), in which, as discussed by Grant and Grant (2009), the ancestral branch is considerably longer than those of the extant species. Grant and Grant (2009) propose that plants that constitute one of the niches used by the extant finch species may have arrived only recently. Similarly resource diversity within the lake may have been insufficient to generate the necessary degree of disruptive selection to drive divergence until recently. For example, *Daphnia* never colonized Barombi Mbo, a Cameroon crater lake containing an endemic cichlid radiation, during its ca. 1 million year existence (Cornen, Bande, Giresse, & Maley, 1992; Green & Kling, 1988).

Second, genetic variation for traits underlying sexual and ecological selection and their associated genetic architecture may initially not have been conducive to speciation. For example, sympatric speciation models predict that there will be a waiting time associated with the initial build-up of linkage disequilibrium between ecological and sexual traits before sympatric divergence can proceed (Bolnick & Fitzpatrick, 2007; Dieckmann & Doebeli, 1999; Kondrashov & Kondrashov, 1999). In this light, it is particularly intriguing that introgression of a block containing eight olfactory receptor genes from *C.* sp. "*Mamfé*," which are likely to be highly relevant for mate choice, was introgressed shortly prior to the first speciation event. Therefore, genetic variation brought in by riverine gene flow may have been necessary to initiate speciation among Lake Ejagham *Coptodon*.

4.7 | Implications for the geographic mode of speciation

In the context of an isolated lake, a classic case of fully sympatric speciation would involve (a) colonization of the lake by a single lineage, effectively in a single event, and (b) no subsequent gene flow with populations outside of the lake prior to or during speciation. Our results suggest that for the Lake Ejagham *Coptodon* radiation, the former is true but the latter is not. Nevertheless, speciation can still be considered sympatric if secondary gene flow was present but did not play a causal role in speciation. While the inferred pattern of ongoing gene flow could be interpreted to be consistent with the absence of a role for gene flow in promoting speciation, the introgression of a cluster of olfactory receptor genes into a pair of sister species (but not the third species) just prior to their divergence indicates that secondary gene flow may have been important to speciation. Such a role for secondary gene flow would exclude fully sympatric speciation. If confirmed in other case studies with secondary gene flow that were formerly considered classic examples of sympatric speciation flow (Nicaraguan cichlids: Kautt, Machado-Schiaffino, Meyer et al., 2016; the other three Cameroon crater lake cichlid radiations: Martin, Cutler et al., 2015; Richards, Poelstra, & Martin, 2017), this would suggest that fully sympatric speciation is rare. Nevertheless, it is important to note that our results do strongly suggest that the overwhelming majority of phenotypic and genetic divergence between Ejagham *Coptodon* has taken place in sympatry, such that this process can still be reasonably characterized as speciation in sympatry.

5 | CONCLUSIONS

We showed that Lake Ejagham was rapidly colonized by ancestors of the extant *Coptodon* radiation in a single major colonization, while also inferring low levels of ongoing secondary gene flow from a riverine species into ancestral as well as extant lake species. This gene flow included the introgression of a cluster of olfactory genes just prior to the first speciation event, which is particularly salient given that Ejagham *Coptodon* species exhibit strong assortative mating, but currently weak disruptive selection, syntopic breeding territories, and no sexual dichromatism within a tiny, shallow lake. Our findings are strongly suggestive of a causal trigger of adaptive radiation in sympatry due to the introgression of olfactory receptors used in mate discrimination.

ACKNOWLEDGEMENTS

We gratefully acknowledge the Cameroonian government and the regional authority and village council of Eyumojock and surrounding communities for permission to conduct this research. We thank Cyrille Dening Touokong, Jackson Waite-Himmelwright, and Patrick Enyang for field assistance and Nono LeGrand Gonwuou for help obtaining permits.

COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY

All sequencing data are deposited in NCBI's Short Read Archive (Accession no. PRJNA453986). The master VCF file and input and output files for all analyses are deposited in the Dryad Digital Repository (https://doi.org/10.5061/dryad.4s5dm31). Scripts to perform all analyses and to produce the figures and tables in the paper can be found at https://github.com/jelmerp/EjaghamCoptodon.

ORCID

Jelmer W. Poelstra https://orcid.org/0000-0002-3514-7462
Emilie J. Richards http://orcid.org/0000-0003-2734-6020
Christopher H. Martin http://orcid.org/0000-0001-7989-9124

REFERENCES

Anderson, E. (1949). *Introgressive hybridization*. New York, NY: Wiley. https://doi.org/10.5962/bhl.title.4553

Arnegard, M. E., & Kondrashov, A. S. (2004). Sympatric speciation by sexual selection alone is unlikely. *Evolution*, *58*, 222–237. https://doi.org/10.1111/j.0014-3820.2004.tb01640.x

Azzouzi, N., Barloy-Hubler, F., & Galibert, F. (2014). Inventory of the cichlid olfactory receptor gene repertoires: Identification of olfactory genes with more than one coding exon. *BMC Genomics*, *15*, 586. https://doi.org/10.1186/1471-2164-15-586

Barluenga, M., Stölting, K. N., Salzburger, W., Muschick, M., & Meyer, A. (2006). Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature*, *439*, 719–723. https://doi.org/10.1038/nature04325

Benjamni, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*, 289–300.

Bergey, C. (2012). vcf-tab-to-fasta. http://code.google.com/p/vcf-tab-to-fasta

Berlocher, S. H., & Feder, J. L. (2002). Sympatric speciation in phytophagous insects: Moving beyond controversy? *Annual Review of Entomology*, *47*, 773–815. https://doi.org/10.1146/annurev.ento.47.091201.145312

Blais, J., Plenderleith, M., Rico, C., Taylor, M. I., Seehausen, O., van Oosterhout, C., & Turner, G. F. (2009). Assortative mating among Lake Malawi cichlid fish populations is not simply predictable from male nuptial colour. *BMC Evolutionary Biology*, *9*, 53. https://doi.org/10.1186/1471-2148-9-53

Bolnick, D. I., & Fitzpatrick, B. M. (2007). Sympatric speciation: Models and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics*, *38*, 459–487. https://doi.org/10.1146/annurev.ecolsys.38.091206.095804

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

Bryant, D., & Moulton, V. (2004). Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, *21*, 255–265.

Choi, J.Y., Platts, A.E., Fuller, D.Q., Hsing, Y.-I., Wing, R.A., & Purugganan, M.D. (2017). The rice paradox: Multiple origins but single domestication in asian rice. *Molecular Biology and Evolution*, *34*, 969–979.

Cornen, G., Bande, Y., Giresse, P., & Maley, J. (1992). The nature and chronostratigraphy of Quaternary pyroclastic accumulations from lake Barombi Mbo (West-Cameroon). *Journal of Volcanology and Geothermal Research*, *51*, 357–374. https://doi.org/10.1016/0377-0273(92)90108-P

Coyne, J.A., & Orr, H.A. (2004). *Speciation*. Sunderland, MA: Sinauer Associates.

Crapon de Caprona, M.-D., & Ryan, M. J. (1990). Conspecific mate recognition in swordtails, *Xiphophorus nigrensis* and *X. pygmaeus* (Poeciliidae): Olfactory and visual cues. *Animal Behaviour*, *39*, 290–296. https://doi.org/10.1016/S0003-3472(05)80873-5

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFTOOLS. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*, 491–498. https://doi.org/10.1038/ng.806

Dieckmann, U., & Doebeli, M. (1999). On the origin of species by sympatric speciation. *Nature*, *400*, 354–357. https://doi.org/10.1038/22521

Dunz, A. R., & Schliewen, U. K. (2010). Description of a *Tilapia* (*Coptodon*) species flock of Lake Ejagham (Cameroon), including a redescription of *Tilapia deckerti* Thys van den Audenaerde, 1967. *Spixiana*, *33*, 251–280.

