Received: 16 February 2017 | Accepted: 22 June 2017

DOI: 10.1111/mec.14248

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

Demography and genome divergence of lake and stream populations of an East African cichlid fish

Bernd Egger[1] | Marius Roesti[1,2] | Astrid Böhne[1] | Olivia Roth[3] | Walter Salzburger[1]

[1]Zoological Institute, University of Basel, Basel, Switzerland

[2]Department of Zoology, Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada

[3]Evolutionary Ecology of Marine Fishes, Helmholtz Zentrum für Ozeanforschung Kiel (GEOMAR), Kiel, Germany

Correspondence
Walter Salzburger and Bernd Egger, Zoological Institute, University of Basel, Basel, Switzerland.
Emails: walter.salzburger@unibas.ch; bernd.egger@unibas.ch

Funding information
H2020 European Research Council, Grant/Award Number: 201067 and 617585; Volkswagen Stiftung; German Research Foundation (DFG); University of Basel; Swiss National Science Foundation (SNF), Grant/Award Number: 3100A0 138224, 3100A0 156405

Abstract

Disentangling the processes and mechanisms underlying adaptive diversification is facilitated by the comparative study of replicate population pairs that have diverged along a similar environmental gradient. Such a setting is realized in a cichlid fish from southern Lake Tanganyika, *Astatotilapia burtoni*, which occurs within the lake proper as well as in various affluent rivers. Previously, we demonstrated that independent lake and stream populations show similar adaptations to the two habitat regimes. However, little is known about the evolutionary and demographic history of the *A. burtoni* populations in question and the patterns of genome divergence among them. Here, we apply restriction site-associated DNA sequencing (RADseq) to examine the evolutionary history, the population structure and genomic differentiation of lake and stream populations in *A. burtoni*. A phylogenetic reconstruction based on genome-wide molecular data largely resolved the evolutionary relationships among populations, allowing us to re-evaluate the independence of replicate lake–stream population clusters. Further, we detected a strong pattern of isolation by distance, with baseline genomic divergence increasing with geographic distance and decreasing with the level of gene flow between lake and stream populations. Genome divergence patterns were heterogeneous and inconsistent among lake-stream population clusters, which is explained by differences in divergence times, levels of gene flow and local selection regimes. In line with the latter, we only detected consistent outlier loci when the most divergent lake–stream population pair was excluded. Several of the thus identified candidate genes have inferred functions in immune and neuronal systems and show differences in gene expression between lake and stream populations.

KEYWORDS
Astatotilapia burtoni, demography, genome divergence, Lake Tanganyika, phylogeny

1 | INTRODUCTION

Since the inception of evolutionary biology, natural selection is acknowledged as a main driver for the divergence of populations and, ultimately, the emergence of novel species (Darwin, 1859; Dobzhansky, 1937; Mayr, 1942). Manifold examples demonstrate that divergent selection is a fundamental evolutionary force responsible for genetic differentiation among populations (reviewed in Schluter, 2000, 2009; Nosil, 2012). The phenotypic differences resulting from local adaptation might eventually facilitate reproductive isolation among populations, up to the point where speciation is complete (Rundle & Nosil, 2005; Schluter, 2009).

In the last few years, genomic approaches have increasingly been utilized to study the molecular underpinnings of diversification (reviewed in Seehausen et al., 2014; Berner & Salzburger, 2015). Examining the patterns of genome-wide divergence between ecologically

distinct populations (or between sister species) has provided important insights into the genomics of adaptive divergence (e.g., Seehausen et al., 2014; Wolf, Lindell, & Backstrom, 2010), the evolution of reproductive barriers during speciation (e.g., Ellegren et al., 2012; Gagnaire, Pavey, Normandeau, & Bernatchez, 2013; Renaut et al., 2013), the role of gene flow in organismal diversification (e.g., Feder, Flaxman, Egan, & Nosil, 2013; Gante et al., 2016; Martin et al., 2013), as well as the nature of genomic architectural features connected to diversification (e.g., Berg et al., 2016; Lohse, Clarke, Ritchie, & Etges, 2015). A common outcome from these studies is that genetic differentiation appears to be heterogeneous across the genome, that is, there are regions in the genome displaying low levels of genetic differentiation, while others are highly differentiated (e.g., Carneiro, Blanco-Aguiar, Villafuerte, Ferrand, & Nachman, 2010; Harr, 2006; Malinsky et al., 2015; Nadeau et al., 2012; Roesti, Hendry, Salzburger, & Berner, 2012). In all these cases, only small fractions of the respective genome turned out to be highly differentiated. Such regions, often termed "genomic islands of speciation," are thought to point to the loci involved in reproductive isolation and ecological specialization (Wu, 2001). Heterogeneity is particularly pronounced between parapatric (and sympatric) populations/species pairs, where occasional gene flow is expected to homogenize the genome except at the loci under divergent selection and their close-by neutral genomic neighborhood (e.g., Emelianov, Marec, & Mallet, 2004; Nosil, Harmon, & Seehausen, 2009; Yeaman & Whitlock, 2011). Geographically isolated (allopatric) populations, on the other hand, are predicted to feature a greater number of differentiated loci, as a consequence of the combined action of selection *and* genetic drift (Feder et al., 2013).

A particular focus has been devoted to the identification of the genes involved in adaptive diversification and speciation (e.g., Colosimo et al., 2005; Lamichhaney et al., 2015). A commonly applied strategy to identify such loci in genome scans is the so-called outlier analysis, in which a certain threshold is applied to a divergence measure (e.g., F_{ST}) across all markers and the markers with the highest values are inspected further, for example, by examining genes linked to the outliers. Due to factors such as mutation, demographic perturbation, recombination rate variation and linked selection, such tests are prone to detect false outlier (e.g., Haasl & Payseur, 2016). On the other hand, the so identified candidate genes serve as a good basis for further examinations, for example, in the form of gene expression assays and/or functional experiments (Faria et al., 2014).

Disentangling the processes and mechanisms underlying divergent selection in diversification is greatly facilitated by the comparative study of multiple population pairs that have repeatedly diverged along the same environmental gradient and that show varying levels of reproductive isolation among them (e.g., Faria et al., 2014; Berner & Salzburger, 2015; but see Nosil, Feder, Flaxman, & Gompert, 2017). Comparing replicate them not only permits the examination of the context dependence of adaptation and reduces the risk of false outlier detection, but enables—in those cases where similar selection regimes resulted into similar phenotypes—the investigation of parallel evolution at the molecular level (Berner & Salzburger, 2015; Wolf & Ellegren, 2017). Prominent examples of natural systems offering replicate conditions comprise *Heliconius* butterflies (Nadeau et al., 2012), threespine stickleback fish (Jones et al., 2012; Roesti et al., 2012; Schluter & McPhail, 1992) and lake whitefish (Gagnaire et al., 2013).

We have recently introduced a cichlid fish model system, the East African haplochromine *Astatotilapia burtoni*, for the study of the factors that enhance and/or constrain diversification (Theis, Ronco, Indermaur, Salzburger, & Egger, 2014; Theis et al., 2017). This generalistic species is one of very few cichlid species that occur both within a large lake in East Africa—in this case Lake Tanganyika—as well as in various affluent rivers. In southern Lake Tanganyika, for example, a series of replicate population clusters of *A. burtoni* can be found, each of them consisting of one lake and one to several riverine population(s) (Figure 1a). These lake and stream "population pairs" show similar adaptations to divergent selection regimes (Theis et al., 2014, 2017): (i) stream fish have a shallower body compared to lake fish; this divergence in body shape is associated with different flow regimes in the two habitat types. (ii) Lake fish have a more superior mouth position and possess longer gill rakers plus slender and more elongated lower pharyngeal jaw bones compared to stream fish; these shifts in trophic structures are linked to differential resource use in the two habitat types. Notably, the trait differences among lake and stream populations reported under (i) and (ii) do not reflect pure plastic responses to different environmental conditions, but have a substantial genetic component. (iii) Stream fish feature fewer but larger and more conspicuous egg-spots—that is, an anal fin pigmentation trait in the form of ovoid markings important in intraspecific interactions; the difference between stream and lake fish is probably to maintain signal efficiency in the different light environments. Taken together, the setting of replicate lake–stream population pairs in *A. burtoni* offers the rare opportunity to examine adaptive differentiation along an environmental gradient in East African cichlid fish, which are otherwise restricted to one of these habitat types.