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*, 1184–1191. https://doi.org/10.1038/nprot.2009.97

Feder, J. L., Berlocher, S. H., Roethele, J. B., Dambroski, H., Smith, J. J., Perry, W. L., … Aluja, M. (2003). Allopatric genetic origins for sympatric host-plant shifts and race formation in Rhagoletis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 10314–10319. https://doi.org/10.1073/pnas.1730757100

Grant, P.R., & Grant, B.R. (2009). Sympatric speciation, immigration, and hybridization in island birds. In J. B. Losos & R. E. Ricklefs (Eds.), *The theory of island biogeography revisited* (pp. 326–357). Princeton, NJ: Princeton University Press.

Green, J., & Kling, G. W. (1988). The genus *Daphnia* in Cameroon, West Africa. *Hydrobiologia*, *160*, 257–261. https://doi.org/10.1007/BF00007140

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science*, *328*, 710–722. https://doi.org/10.1126/science.1188021

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, *43*, 1031–1034. https://doi.org/10.1038/ng.937

Guo, B., Chain, F. J. J., Bornberg-Bauer, E., Leder, E. H., & Merilä, J. (2013). Genomic divergence between nine- and three-spined sticklebacks. *BMC Genomics*, *14*, 756. https://doi.org/10.1186/1471-2164-14-756

Hadid, Y., Pavlíček, T., Beiles, A., Ianovici, R., Raz, S., & Nevo, E. (2014). Sympatric incipient speciation of spiny mice Acomys at "Evolution Canyon", Israel. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 1043–1048. https://doi.org/10.1073/pnas.1322301111

Hadid, Y., Tzur, S., Pavlíček, T., Šumbera, R., Šklíba, J., Lövy, M., … Nevo, E. (2013). Possible incipient sympatric ecological speciation in blind mole rats (*Spalax*). *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 2587–2592. https://doi.org/10.1073/pnas.1222588110

Heliconius Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, *487*, 94–98.

Hung, C.-M., Shaner, P.-J. L., Zink, R. M., Liu, W.-C., Chu, T.-C., Huang, W.-S., & Li, S.-H. (2014). Drastic population fluctuations explain the rapid extinction of the passenger pigeon. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 10636–10641. https://doi.org/10.1073/pnas.1401526111

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*, 254–267. https://doi.org/10.1093/molbev/msj030

Kagawa, K., & Takimoto, G. (2018). Hybridization can promote adaptive radiation by means of transgressive segregation. *Ecology Letters*, *21*, 264–274. https://doi.org/10.1111/ele.12891

Kautt, A. F., Machado-Schiaffino, G., & Meyer, A. (2016). Multispecies outcomes of sympatric speciation after admixture with the source population in two radiations of nicaraguan crater lake cichlids. *PLOS Genetics*, *12*, e1006157. https://doi.org/10.1371/journal.pgen.1006157

Kautt, A. F., Machado-Schiaffino, G., Torres-Dowdall, J., & Meyer, A. (2016). Incipient sympatric speciation in Midas cichlid fish from the youngest and one of the smallest crater lakes in Nicaragua due to differential use of the benthic and limnetic habitats? *Ecology and Evolution*, *6*, 5342–5357. https://doi.org/10.1002/ece3.2287

Keijman, M. (2010). Tilapia & Co — Enkele onbeschreven en minder bekende Tilapia-soorten globaal voorgesteld. *Cichlidae (Nederlandse Vereniging van Cichlidenliefhebbers)*, *36*, 19–29.