Our initial study further revealed an unexpectedly high degree of genetic diversity among the 22 populations examined from the southern part of Lake Tanganyika, and we reported a deep split between populations from the eastern shoreline, the western shoreline and the headwaters of the Lufubu River (Theis et al., 2014). However, genetic differentiation, population structure and phylogenetic relationships were inferred from microsatellite and mtDNA markers only, limiting the informative value of our previous study. No information is, as of yet, available regarding the demographic histories of the different lake and stream populations, nor on the patterns of genomic differentiation along the lake–stream environmental gradient in *A. burtoni*. This kind of information would be crucial to (i) understand the dynamics of divergence along the lake–stream environmental gradient; (ii) identify source populations and reconstruct the direction of colonization (lake into streams or vice versa); (iii) examine the patterns of differentiation in the genome associated with divergence along a habitat

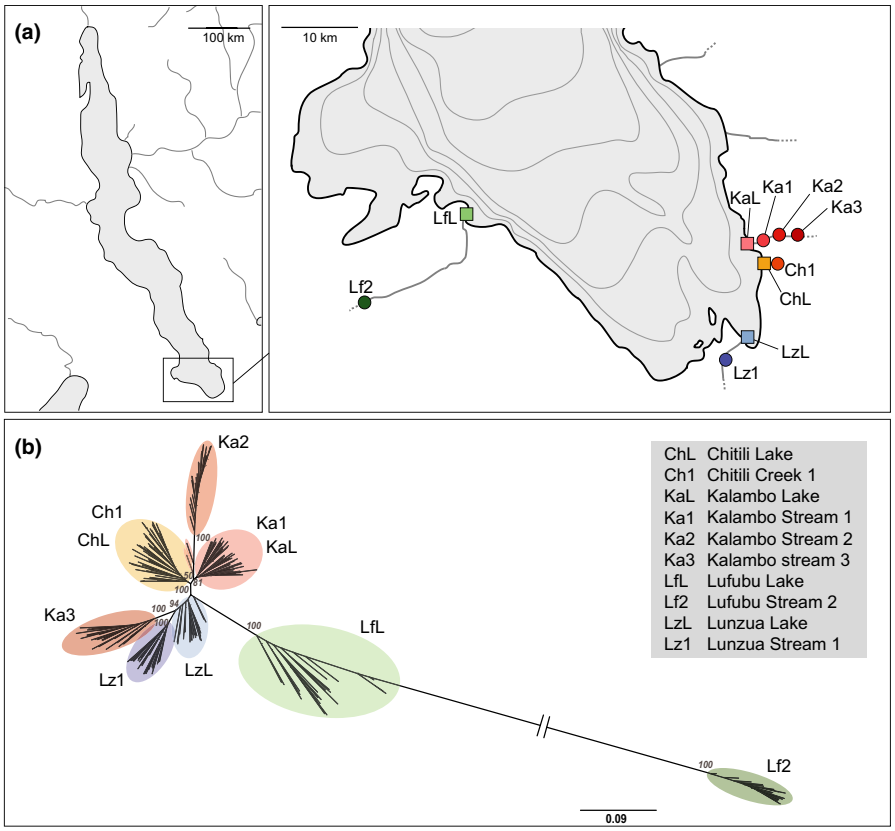

FIGURE 1 Sampling locations and phylogeny of study populations. (a) The 10 sampling localities in the southern part of lake tanganyika (LT). Squares represent lake and circles stream populations; full names of localities are listed in the grey box in (b). (b) Unrooted maximum-likelihood tree based on 9,529 SNPs. Bootstrap support in per cent is given for the key nodes

gradient; as well as (iv) identify candidate loci, and ultimately the molecular mechanisms, involved in adaptive divergence.

In this study, we apply restriction site-associated DNA sequencing (RADseq) to examine the evolutionary history, the population structure and the patterns of genomic differentiation of *A. burtoni* lake and stream populations in the south of Lake Tanganyika. We first establish a phylogenetic hypothesis based on the maximum-likelihood method and 9,529 SNP markers, and examine the demographic history of the ten lake and stream populations under investigation. We then explore the population structure via nearest neighbour haplotype co-ancestry analyses. With this information at hand, we propose a colonization scenario for *A. burtoni* in southern Lake Tanganyika and confirm the evolutionary independency of lake–stream replicates. We then evaluate patterns of genomic differentiation among evolutionary-independent lake–stream replicates and show that geographically adjacent populations are generally more admixed genetically. Further, we ask whether the genetic basis of divergence is shared across the lake–stream gradient, and identify consistent outlier loci across independent lake–stream replicates. Finally, we conduct gene expression analyses for the thus identified candidate genes and show that candidate genes with inferred functions in immune and neuronal systems show differences in gene expression between lake and stream populations.

2 | MATERIALS AND METHODS

2.1 | Study sites and sampling strategy

Sampling was carried out between July 2011 and July 2013 in the southern basin of Lake Tanganyika as well as in inflowing rivers and streams, using hook and line fishing, minnow traps and gill nets under the permission of the Lake Tanganyika Research Unit, Department of Fisheries, Republic of Zambia. Fish were collected from four lake–stream systems: the Kalambo River at locations KaL, Ka1, Ka2, and Ka3; the Chitili Creek, ChL and Ch1; the Lunzua River, LzL and Lz1; and the Lufubu River, LfL and Lf2 (see Figure 1a and Table 1 for GPS coordinates). We collected 25 adult individuals each from KaL, Ka1, Ka2, Ka3, ChL, Ch1, LzL, Lz1 and 20 individuals each from LfL and Lf2 for subsequent RADseq, resulting in a total number of 240 individuals from 10 populations. Each specimen was measured, weighted and photographed in the field; a fin clip was taken as tissue sample and stored in 96% ethanol. In addition, between 9 and 13 individuals were sampled from populations KaL, Ka1, Ka2, LzL, Lz1, Lfl and Lf2 for gene expression analysis. To this end, gill rakers were dissected in the field and immediately stored in RNAlater (see Table 1 for details).

TABLE 1 Geographic coordinates of sampling locations, and sample size details for RADseq and gene expression analyses

		GPS coordinates		Sample size RADseq			Sample size qPCR			
Location	Habitat	Latitude	Longitude	Males	Females	Total	Males	Females	Juveniles	Total
KaL	Lake	8°36′6.27″S	31°11′13.24″E	14	11	25	7	6	0	13
Ka1	Stream	8°35′35.23″S	31°11′6.18″E	13	12	25	6	4	0	10
Ka2	Stream	8°35′6.24″S	31°12′29.32″E	12	13	25	6	6	0	12
Ka3	Stream	8°35′41.59″S	31°14′50.32″E	13	12	25	8	4	0	12
ChL	Lake	8°38′18.42″S	31°11′55.34″E	11	14	25	na	na	na	na
Ch1	Stream	8°38′16.91″S	31°12′4.02″E	13	12	25	na	na	na	na
LzL	Lake	8°44′57.13″S	31°10′21.86″E	12	13	25	6	6	0	12
Lz1	Stream	8°47′23.51″S	31°8′14.33″E	13	12	25	7	2	0	9
LfL	Lake	8°33′36.56″S	30°43′33.79″E	17	3	20	8	4	0	12
Lf2	Stream	8°41′9.37″S	30°33′51.90″E	10	10	20	5	2	4	11

2.2 | RADseq library preparation and marker generation

Total DNA was extracted from fin clips preserved in ethanol applying a proteinase K digestion followed by a MagNA Pure extraction using a robotic device (MagNA Pure LC278, Roche Diagnostics, Switzerland) and following the manufacturer's protocol. Libraries for RADseq were prepared following the protocol described in Roesti et al. (2012). In brief, genomic DNA was digested with the *Sbf1* restriction enzyme and 5-mer barcoded. Libraries from 40 individuals were pooled and single-end-sequenced to 100-bp reads in six Illumina HiSeq2000 lanes at the Genomics Facility Basel jointly operated by ETH Zurich Department of Biosystems Science and Engineering and the University of Basel (see Table S5 for information on read numbers and coverage per specimen). Illumina reads are available from the Sequence Read Archive (SRA) at NCBI under the Accession nos SRX2967972–SRX2968211.

The raw Illumina sequence reads were quality-filtered, sorted according to barcode and aligned to the *Astatotilapia burtoni* reference genome (release Broad HAPBUR1.0, Brawand et al., 2014) using NOVOALIGN v2.08.03 (http://novocraft.com). We accepted a total of 6 to 8 high-quality mismatches and/or indels along a read (flags: −t200, −g40, −x15). Alignments were converted to BAM format using SAMTOOLS v0.1.18 (Li et al., 2009). Consensus genotypes at individual RAD loci were determined using a "genotype-haplotype" (sensu Nevado, Ramos-Onsins, & Perez-Enciso, 2014) calling approach introduced by Roesti, Kueng, Moser, and Berner (2015). Diploids were called if the dominant haplotype occurred in at least 18 copies. A lighter representation of the dominant haplotype resulted in a haploid call, provided this haplotype was still present in more than two copies. For diploid loci, a RAD locus was considered heterozygous if the ratio of the dominant to the second most frequent haplotype was lower than 0.25. To avoid the unspecific alignment of sequence reads to several sites in the genome, we excluded RAD loci with a sequence coverage exceeding 3.5 times the expected mean coverage across all genome-wide RAD loci (Roesti et al., 2015). Single nucleotide polymorphisms (SNPs) with insufficient representation across individuals (threshold: 40 nucleotide calls from each population) and individuals with more than 75% missing data per haplotype were also excluded from further analyses. The resulting matrix comprised 61,291 SNPs in 228 individuals.

2.3 | Phylogenetic analyses

To obtain phylogenetic relationships among the individuals from the ten lake and stream populations, the initial SNP matrix was again quality-filtered by excluding SNPs with more than 20% missing data across all individuals. Furthermore, the SNP matrix was reduced to one SNP per RADtag only and filtered for SNPs with a minor allele frequency (MAF) >0.05 across all populations, before individual genotypes were transformed to single-letter code. This resulted in a data set containing 9,529 SNPs in 228 individuals. We then used the PHANGORN package (Schliep, 2011) in R (version 3.1.2; R Core Team, 2012) to determine the most appropriate substitution model (GTR+G) and to generate an unrooted maximum-likelihood tree. Bootstrap support was calculated on the basis of 200 replicates.