Keller-Costa, T., Canário, A. V. M., & Hubbard, P. C. (2015). Chemical communication in cichlids: A mini-review. *General and Comparative Endocrinology*, *221*, 64–74. https://doi.org/10.1016/j.ygcen.2015.01.001

Kodric-Brown, A., & Strecker, U. (2001). Responses of *Cyprinodon maya* and *C. labiosus* females to visual and olfactory cues of conspecific and heterospecific males. *Biological Journal of the Linnean Society*, *74*, 541–548. https://doi.org/10.1111/j.1095-8312.2001.tb01411.x

Kondrashov, A. S., & Kondrashov, F. A. (1999). Interactions among quantitative traits in the course of sympatric speciation. *Nature*, *400*, 351–354. https://doi.org/10.1038/22514

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., … Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*, 371–375. https://doi.org/10.1038/nature14181

Li, H. (2009). SNPable. Retrieved from http://lh3lh3.users.sourceforge.net/snpable.shtml.

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. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 [q-Bio].

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, *10*, 302. https://doi.org/10.1186/1471-2148-10-302

Mailund, A.T. (2014). Estimating admixture proportions. Retrieved from http://www.mailund.dk/index.php/2014/12/17/estimating-admixture-proportions/

Malinsky, M., Challis, R. J., Tyers, A. M., Schiffels, S., Terai, Y., Ngatunga, B. P., … Turner, G. F. (2015). Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, *350*, 1493–1498. https://doi.org/10.1126/science.aac9927

Martin, C. H. (2012). Weak disruptive selection and incomplete phenotypic divergence in two classic examples of sympatric speciation: Cameroon crater lake cichlids. *The American Naturalist*, *180*, E90–E109. https://doi.org/10.1086/667586

Martin, C. H. (2013). Strong assortative mating by diet, color, size, and morphology but limited progress toward sympatric speciation in a classic example: Cameroon crater lake cichlids. *Evolution*, *67*, 2114–2123. https://doi.org/10.1111/evo.12090

Martin, S. (2015). Genomics_general. Retrieved from https://github.com/simonhmartin/genomics_general

Martin, C. H., Cutler, J. S., Friel, J. P., Dening Touokong, C., Coop, G., & Wainwright, P. C. (2015). Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution*, *69*, 1406–1422. https://doi.org/10.1111/evo.12674

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2015). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, *32*, 244–257. https://doi.org/10.1093/molbev/msu269

Martin, C.H., & Höhna, S. (2018). New evidence for the recent divergence of Devil's Hole pupfish and the plausibility of elevated mutation rates in endangered taxa. *Molecular Ecology*, *27*, 831–838. https://doi.org/10.1111/mec.14404

Martin, C. H., Höhna, S., Crawford, J. E., Turner, B. J., Richards, E. J., & Simons, L. H. (2017). The complex effects of demographic history on the estimation of substitution rate: Concatenated gene analysis results in no more than twofold overestimation. *Proceedings of the Royal Society B: Biological Sciences*, *284*, 20170537. https://doi.org/10.1098/rspb.2017.0537

McLennan, D. A. (2004). Male Brook Sticklebacks' (*Culaea inconstans*) response to olfactory cues. *Behaviour*, *141*, 1411–1424. https://doi.org/10.1163/1568539042948132

McLennan, D. A., & Ryan, M. J. (1999). Interspecific recognition and discrimination based upon olfactory cues in northern swordtails. *Evolution*, *53*, 880–888. https://doi.org/10.1111/j.1558-5646.1999.tb05382.x

McManus, K. F., Kelley, J. L., Song, S., Veeramah, K. R., Woerner, A. E., Stevison, L. S., … Hammer, MF (2015). Inference of gorilla demographic and selective history from whole-genome sequence data. *Molecular Biology and Evolution*, *32*, 600–612. https://doi.org/10.1093/molbev/msu394

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. https://doi.org/10.1038/ncomms14363

Meier, J.I., Marques, D.A., Wagner, C.E., Excoffier, L., & Seehausen, O. (2018). Genomics of parallel ecological speciation in Lake Victoria cichlids. *Molecular Biology and Evolution*, *35*, 1489–1506. https://doi.org/10.1093/molbev/msy051

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, *30*, i541–i548. https://doi.org/10.1093/bioinformatics/btu462

Neumann, D. (2011). Two new sympatric Sarotherodon species (pisces: Cichlidae) endemic to Lake Ejagham, Cameroon, west-central Africa, with comments on the *Sarotherodon galilaeus* species complex. *Zootaxa*, *20*, 5326.