2.4 | Demographic analyses

Demographic parameters for *A. burtoni* lake and stream populations were estimated based on the observed joint site frequency spectrum (SFS) using the software FASTSIMCOAL 2.1 (Excoffier, Dupanloup, Huerta-Sánchez, Sousa, & Foll, 2013) and following the strategy described in Roesti et al. (2015) with some modifications. The SFS was calculated for six pairwise lake–stream population comparisons (KaL vs. Ka1, KaL vs. Ka2, KaL vs. Ka3, ChL vs. Ch1, LzL vs. Lz1, LfL vs. Lf2), four lake–lake population comparisons (LfL vs. KaL, LfL vs. LzL, ChL vs. KaL, LzL vs. KaL) and three stream–stream comparisons (Lf2 vs. Ka2, Lz1 vs. Ka2, Ka3 vs. Lz1). To this end, 30 haploid consensus genotypes per RAD locus were sampled randomly from both the lake and the stream population of each of the 13 population pairs. Loci with a lower coverage and with more than two polymorphisms with identical MAF across the last 30 positions were ignored. The latter excluded uninformative sequential pseudo-SNPs from

RAD loci harbouring a micro-indel polymorphism to ensure that only true SNPs were considered (see Roesti et al., 2015). To generate the SFS, the occurrence of the minor allele at each of the 88 positions per RAD locus in each population was counted, thereby considering both monomorphic and biallelic SNPs. The numbers of base positions and RAD loci for each joint SFS are shown in Table S1. The pairwise joint SFS was then used to estimate divergence times, effective population sizes and migration rates, applying a divergence with geneflow model. As no genome-wide mutation rate estimate for haplochromine cichlids is available to date, we applied the human mutation rate (Ségurel, Wyman, & Przeworski, 2014), assuming that it is similar to the one in cichlids (see Malinsky et al., 2015). For each of the 13 lake–stream population comparisons, 80 replicate runs including 40 estimation loops with 100,000 coalescence simulations were performed. The best parameter estimates were determined by selecting those 10 runs with the smallest difference between the estimated and observed likelihood (Δlikelihood) for each comparison. We then used this subset to calculate the mean and the 95% confidence intervals (95 percentiles from bootstrap distributions based on 100,000 resamples) for all estimated parameters for each comparison.

2.5 | Population genomic analyses

We used the program FINERADSTRUCTURE (v0.1; Malinsky, Trucchi, Lawson, & Falush, 2016) to infer population structure via shared ancestry among *A. burtoni* individuals from lake and stream sampling locations. This program is a modified version of the FINESTRUCTURE package (Lawson, Hellenthal, Myers, & Falush, 2012), specifically adopted for RADseq data, and does not require information about location of loci on chromosomes or phased haplotypes. To this end, the original SNP matrix was quality-filtered by only allowing 10% missing data per SNP across all individuals, resulting in a matrix comprising 16,998 SNPs. SNPs from the same RADtag were merged using a custom R script to generate the input file. We then ran RADPAINTER, implemented in the FINERADSTRUCTURE package, to calculate the co-ancestry matrix. As a next step, individuals were assigned to populations, with a burn-in period of 100,000 and 100,000 Markov chain Monte Carlo iterations. Tree building was performed using default parameters. To visualize the results, we used the R scripts FINERADSTRUCTUREPLOT.R and FINESTRUCTURELIBRARY.R (available at http://cichlid.gurdon.cam.ac.uk/fineRADstructure.html).

We then calculated F_{ST} values based on haplotype diversity (Nei & Tajima, 1981; equation 7) using all informative SNPs. For RAD loci that comprised multiple SNPs, only the one yielding the highest F_{ST} value was retained. The number of polymorphic SNPs per population comparison ranged between 3,840 and 5,957 (see Table 2 for details).

As the *A. burtoni* reference genome is not assembled into chromosomes, we projected our RADtags onto the Nile tilapia reference genome (*Oreochromis niloticus*; ORENIL 1.0 at Ensembl; GENBANK Assembly ID GCA_000188235.1) to visualize patterns of divergence within pairwise population comparisons along chromosomes. To this end, we generated pseudolinkage groups using the information provided by SATSUMA (Grabherr et al., 2010) synteny mapping between *A. burtoni* genomic scaffolds and the *O. niloticus* reference genome (provided by BROAD as part of the cichlid genome sequencing project). Projection onto the 22 Nile tilapia linkage groups resulted in a marked reduction in SNPs, ranging between 1,421 and 2,595 polymorphic sites in pairwise comparisons. Note that *A. burtoni* has a different karyotype (1n=20) compared to *O. niloticus* (1n=22) due to two fused chromosomes (Mazzuchelli, Kocher, Yang, & Martins, 2012).

TABLE 2 The number of polymorphic loci and genome-wide median F_{ST} values for each lake–stream population comparison

Comparison	Number of loci	Median F_{ST}
KaL vs. Ka1	3,887	0
KaL vs. Ka2	3,875	0.051
KaL vs. Ka3	3,869	0.056
ChL vs. Ch1	3,840	0
LzL vs. Lz1	4,094	0.012
LfL vs. Lf2	5,957	0.407

To test for isolation by distance (IBD), we conducted a Mantel test in R (package ECODIST v1.2.9; Goslee, & Urban, 2007) using the median genome-wide F_{ST} and the geographic distance in metres between sample sites measured in Google Earth (Table S2). Partial Mantel tests were applied to compare differences in phenotypic traits between lake and stream populations with the corresponding genome-wide median F_{ST}, while correcting for geographic distances ("isolation by adaptation"; see Nosil, 2012). To this end, we used Mahalanobis distances for body shape, mouth position and lower pharyngeal jaw bone, as well as metric measurements for gill rakers taken from Theis et al. (2014).

2.6 | F_{ST} outlier detection

To determine F_{ST} outlier loci, we screened for F_{ST} values that were above the baseline divergence (calculated as the median F_{ST} across all loci) in each lake–stream comparison, but that were equal to or below the baseline divergence in lake–lake comparisons. For each detected outlier SNP, a 10-kb window up- and downstream of the locus was extracted from the *A. burtoni* genome and back-blasted against the *A. burtoni* and *O. niloticus* genomes using BLASTN with default settings (http://blast.ncbi.nlm.nih.gov). We then exported all annotated genes within these 20-kb windows, as well as the most adjacent up- and downstream genes (Table S3). The putative function of these genes was then assessed from annotations of their human orthologs in UNIPROT (the most complete functional data set available to date; http://www.uniprot.org). For genes annotated as "uncharacterized," a BLASTP was performed against the nr database with default settings (https://blast.ncbi.nlm.nih.gov); functional annotation was subsequently derived from the, respectively, best blast hit.

2.7 | Gene expression assays

RNA extractions from gill tissue stored in RNAlater were performed with the RNeasy 96 Universal Tissue Kit (Qiagen) following the manufacturer's protocol. RNA yield was measured by spectrometry (NanoDrop ND-1000; peQLab); a total of 1,200 ng (diluted in 6 μl) were used for reverse transcription with QuantiTect®Reverse-Transcription Kit (Qiagen). Primers for candidate genes in F_{ST} outlier regions were designed on the *A. burtoni* coding sequence, using PRIMER-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) and placed to span an exon–exon boundary to disable amplification from potential DNA contamination. Pooled cDNA was used to test primer efficiency using 5× HOT FIREPol® EvaGreen® qPCR Mix Plus (ROX) (Solis BioDyne). In total, we tested primers for 34 candidate genes on a StepOnePlus (Thermo Fisher Scientific), allowing efficiencies between 90% and 100% and a slope of the standard curves of log quality vs. threshold cycle (Ct) between −3.5 and −3.2 (for a list of all primers, see Table S3). Twenty-four of the 32 candidate gene primer pairs passed this quality filtering. The expression of these 24 genes was then measured simultaneously for all RNA samples using a BioMark™ HD system (Fluidigm, South San Francisco, CA, USA) based on 96.96 dynamic arrays (GE chips). According to Beemelmanns and Roth (2016), preamplification products were diluted 1:10. Sample and assay mix were filled into the GE chips and measured in the BioMark system, applying the GE-fast 96.96 PCR protocol according to the manufacturer's instructions (Fluidigm). The chip included controls without template (NTC), controls for gDNA contamination (−RT) and standards and three technical replicates per sample and gene.

For each of the three technical replicates per sample, the Ct, the standard deviation (*SD*) and the coefficient of variance (CV) were calculated. Samples with a CV larger than 4% were removed due to potential measurement errors (see Bookout & Mangelsdorf, 2003). The genes *slc22a15* and *plaa* showed the highest stability in expression among samples (geNorm $M > 0.85$) (Hellemans, Mortier, de Paepe, Speleman, & Vandesompele, 2007). Their geometric mean was, thus, used to quantify relative expression of each target gene by calculating −ΔCt values. Data analysis was performed in R (v3.2.2). Statistical univariate approaches were applied to test for differences in expression between lake and stream populations within systems for each gene. An ANOVA (NMLE package—LMER function in R) was fitted using population as fixed factor. Prior to the analysis, data and residuals were tested for normal distribution and variance homogeneity (Shapiro–Wilk test, Levene's test). ANOVAs and post hoc Tukey HSD tests were used to test for differences between lake and stream populations within systems. For visualization purposes, we generated heatmaps depicting relative differences in gene expression with the R package NMF (version 0.20.6; Gaujoux & Seoighe, 2010). For normalization (−ΔΔCt), the −ΔCt value of each sample per gene was subtracted from the average −ΔCt value of that gene over all samples. Means of −ΔΔCt values of either the significant main effects or the interactions are shown.

3 | RESULTS

3.1 | Genome-wide phylogeny and population structure of *A. burtoni* lake and stream populations

The unrooted maximum-likelihood phylogeny based on 9,529 RAD loci (Figure 1b) revealed a deep split between the Lufubu stream population (Lf2) and the remaining populations including the specimens collected at the Lufubu lake site (LfL). The populations from the eastern shoreline of Lake Tanganyika clustered together. Not all lake–stream population pairs were resolved as, monophyletic: the most upstream population from the Kalambo system (Ka3), situated above a more than 220 m high waterfall, was grouped together with the populations from the Lunzua River system (LzL and Lz1).

The clustered co-ancestry matrix and the cladogram resulting from the FINERADSTRUCTURE analysis (Figure 2) confirmed these findings: the upstream Lufubu population (Lf2) formed the most distinct cluster and there was substantial co-ancestry sharing among the Lf2 specimens (as indicated by purple and blue coloration in Figure 2). A strong signal of co-ancestry sharing was also observed between the Lf2 individuals and fish caught at the Lufubu lake site (LfL; red coloration in Figure 2), whereas very low levels of co-ancestry sharing were found between Lf2 and all remaining populations (yellow coloration in Figure 2). Additionally, according to FINERADSTRUCTURE, we detected subpopulation structure within Lf2. Co-ancestry sharing among the individuals from LfL reached similar levels as between the individuals from LfL and Lf2. Within the eastern populations, the Lunzua lake and stream populations (LzL and Lz1) and the Kalambo upstream populations (Ka2 and Ka3) each form distinct genetic clusters (i.e., all individuals from a sampling location were assigned to one genetic cluster), whereas geographically close populations from the Chitili (ChL and Ch1) and the Kalambo River (KaL and Ka1) were indicated to be admixed. The FINERADSTRUCTURE analysis also confirmed the genetic affinity between Ka3 and Lz1, as indicated by relatively high co-ancestry sharing between individuals from these two populations.

The level of overall baseline genomic divergence varied substantially among the *A. burtoni* lake–stream population pairs, and also among the different populations of the Kalambo River system (the only river system with more than one sampled stream population), ranging from zero median F_{ST} in systems with geographically proximate lake and stream populations (KaL vs. Ka1, ChL vs. Ch1), to intermediate differentiation in the Lunzua system, and substantial genomic differentiation (median $F_{ST} = 0.407$, 1% of loci fixed) between the geographically most distant lake and stream population pair from the Lufubu system (Table 2).

There was a strong pattern of isolation by distance when all lake and stream populations were included (Mantel-R = 0.8001, $p = .0010$), as well as when the population Ka3 was excluded (Mantel-R = 0.8018, $p = .0010$). Partial Mantel tests indicated a pattern of isolation by adaptation with regard to body shape (Mantel-R = 0.4768, $p = .0390$), but not for mouth position (Mantel-

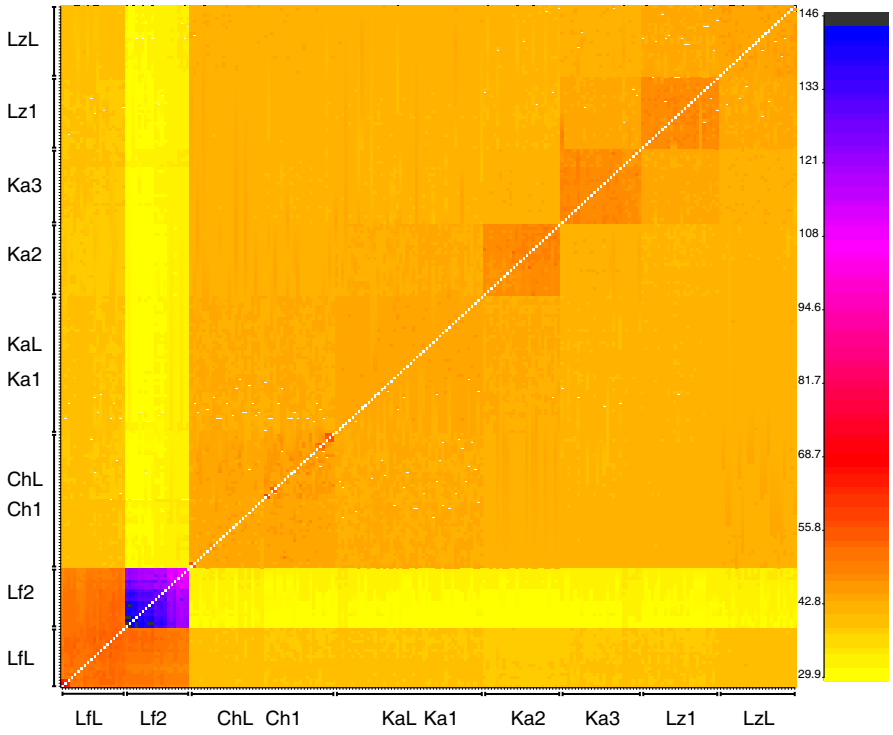

FIGURE 2 Clustered FINERADSTRUCTURE co-ancestry matrix. The highest levels of co-ancestry is evident among individuals from the Lufubu stream population (Lf2), indicated by black, blue and purple colours. The lowest levels of co-ancestry sharing are given between Lf2 and the eastern populations (KaL, Ka1, Ka2, Ka3, ChL, Ch1, LzL and Lz1), indicated by yellow coloration

R = 0.2353, p = .1760), pharyngeal jaw bone (Mantel-R = −0.1224, p = .6670), nor gill raker length (Mantel-R = 0.1368, p = .2240).

3.2 | Demographic analyses

To explore the demographic history of lake–stream population pairs, we ran coalescent simulations based on joint allele site frequency spectra and applying a divergence with geneflow model. Means and 95% confidence intervals of parameter estimates for lake–stream comparisons are shown in Table 3, and for lake–lake and stream–stream comparisons in Table S4. In general, our coalescent simulations revealed that migration rates decreased with geographic distance between lake and stream populations; that is, migration rate was lowest between the geographically most distant lake and stream populations of the Lufubu system (LfL and Lf2), yet substantial between the geographically most adjacent lake and stream populations of the Chitili system (ChL and Ch1) and between KaL and Ka1 (see Table 3). Moreover—with the exception of the comparisons ChL and Ch1 and KaL and Ka1, which each essentially form a panmictic population according to FINERADSTRUCTURE (Figure 2) and show a median F_{ST} of 0 (Table 2)—estimated migration rates were higher in the direction stream into lake as compared to the opposite direction. With the exception of the Lufubu system, estimated ancestral population sizes roughly scaled with the size of the river system under investigation; that is, they were largest in the Kalambo River system and smallest in the Chitili and Lunzua systems. Effective population sizes were generally lower in stream compared to lake populations (with the exception of the KaL and Ka1 comparison, but recall that these are basically panmictic). Estimated divergence times indicated that the oldest split of a lake and stream population from an ancestral population in our study area occurred within the Lufubu system (181,336 generations ago, Table 3), followed by the Chitili and the Lunzua systems.

Comparisons among lake population samples indicated asymmetric gene flow between the Lufubu lake population (LfL, on the western shoreline of lake tanganyika [LT]) and the eastern lake populations (KaL and LzL), with substantially higher migration rates in the direction west towards east (Table S4). In contrast, migration rates were balanced within eastern lake population comparisons (ChL and KaL, LzL and KaL). Divergence times in lake–lake comparisons roughly scaled with geographic distance. Within stream–stream comparisons, the lowest migration rates and the oldest split were found between Lf2 and Ka2, again with higher levels of migration in the direction west to east (from Lf2 towards Ka2). Comparisons among riverine populations from the Kalambo (Ka2 or Ka3) and Lunzua rivers inferred slightly higher migration rates from Lz1 into either the Ka3 or Ka2, and very recent divergence times (2,792 generations ago for Ka3 and Lz1 and 9,233 generations ago for Lz1 and Ka2).

3.3 | Patterns of genome divergence and F_{ST} outlier screening

The patterns of genome divergence differed substantially between replicate lake–stream population pairs in *A. burtoni* (Figure 3). Low levels of genome divergence between lake and stream populations, with comparably few outlier loci, were found in the KaL vs. Ka1 (the maximum F_{ST} value observed for a single SNP was 0.35) and the ChL vs. Ch1 comparison (maximum F_{ST} = 0.31); moderate levels of genome divergence were found between LzL vs. Lz1 (maximum F_{ST} = 0.76) and KaL vs. Ka2 (maximum F_{ST} = 0.86), whereas high levels of divergence with maximum F_{ST} values of 1 were detected in the Lufubu comparison (Figure 3). (Note that we refrained from the

TABLE 3 Results from the demographic analysis. Ancestral effective population sizes and effective population sizes of lake and stream populations, bidirectional migration rates between each lake and stream population pair and estimated ages of the split of lake and stream populations from their common ancestor

Population pair	Ne (ancestor)	Ne (lake)	Ne (stream)	*m* (lake->stream)	*m* (stream->lake)	TDIV
KaL & Ka1						
Mean	36,850	28,750	46,748	5.23E-03	1.83E-03	20,232
Lower CI	34,903	18,456	35,412	2.92E-03	6.80E-05	15,550
Upper CI	38,445	40,739	56,923	7.35E-03	3.70E-03	26,183
KaL & Ka2						
Mean	37,893	51,726	8,967	6.02E-05	2.02E-04	13,584
Lower CI	37,301	50,780	8,754	5.67E-05	1.97E-04	12,072
Upper CI	38,454	52,547	9,176	6.34E-05	2.10E-04	15,342
KaL & Ka3						
Mean	42,431	47,326	16,785	2.60E-05	4.81E-05	7,987
Lower CI	41,634	46,583	16,495	2.43E-05	4.41E-05	7,469
Upper CI	43,260	48,180	17,078	2.74E-05	5.19E-05	8,524
ChL & Ch1						
Mean	19,772	54,064	1,857	4.64E-04	2.85E-02	52,659
Lower CI	11,129	53,011	1,023	2.70E-04	1.26E-02	40,141
Upper CI	27,704	55,153	2,690	6.61E-04	4.76E-02	65,370
LzL & Lz1						
Mean	25,647	119,730	1,382	1.99E-04	7.40E-03	43,084
Lower CI	13,947	108,556	588	1.36E-04	2.72E-03	24,359
Upper CI	36,509	132,309	2,478	2.60E-04	1.45E-02	62,850
LfL & Lf2						
Mean	6,975	48,434	8,438	7.51E-06	3.78E-05	181,336
Lower CI	1,033	47,700	8,210	7.24E-06	3.65E-05	161,925
Upper CI	15,963	49,133	8,683	7.79E-06	3.91E-05	213,190

comparison between KaL and Ka3 on the basis of the distinctiveness of Ka3 from all other populations of the Kalambo River system, which is suggestive of an independent origin of the *A. burtoni* stocks at the Ka3 site.) In all pairwise comparisons, multiple F_{ST} outliers distributed across the entire genome were observed.