Niimura, Y., & Nei, M. (2005). Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 6039–6044. https://doi.org/10.1073/pnas.0501922102

Nikaido, M., Ota, T., Hirata, T., Suzuki, H., Satta, Y., Aibara, M., … Okada, N. (2014). Multiple episodic evolution events in V1R receptor genes of east-african cichlids. *Genome Biology and Evolution*, *6*, 1135–1144. https://doi.org/10.1093/gbe/evu086

Nikaido, M., Suzuki, H., Toyoda, A., Fujiyama, A., Hagino-Yamagishi, K., Kocher, T. D., … Okada, N. (2013). Lineage-specific expansion of vomeronasal type 2 receptor-like (OlfC) genes in cichlids may contribute to diversification of amino acid detection systems. *Genome Biology and Evolution*, *5*, 711–722. https://doi.org/10.1093/gbe/evt041

Pardo-Diaz, C., Salazar, C., Baxter, S. W., Merot, C., Figueiredo-Ready, W., Joron, M., … Jiggins, C. D. (2012). Adaptive introgression across species boundaries in heliconius butterflies. *PLOS Genetics*, *8*, e1002752. https://doi.org/10.1371/journal.pgen.1002752

Patterson, N. J., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., … Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*, 1065–1093. https://doi.org/10.1534/genetics.112.145037

Pease, J. B., & Hahn, M. W. (2015). Detection and polarization of introgression in a five-taxon phylogeny. *Systematic Biology*, *64*, 651–662. https://doi.org/10.1093/sysbio/syv023

Plenderleith, M., van Oosterhout, C., Robinson, R. L., & Turner, G. F. (2005). Female preference for conspecific males based on olfactory cues in a Lake Malawi cichlid fish. *Biology Letters*, *1*, 411–414. https://doi.org/10.1098/rsbl.2005.0355

Rambaut, A., Suchard, M., Xie, D., & Drummond, A. (2014). TRACER v1.6. Retrieved from http://beast.bio.ed.ac.uk/Tracer

Recknagel, H., Elmer, K. R., & Meyer, A. (2013). A hybrid genetic linkage map of two ecologically and morphologically divergent Midas cichlid fishes (*Amphilophus* spp.) obtained by massively parallel DNA sequencing (ddRADSeq). *G3: Genes, Genomes, Genetics*, *3*, 65–74. https://doi.org/10.1534/g3.112.003897

Richards, E. J., & Martin, C. H. (2017). Adaptive introgression from distant Caribbean islands contributed to the diversification of a microendemic adaptive radiation of trophic specialist pupfishes. *PLOS Genetics*, *13*, e1006919. https://doi.org/10.1371/journal.pgen.1006919

Richards, E., Poelstra, J., & Martin, C. (2017). Don't throw out the sympatric species with the crater lake water: fine-scale investigation of introgression provides weak support for functional role of secondary gene flow in one of the clearest examples of sympatric speciation. BioRxiv 217984.

Ryan, P. G., Bloomer, P., Moloney, C. L., Grant, T. J., & Delport, W. (2007). Ecological speciation in South Atlantic island finches. *Science*, *315*, 1420–1423. https://doi.org/10.1126/science.1138829

Salichos, L., Stamatakis, A., & Rokas, A. (2014). Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution*, *31*, 1261–1271.