When comparing across all four lake–stream population pairs (ChL vs. Ch1, KaL vs. Ka1, LfL vs. Lf2, and LzL vs. Lz1), we did not find a single consistent RAD outlier locus. However, when the very divergent Lufubu system was excluded (i.e., focusing on ChL vs. Ch1, KaL vs. Ka1, KaL vs. Ka2 and LzL vs. Lz1), eight outlier loci were consistently retrieved. We identified a total of 32 annotated genes in the genomic regions surrounding these eight outlier SNPs (with a minimum of two and a maximum of six genes per region; see Table S3). Of the 32 genes, six genes were tentatively implicated with neuronal and another six genes with immune functions; five genes were uncharacterized (named "uncharact1" to "uncharact5") so that no function could be assigned.

3.4 | Gene expression

For quantitative real-time PCR analysis, the Lufubu lake and stream populations were included, as our aim was to test for habitat-specific gene expression, which benefits from including more replicates (but note that we also present statistical analyses excluding the populations LfL and Lf2 in Fig. S1).

We detected a general pattern of relatively higher levels of gene expression in stream populations compared to lake populations within the Lunzua and Lufubu systems, whereas lake and stream populations from the Kalambo River system showed less pronounced differences (Figure 4; Fig. S2). In more detail, gene-by-gene analyses revealed that of 24 genes in total, five genes showed differential gene expression between LzL and Lz1 (*Flec4*, *fam83a*, *PERK4*, *glud1* and "uncharact5"; placed in two outlier regions), five genes between LfL and Lf2 (*glud1*, *zdhhc21*, *PERK4*, *dennd4c* and *haus6*; placed in two outlier regions, one shared with the Lunzua system) and three genes between KaL and Ka2 ("uncharact2," "uncharact3" and *pro-MCH*; placed on two different scaffolds) (Figure 4; Fig. S2 and Table S3).

4 | DISCUSSION

In this study, we applied RADseq to infer phylogenetic relationships, demographic histories, the population structure and patterns of genomic divergence among lake and stream populations of the haplochromine cichlid fish *Astatotilapia burtoni* from the southern part of

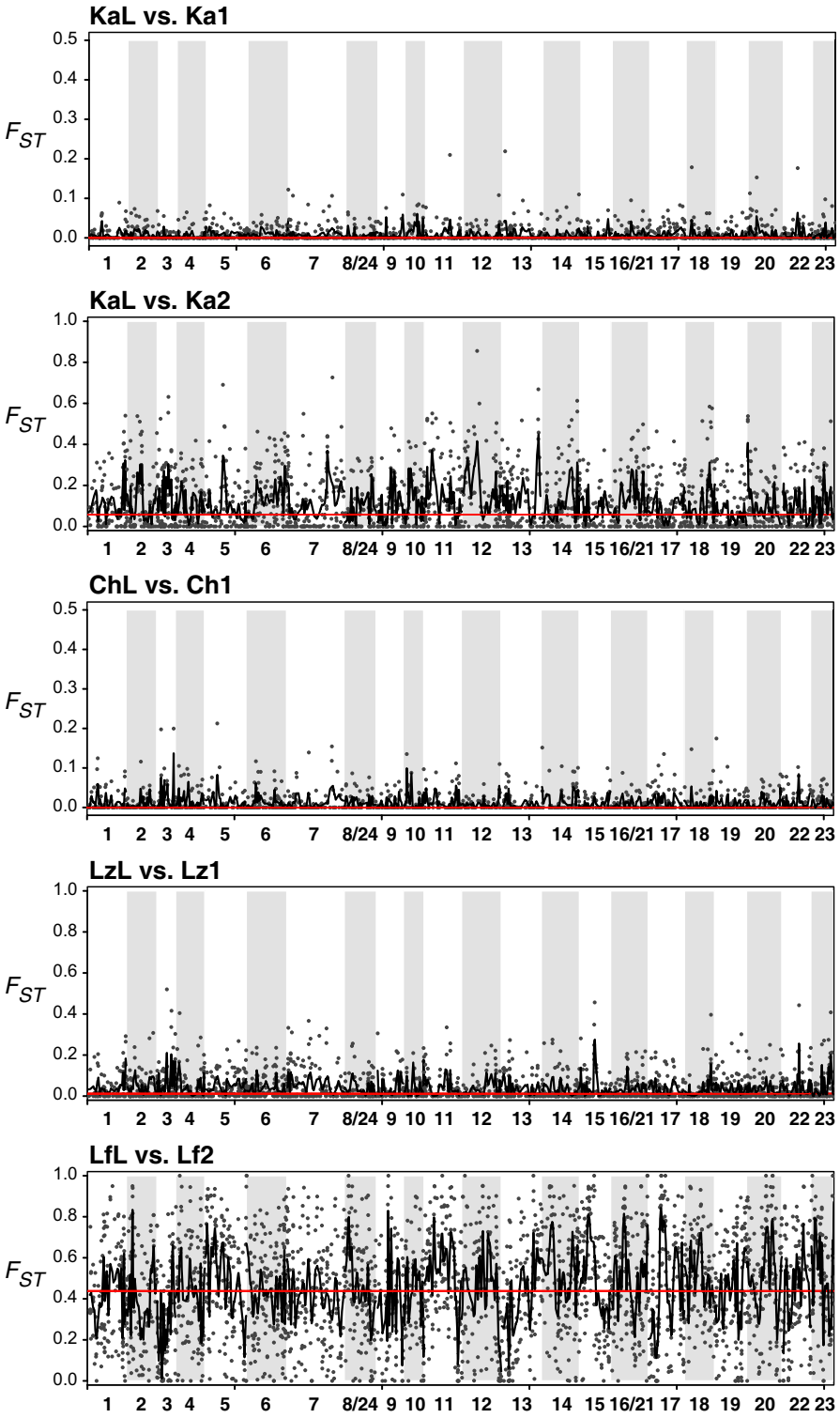

FIGURE 3 Genome-wide differentiation in five independent lake–stream population pairs. Dark dots show F_{ST} values for each marker on different linkage groups (x-axis); linkage groups are separated by white and grey background shading. The black line represents a sliding window analysis to visualize broad-scale divergence patterns; the red horizontal line represents baseline divergence defined as genome-wide median F_{ST}. Note that we used different FST-scales for the weakly divergent KaL vs. Ka1 and ChL vs. Ch1 comparisons

Lake Tanganyika in East Africa. In a previous study (Theis et al., 2014), we have reported consistent morphological and ecological differences between these and other lake and stream populations of *A. burtoni*, and we could show—using common garden experiments—that these differences are at least partly genetically controlled. However, with the mitochondrial (mt) DNA and microsatellite markers analysed by that time, we could not adequately resolve the evolutionary relationships between the populations under investigation, nor could we examine the genomic underpinnings of adaptive divergence in *A. burtoni*. With our new analyses based on thousands of SNP markers sampled across the genome, we obtain a much better understanding of the complex evolutionary history of *A. burtoni* in southern Lake Tanganyika. We first discuss new insights regarding the phylogeography of this species and re-evaluate the evolutionary

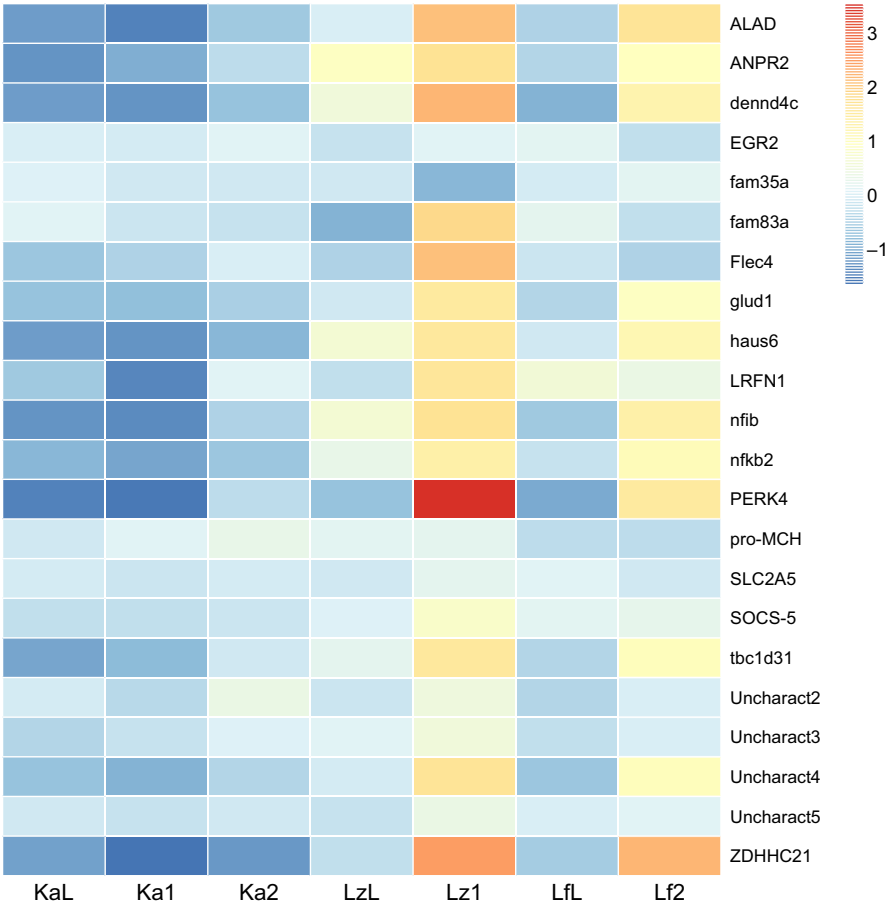

FIGURE 4 Heatmap showing gene expression means of 22 candidate genes in seven lake and stream populations, normalized by the overall mean of the gene (−ΔΔCt). Stream populations from the Lunzua (Lz1) and Lufubu (Lf2) systems show elevated gene expression (indicated by yellow, orange and red coloration) compared to the respective lake populations, whereas lake and stream populations from the Kalambo system show less pronounced differences. Note that we excluded the Ka3 population due to its phylogenetic position and no data was available for the Chitili system

independence of different lake–stream population clusters. We then focus on genome divergence between lake and stream populations and candidate outlier loci for lake–stream divergence.

4.1 | Phylogeography and population structure of *A. burtoni* in southern Lake Tanganyika

Our initial examination of *A. burtoni* populations in the south of Lake Tanganyika (Theis et al., 2014) revealed an unexpectedly high degree of differentiation in the mtDNA control region, resembling the genetic diversity observed in the same marker in the entire haplochromine cichlid assemblage of Lake Victoria (Verheyen, Salzburger, Snoeks, & Meyer, 2003). More precisely, we identified three mtDNA haplotype clusters in *A. burtoni*, grouping together (i) specimens sampled at the shoreline to the west of Crocodile Island near Mpulungu (populations 15–17 and 19–20 in Theis et al., 2014); (ii) specimens collected to the east of Crocodile Island as far as Ninde in Tanzania (populations 1–14); and (iii) specimens collected at an upstream site in the Lufubu River (Lf2; population 18). Our new phylogenetic hypothesis based on the maximum-likelihood method and on a concatenation of 9,529 SNPs (Figure 1b) confirms a deep divergence in *A. burtoni*, in this case, however, between the upstream Lufubu population (Lf2) and all remaining populations including the fish sampled at the estuary of the Lufubu River (LfL). Despite the deep split between the LfL and Lf2 specimens in the phylogeny (Figure 1b) and the high median F_{ST} between LfL and Lf2 (Table 2), there is nonetheless evidence for a shared evolutionary history between these two populations, as indicated by nonzero migration rate estimates (Table 3) and, more importantly, substantial co-ancestry sharing (Figure 2). Interestingly, the LfL individuals share similar levels of co-ancestry with individuals from their own population as with specimens collected at Lf2; yet, unlike the Lf2 fish, the LfL individuals also share co-ancestry with the remaining lake and stream populations (Figure 2). A possible explanation for this pattern is that southern Lake Tanganyika was originally colonized by *A. burtoni* individuals from upstream Lufubu River stocks, whereas subsequent gene flow primarily occurred between adjacent populations. That migration rates were generally found to be higher from either Lufubu population into the eastern populations as compared to the opposite direction (Table S4), lends additional support to this scenario. Interestingly, this also fits the colonization scenario proposed for *Tylochromis polylepis*, another comparatively recent addition to the fauna of Lake Tanganyika (Koch et al., 2007).

Another interesting finding is the apparent polyphyly of the Kalambo River populations. While other populations sampled from a certain river system cluster together (Figure 1b) and/or show the highest degree of co-ancestry sharing (Figure 2), the most upstream Kalambo River population (Ka3) is placed as sister group to the Lunzua riverine population (Lz1) in our phylogeny (Figure 1b), technically rendering the Lunzua populations as paraphyletic. Ka3 is separated

from the more downstream Kalambo River populations (Ka2 and Ka1) and the fish collected from within Lake Tanganyika yet near the Kalambo River delta (KaL) by the Kalambo Falls, which obviously act—with a drop of more than 220 m—as a strong barrier to gene flow in both directions (Figures 1 and 2; Table 2). The *A. burtoni* stocks from upstream and downstream the Kalambo Falls thus derive from different founder populations. While the downstream fish (populations Ka1 and Ka2) appear to be derived from from lake fish (Figure 1b), the close relatedness between the upstream fish (Ka3) and the Lunzua River populations (LzL and Lz1) suggests past migration between the upper Kalambo and the Lunzua River via a past river connection somewhere in the rivers' hinterlands. Temporary historical connections between the headwaters of the Kalambo and Lunzua River system are not unlikely, given that the area is geologically rather active and that past river captures mediated by tectonic movements have been suggested as agents for faunal exchange between otherwise isolated basins, including the Kalambo River (e.g., Delvaux, Kervyn, Vittori, Kajara, & Kilembe, 1998; Cohen et al., 2013; and Meyer et al., 2015 for an example involving another haplochromine cichlid fish discovered at the Kalambo estuary, *Haplochromis* sp. "Chipwa"). That this pattern is due to human translocation cannot be ruled out but is unlikely, given the monophyly of the Ka3 individuals and the long branch leading to this clade. In any case, as Ka3 and the downstream Kalambo populations (KaL, Ka1, Ka2) of *A. burtoni* have different evolutionary origins, the comparison KaL-Ka3 should no longer be considered an independent replicate of a lake–stream population pair, and was consequently excluded from further analyses. It remains a matter of future research to disentangle the precise colonization history of *A. burtoni* in southern Lake Tanganyika.

When comparing the five remaining lake–stream population pairs (Chitili: ChL vs. Ch1; Kalambo: KaL vs. Ka1 and KaL vs. Ka2; Lufubu: LfL vs. Lf2; Lunzua: LzL vs. Lz1), a strong geographic component becomes evident. Geographically proximate lake–stream population pairs (ChL/Ch1 and KaL/Ka1, sampled within a distance of 340 m and 2 km, respectively; Figure 1b) showed high levels of gene flow as indicated by median F_{ST} values of 0 (Table 2), the highest estimated migration rates (Table 3), they were not reciprocally monophyletic (Figure 1b), and the individuals essentially showed indistinguishable levels of co-ancestry sharing within and between the two compared populations (Figure 2). The lake–stream population pairs KaL/Ka2 and LzL/Lz1 (sampled at a distance of ~5 and ~7 km, respectively) showed intermediate medium F_{ST} values (Table 1), the riverine populations (Ka2 and Lz1) were monophyletic (Figure 1c) and clearly separated according to the FINERADSTRUCTURE analysis (Figure 2), whereas the Lufubu lake–stream population pair (LfL/Lf2, distance between the sampling locations: ~38 km) was the most genetically distinct according to these parameters (see also above). This is also reflected in the magnitude of genome-wide divergence between lake and stream populations within river systems (Figure 3), with essentially no baseline divergence between ChL and Ch1 and between KaL and Ka1, moderate levels of divergence between KaL and Ka2 and between LzL and Lz2, and substantial divergence (and fixation of ~1% of the loci) within the Lufubu system. Together with the strong signal of IBD observed across all comparisons, this suggests that baseline genomic divergence increases with geographic distance and decreases with the level of gene flow between lake and stream populations of *A. burtoni*.

In our previous study on lake and stream *A. burtoni* (Theis et al., 2014), we tested for a positive correlation between the extent of adaptive divergence in phenotypic traits and of neutral genetic differentiation, thereby controlling for geographic distance between populations. Such an association has been termed "isolation by adaptation" (IBA) and is interpreted as evidence for (ecological) diversification (Nosil et al., 2009; Thibert-Plante & Hendry, 2009). None of the morphological traits measured in Theis et al., 2014 (body shape, lower pharyngeal jaw shape and gill raker length) correlated positively with F_{ST} values based on microsatellite data when controlling for geographic distance. A reanalysis using median genome-wide F_{ST} values from the current study, however, indicated that higher levels of baseline divergence were associated with increased body shape differentiation. Yet, no IBA was detected with regard to the trophic traits gill raker length and lower pharyngeal jaw morphology. That, overall, geographically adjacent (and genetically more admixed) population pairs show less trait divergence compared to geographically separated (and genetically more differentiated) population pairs, provides further evidence to the view that the different lake–stream population pairs in *A. burtoni* rest at different stages along the "speciation continuum" (see Theis et al., 2014).

4.2 | Demographic inferences and patterns of genome divergence in *A. burtoni*

Time since divergence from a common ancestral population is, in general, acknowledged as an important factor influencing patterns of genome differentiation, resulting in highly heterogeneous divergence patterns between populations that split at different times and/or feature different levels of between-population gene flow (e.g., Feder et al., 2013; Nosil et al., 2009). Accordingly, selection on few loci would be the major determinant of genomic differentiation in "young" populations diverging along an environmental gradient (in this case, gene flow between populations would be prevalent). In "older" population pairs, on the other hand, which experience low levels of gene flow, selection, drift and new mutations would jointly shape the genomic landscape of divergence, thereby affecting many regions in the genome (see Figure 1 in Feder, Egan, & Nosil, 2012).

Our estimates of demographic parameters on the basis of joint site frequency spectra combined with the use of FASTSIMCOAL uncovered substantial differences between the replicate lake–stream population pairs in *A. burtoni* in the south of Lake Tanganyika (Table 3). It is important to note here that the estimated demographic parameters should be taken with caution. The influence of past climatic changes on the evolutionary history of species inhabiting the shallow, littoral zone of Lake Tanganyika has been well documented (e.g., Koblmüller et al., 2011; Sturmbauer, Baric, Salzburger, Rüber, &

Verheyen, 2001; Sturmbauer, Koblmüller, Sefc, & Duftner, 2005). Similarly, riverine and lacustrine *A. burtoni* populations were likely affected by past climatic and tectonic changes, as, for example, evidenced by the close relationship between populations from the eastern and western shore of Lake Tanganyika based on mtDNA in Theis et al. (2014). Thus, our model assumptions (i.e., the split of two populations from one ancestral population, divergence with gene flow) might be violated in some of the population comparisons, which could lead to deviating parameter estimates. We are also aware that a divergence with geneflow model for geographically close, genetically undifferentiated lake–stream population pairs (KaL/Ka1 and ChL/Ch1) is inappropriate. However, due to the fact that more complex models would involve defining a priori evolutionary scenarios not backed by empirical data, and for the sake of completion, we also included the latter two comparisons.