Savolainen, V., Anstett, M.-C., Lexer, C., Hutton, I., Clarkson, J. J., Norup, M. V., … Baker, W. J. (2006). Sympatric speciation in palms on an oceanic island. *Nature*, *441*, 210–213. https://doi.org/10.1038/nature04566

Schliewen, U. K., & Klee, B. (2004). Reticulate sympatric speciation in Cameroonian crater lake cichlids. *Frontiers in Zoology*, *1*, 5. https://doi.org/10.1186/1742-9994-1-5

Schliewen, U., Rassmann, K., Markmann, M., Markert, J., Kocher, T., & Tautz, D. (2001). Genetic and ecological divergence of a monophyletic cichlid species pair under fully sympatric conditions in Lake Ejagham, Cameroon. *Molecular Ecology*, *10*, 1471–1488. https://doi.org/10.1046/j.1365-294X.2001.01276.x

Schliewen, U. K., Tautz, D., & Pääbo, S. (1994). Sympatric speciation suggested by monophyly of crater lake cichlids. *Nature*, *368*, 629–632. https://doi.org/10.1038/368629a0

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends in Ecology & Evolution*, *19*, 198–207. https://doi.org/10.1016/j.tree.2004.01.003

Seehausen, O. (2006). African cichlid fish: A model system in adaptive radiation research. *Proceedings of the Royal Society B: Biological Sciences*, *273*, 1987–1998.

Smadja, C., & Butlin, R.K. (2008). On the scent of speciation: The chemosensory system and its role in premating isolation. *Heredity*, *102*, 77–97.

Snow, G. (2016). TEACHINGDEMOS. R package version 2.10.

Sorenson, M. D., Sefc, K. M., & Payne, R. B. (2003). Speciation by host switch in brood parasitic indigobirds. *Nature*, *424*, 928–931. https://doi.org/10.1038/nature01863

Stager, J.C., Alton, K., Martin, C.H., King, D.T., Petruny, L.W., Wiltse, B., & Livingstone, D.A. (2017). On the age and origin of Lake Ejagham, Cameroon, and its endemic fishes. *Quaternary Research*, *89*, 1–12.

Stamatakis, A. (2014). RAXML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stankowski, S., & Streisfeld, M. A. (2015). Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. *Proceedings of the Royal Society B: Biological Sciences*, *282*, 20151666. https://doi.org/10.1098/rspb.2015.1666

Than, C., & Nakhleh, L. (2009). Species tree inference by minimizing deep coalescences. *PLOS Computational Biology*, *5*, e1000501.

Than, C., Ruths, D., & Nakhleh, L. (2008). PHYLONET: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, *9*, 322. https://doi.org/10.1186/1471-2105-9-322

Turelli, M., Barton, N. H., & Coyne, J. A. (2001). Theory and speciation. *Trends in Ecology & Evolution*, *16*, 330–343. https://doi.org/10.1016/S0169-5347(01)02177-2

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., … DePristo, MA (2013). From FASTQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*, 11.10.1–11.10.33.

Wagner, C. E., Harmon, L. J., & Seehausen, O. (2012). Ecological opportunity and sexual selection together predict adaptive radiation. *Nature*, *487*, 366. https://doi.org/10.1038/nature11144

Ward, B. J., & van Oosterhout, C. (2016). Hybridcheck: Software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. *Molecular Ecology Resources*, *16*, 534–539. https://doi.org/10.1111/1755-0998.12469

Wen, D., Yu, Y., Zhu, J., & Nakhleh, L. (2018). Inferring phylogenetic networks using PHYLONET. *System Biology*, *67*, 735–740.

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biology*, *11*, R14. https://doi.org/10.1186/gb-2010-11-2-r14

Zamani, N., Russell, P., Lantz, H., Hoeppner, M. P., Meadows, J. R., Vijay, N., … Grabherr, MG (2013). Unsupervised genome-wide recognition of local relationship patterns. *BMC Genomics*, *14*, 347. https://doi.org/10.1186/1471-2164-14-347

SUPPORTING INFORMATION

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

How to cite this article: Poelstra JW, Richards EJ, Martin CH. Speciation in sympatry with ongoing secondary gene flow and a potential olfactory trigger in a radiation of Cameroon cichlids. *Mol Ecol*. 2018;27:4270–4288. https://doi.org/10.1111/mec.14784