According to our estimates, the genome-wide most divergent and geographically most distant lake–stream pair, LfL/Lf2, also revealed the oldest split from an ancestral population, followed by LzL/Lz1 and KaL/Ka2 (when ignoring estimates for the undifferentiated population pairs ChL/Ch1 and KaL/Ka1). Another observation from the demographic analyses is that effective population sizes were consistently smaller in stream populations compared to lake populations (except for KaL and Ka1; Table 3), suggesting that riverine populations underwent bottlenecks and/or experienced more severe selection regimes. In line with this, pairwise population-specific allele frequency spectra revealed that genetic diversity is higher in lake compared to stream populations, especially in the more divergent systems (Fig. S3).

The differences in demographic parameters are also reflected in the patterns of genome divergence, which differed substantially between the replicate lake–stream population pairs (Figure 3). The most divergent and geographically most separated population pair (LfL/Lf2) shows substantial divergence across the genome, with multiple regions of high divergence and fixation of about 1% of the SNPs, whereas the least divergent and geographically most proximate lake-stream population pairs (ChL/Ch1, KaL/Ka1) show very little divergence and fewer "outlier" loci with F_{ST} values much smaller than 1. A similar pattern has previously been observed in other replicate lake–stream population pairs in fish (threespine sticklebacks: Roesti et al., 2012) and, hence, appears to be a common feature under such environmental settings, that is, along a lake–stream environmental gradient in fish. In the case of sticklebacks, the differences in the patterns of genome divergence have been attributed to differences in selection regimes among replicates, the involvement of different QTLs in responses to similar selection regimes or fundamental limitations of genome scans; that is, diverged regions in replicate genome scans might not necessarily be related to selection mediated by ecological differences (e.g., Deagle et al., 2012; Roesti et al., 2012). The streams inhabited by our study populations, ranging from a small creek (Chitili) to the largest tributary to Lake Tanganyika in the south (Lufubu), differ substantially in a variety of parameters (e.g., water current, ambient light, pH, temperature, water chemistry; Theis et al., 2014, 2017; B. Egger, A. Theis, & W. Salzburger, unpublished data), suggesting that selection plays a major role in shaping the genomic landscape of divergence.

4.3 | Outlier analyses

An advantage of using high-density genome scans (such as RADseq) at the population level is the potential to identify and characterize loci (and regions) in the genome showing high divergence (e.g., F_{ST}) between populations, so-called outlier loci. This strategy becomes even more powerful, when several replicates are considered, as adaptation loci consistently involved in divergence can be detected with higher confidence (Berner & Salzburger, 2015). We have thus filtered for loci that were above the baseline divergence in contrasting habitats (lake–stream comparisons), but below or equal to the baseline divergence within habitats (lake–lake comparisons).

Interestingly, our outlier analysis did not reveal a single consistent outlier locus across all investigated lake–stream population pairs. However, when the most divergent pair (LfL/Lf2) was excluded, we retrieved eight outlier loci consistently present in the pairwise comparisons within the Chitili, Kalambo and Lunzua rivers located in eight distinct nonoverlapping genome regions. The low number of consistent outliers is somewhat surprising, given the differentiation across lake and stream populations in several morphological traits (body shape, gill raker length and pharyngeal jaws), which represent quantitative traits with an underlying polygenic basis (e.g., Albertson & Kocher, 2006; Berner, Moser, Roesti, Buescher, & Salzburger, 2014; Franchini et al., 2014; Miller et al., 2014). On the other hand, as within each population pair many more outliers have been found, this suggests that system-specific adaptations have occurred as well, which remain to be characterized.

The human orthologs of five of the 34 candidates located in the outlier regions have functions in the immune system, which motivated us to focus our gene expression analysis on gill tissue. Furthermore, five other genes in the candidate regions act in the nervous system. However, none of these are clustered in a specific genomic region. Expression of half of the successfully amplified outlier candidate genes differed at least in one of the lake–stream pairs, with a consistent upregulation in the stream population as compared to the lake population. The differentially expressed genes function in immune defense (*Flec4*), brain development and learning (*glud 1*, *pro-MCH*, *zdhhc1*), cell division and proliferation (*fam83a*, *haus6*, *dennd4c*) or their function is unknown (uncharact2, uncharact3, uncharact5, *PERK4*). The Lufubu and Lunzua stream populations (Lf2 & Lz1) show very similar expression patterns (Figure 4), and within the two systems, more genes are differentially expressed (five each) as compared to the Kalambo system (three genes between KaL and Ka2). Among them are *PERK4*, a receptor-like protein kinase, which is not functionally characterized in vertebrates, and *Glud1*, which is one of the genes implicated with the nervous system—a mitochondrial glutamate dehydrogenase that may be involved in learning and memory reactions. Interestingly, genes differentially expressed within a lake–stream system were located on only one or two scaffolds each (Table S3), implying that requirements for successful adaptation may

be distinct depending on the environmental condition of the specific system.

We would like to note here that our outlier detection approach remains limited due to (i) the fact that only one tissue was analysed, (ii) a rather low marker density (compared to whole genome sequencing), (iii) a rather large window size (20 kb) used for blast searches and (iv) a possible failure to detect genes in outlier regions due to the somewhat incomplete genome annotation of *A. burtoni* (for a recent discussion on the utility of RADseq for genome scans of adaptation, see Catchen et al., 2017; Lowry et al., 2017; McKinney, Larson, Seeb, & Seeb, 2017). Still, our results suggest that selection for diversification may be particularly strong on genes involved in the adaptation to new habitats such as immune defence and cell proliferation genes, brain development and learning.

5 | CONCLUSIONS

In summary, our phylogenetic reconstruction based on genome-wide molecular data largely resolved the evolutionary relationships among lake and stream *A. burtoni* populations in southern Lake Tanganyika. We detected a deep divergence between the upstream Lufubu (Lf2) population and all remaining populations, and distinct evolutionary origins for populations from the Kalambo River, implicating that KaL-Ka3 does not constitute an independent replicate of a lake–stream population pair. Further, there was a strong pattern of IBD, with baseline genomic divergence increasing with geographic distance and decreasing with the level of gene flow between lake and stream population clusters. Genome divergence between lake and stream populations was generally heterogeneous and inconsistent among replicates, which may be explained by differences in divergence times, levels of gene flow and local selection regimes. In line with the latter, we did not find a single outlier when taking all independent replicates into account. However, when the divergent Lufubu system was excluded, we detected eight consistent outlier loci among the remaining lake and stream population comparisons. The candidate genes identified in the outlier regions have inferred functions in immune and neuronal systems, and interestingly, half of the successfully amplified outlier candidate genes were differently expressed within at least one of the lake–stream replicates, with a consistent upregulation in stream populations as compared to lake populations. Overall, however, while the RADseq data provided valuable novel insights with respect to phylogenetic and demographic patterns, the relatively low genomic resolution achieved by this method limits interpretations related to the molecular basis of adaptive divergence in *A. burtoni*.

ACKNOWLEDGEMENTS

We thank Daniel Berner for providing the RADseq analyses pipeline and for help with demographic analyses and Milan Malinsky for sharing the code of and help with FINERADSTRUCTURE. We further thank our helpers in the field, Adrian Indermaur, Anya Theis, Fabrizia Ronco, Isabel Tanger (née Keller) Heinz H. Büscher, Craig Zytkow (Conservation Lake Tanganyika) and Gilbert Tembo and his crew for their logistic support in Zambia; the Lake Tanganyika Research Unit, Department of Fisheries, Republic of Zambia, for research permits; and the subject editor and the anonymous referees for valuable comments. This study was supported by grants from the Volkswagen Stiftung and the German Research Foundation (DFG) to O.R.; the European Research Council (ERC, Starting Grant "INTERGENADAPT" and Consolidator Grant "CICHLID~X"), the University of Basel and the Swiss National Science Foundation (SNF) (Grants 3100A0 138224 and 3100A0 156405) to W.S.

DATA ACCESSIBILITY

Raw RAD sequencing reads associated with this study are available from the Sequence Read Archive (SRA) at NCBI under the Accession nos SRX2967972–SRX2968211 (SRA Study Number: SRP110734).

AUTHOR CONTRIBUTIONS

B.E. and W.S. designed the research; B.E., M.R. and O.R. performed the wet laboratory work; B.E., M.R., O.R. and A.B. analysed the data; B.E. and W.S. wrote the manuscript with input from all co-authors.

REFERENCES

Albertson, R. C., & Kocher, T. D. (2006). Genetic and developmental basis of cichlid trophic diversity. *Heredity*, *97*, 211–221.

Beemelmanns, A., & Roth, O. (2016). Biparental immune priming in the pipefish *Syngnathus typhle*. *Zoology*, *119*, 262–272.

Berg, P. R., Star, B., Pampoulie, C., Sodeland, M., Barth, J. M. I., Knutsen, H., … Jentoft, S. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scientific Reports*, *6*, 23246.

Berner, D., Moser, D., Roesti, M., Buescher, H., & Salzburger, W. (2014). Genetic architecture of skeletal evolution in European lake and stream stickleback. *Evolution*, *68*, 1792–1805.

Berner, D., & Salzburger, W. (2015). The genomics of organismal diversification illuminated by adaptive radiations. *Trends in Genetics*, *31*, 491–499.

Bookout, A. L., & Mangelsdorf, D. J. (2003). Quantitative real-time PCR protocol for analysis of nuclear receptor signaling pathways. *Nuclear Receptor Signaling*, *1*, e012.

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.

Carneiro, M., Blanco-Aguiar, J. A., Villafuerte, R., Ferrand, N., & Nachman, M. W. (2010). Speciation in the European rabbit (*Oryctolagus cuniculus*): Islands of differentiation on the X chromosome and autosomes. *Evolution*, *64*, 3443–3460.

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, *17*, 362–365.

Cohen, A. S., Van Bocxlaer, B., Todd, J. A., McGlue, M., Michel, E., Nkotagu, H. H., … Delvaux, D. (2013). Quaternary ostracodes and molluscs from the Rukwa Basin (Tanzania) and their evolutionary and

palaeobiogeographic implications. *Palaeogeography Palaeoclimatology Palaeoecology*, *392*, 97.

Colosimo, P. F., Hosemann, K. E., Balabhadra, S., Villarreal, G., Dickson, M., Grimwood, J., … Kingsley, D. M. (2005). Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, *307*, 1928–1933.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: J. Murray.

Deagle, B. E., Jones, F. C., Chan, Y. F., Absher, D. M., Kingsley, D. M., & Reimchen, T. E. (2012). Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Proceedings of the Royal Society B: Biological Sciences*, *279*, 1277–1286.

Delvaux, D., Kervyn, R., Vittori, E., Kajara, R. S. A., & Kilembe, E. (1998). Late Quaternary tectonic activity and lake level fluctuation in the Rukwa rift basin, East Africa. *Journal of African Earth Sciences*, *26*, 397–421.

Dobzhansky, T. (1937). *Genetics and the origin of species*. New York: Columbia University Press.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., … Wolf, J. B. W. (2012). The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, *491*, 756–760.

Emelianov, I., Marec, F., & Mallet, J. (2004). Genomic evidence for divergence with gene flow in host races of the larch budmoth. *Proceedings of the Royal Society B: Biological Sciences*, *271*, 97–105.

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

Faria, R., Renaut, S., Galindo, J., Pinho, C., Melo Ferreira, J., Melo, M., … Bultin, R. (2014). Advances in ecological speciation: An integrative approach. *Molecular Ecology*, *23*, 513–521.

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, *28*, 342–350.

Feder, J. L., Flaxman, S. M., Egan, S. P., & Nosil, P. (2013). Hybridization and the buildup of genomic divergence during speciation. *Journal of Evolutionary Biology*, *26*, 261–266.

Franchini, P., Fruciano, C., Spreitzer, M. L., Jones, J. C., Elmer, K. R., Henning, F., & Meyer, A. (2014). Genomic architecture of ecologically divergent body shape in a pair of sympatric Crater Lake cichlid fishes. *Molecular Ecology*, *23*, 1828–1845.

Gagnaire, P.-A., Pavey, S. A., Normandeau, E., & Bernatchez, L. (2013). The genomic architecture of reproductive isolation across the speciation continuum in Lake Whitefish species pairs assessed by RAD-sequencing. *Evolution*, *67*, 2483–2497.

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

Gaujoux, R., & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, *11*, 367.

Goslee, S. C., & Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, *22*, 1–19.

Grabherr, M. G., Russel, P., Meyer, M., Mauceli, E., Alföldi, J., Di Palma, F., & Lindblad-Toh, K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*, *26*, 1145–1151.

Haasl, R. J., & Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, *25*, 5–23.

Harr, B. (2006). Genomic islands of differentiation between house mouse subspecies. *Genome Research*, *16*, 730–737.

Hellemans, J., Mortier, G., de Paepe, A., Speleman, F., & Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biology*, *8*(2), R19.

Jones, F. C., Grabherr, M. G., Chan, F. Y., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*, 55–61.

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

Koch, M., Koblmüller, S., Sefc, K. M., Duftner, N., Katongo, C., & Sturmbauer, C. (2007). Evolutionary history of the endemic Lake Tanganyika cichlid fish *Tylochromis polylepis*: A recent intruder to a mature adaptive radiation. *Journal of Zoological Systematics and Evolutionary Research*, *45*, 64–71.

Lamichhaney, S., Berglund, J., Almen, M. S., Mawbool, K., Grabherr, M., Martinez-Barrio, A., … Anderson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*, 371–375.

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, *8*(1), e1002453.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Lohse, K., Clarke, M., Ritchie, M. G., & Etges, W. J. (2015). Genome-wide tests for introgression between cactophilic Drosophila implicate a role of inversions during speciation. *Evolution*, *69*(5), 1178–1190.

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: An evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, *17*, 142–152.

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.

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

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., … Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, *23*, 1817–1828.

Mayr, E. (1942). *Systematics and the origin of species*. New York: Columbia University Press.

Mazzuchelli, J., Kocher, T., Yang, F., & Martins, C. (2012). Integrating cytogenetics and genomics in comparative evolutionary studies of cichlid fish. *BMC Genomics*, *13*(1), 463.

McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: Comment on Breaking RAD by Lowry et al. (2017). *Molecular Ecology Resources*, *17*, 356–361.

Meyer, B. S., Indermaur, A., Ehrensperger, X., Egger, B., Banyankimbona, G., Snoeks, J., & Salzburger, W. (2015). Back to Tanganyika: A case of recent trans-species-flock dispersal in East African haplochromine cichlid fish. *Royal Society Open Science*, *2*, 140498.

Miller, C. T., Glazer, A. M., Summers, B. R., Blackman, B. K., Norman, A. R., Shapiro, M. D., … Kingsley, D. M. (2014). Modular skeletal evolution in sticklebacks is controlled by additive and clustered quantitative trait loci. *Genetics*, *197*, 405–420.

Nadeau, N. J., Whibley, A., Jones, R. T., Davey, J. W., Dasmahapatra, K. K., Baxter, S. W., … Jiggins, C. D. (2012). Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale

targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 343–353.

Nei, M., & Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics*, *97*, 145–163.

Nevado, B., Ramos-Onsins, S. E., & Perez-Enciso, M. (2014). Resequencing studies of nonmodel organisms using closely related reference genomes: Optimal experimental designs and bioinformatics approaches for population genomics. *Molecular Ecology*, *23*, 1764–1779.

Nosil, P. (2012). *Ecological speciation*. Oxford: Oxford University Press.

Nosil, P., Feder, J., Flaxman, S. M., & Gompert, Z. (2017). Tipping points in the dynamics of speciation. *Nature Ecology and Evolution*, *1*, 0001.

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

R Core Team (2012). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org/

Renaut, S., Grassa, C. J., Yeaman, S., Moyers, B. T., Lai, Z., Kane, N. C., … Rieseberg, L. H. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, *4*, 1827.

Roesti, M., Hendry, A. P., Salzburger, W., & Berner, D. (2012). Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology*, *21*, 2852–2862.

Roesti, M., Kueng, B., Moser, M., & Berner, B. (2015). The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications*, *6*, 8767. https://doi.org/10.1038/ncomms9767

Rundle, H. D., & Nosil, P. (2005). Ecological speciation. *Ecology Letters*, *8*, 336–352.

Schliep, K. (2011). phangorn: Phylogenetic analysis in r. *Bioinformatics*, *27*, 592–593.

Schluter, D. (2000). Ecological character displacement in adaptive radiation. *The American Naturalist*, *156*, S4–S16.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. *Science*, *323*, 737–741.

Schluter, D., & McPhail, J. D. (1992). Ecological character displacement and speciation in sticklebacks. *The American Naturalist*, *140*, 85–108.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., … Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, *15*, 176–192.

Ségurel, L., Wyman, M. J., & Przeworski, M. (2014). Determinants of mutation rate variation in the human germline. *Annual Review of Genomics and Human Genetics*, *15*, 47–70.

Sturmbauer, C., Baric, S., Salzburger, W., Rüber, L., & Verheyen, E. (2001). Lake level fluctuations synchronize genetic divergences of cichlid fishes in African lakes. *Molecular Biology and Evolution*, *18*, 144–154.

Sturmbauer, C., Koblmüller, S., Sefc, K. M., & Duftner, N. (2005). Phylogeographic history of the genus Tropheus, a lineage of rock-dwelling cichlid fishes endemic to Lake Tanganyika. *Hydrobiologia*, *542*, 335–366.

Theis, A., Ronco, F., Indermaur, A., Salzburger, W., & Egger, B. (2014). Adaptive divergence between lake and stream populations of an East African cichlid fish. *Molecular Ecology*, *23*, 5304–5322.

Theis, A., Roth, O., Cortesi, F., Ronco, F., Salzburger, W., & Egger, B. (2017). Variation of anal fin egg-spots along an environmental gradient in a haplochromine cichlid fish. *Evolution*, *71*, 766–777

Thibert-Plante, X., & Hendry, A. P. (2009). Five questions on ecological speciation addressed with individual-based simulations. *Journal of Evolutionary Biology*, *22*(1), 109–123.

Verheyen, E., Salzburger, W., Snoeks, J., & Meyer, A. (2003). Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. *Science*, *300*, 325–329.

Wolf, J. B., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, *18*, 87–100.

Wolf, J. B. W., Lindell, J., & Backstrom, N. (2010). Speciation genetics: Current status and evolving approaches. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*, 1717–1733.

Wu, C. I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*, 851–865.

Yeaman, S., & Whitlock, M. C. (2011). The genetic architecture of adaptation under migration-selection balance. *Evolution*, *65*, 1897–1911.

SUPPORTING INFORMATION

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

How to cite this article: Egger B, Roesti M, Böhne A, Roth O, Salzburger W. Demography and genome divergence of lake and stream populations of an East African cichlid fish. *Mol Ecol*. 2017;26:5016–5030. https://doi.org/10.1111/mec.14248