Title:

Admixture and environmental fluctuations shape the evolutionary history of a predator radiation in East Africa's Lake Tanganyika

Authors and Affiliations:

Jessica A. Rick[1,*], Julian Junker[2,3], Alexander L. Lewanski[4], Brittany Swope[5], Michael M. McGlue[6], Emmanuel A. Sweke[7], Ismael A. Kimirei[8], Ole Seehausen[3,4], and Catherine E. Wagner[4,9*]

[1]School of Natural Resources and the Environment, The University of Arizona, Tucson, AZ 85721, USA
[2]Center for Ecology, Evolution, and Biogeochemistry, EAWAG Swiss Federal Institute of Aquatic Science and Technology, CH-6047 Kastanienbaum, Switzerland
[3]Division of Aquatic Ecology & Evolution, Institute of Ecology & Evolution, University of Bern, CH-3012 Bern, Switzerland
[4]Department of Integrative Biology, Michigan State University, East Lansing, MI, 48824, USA
[5]Department of Botany, University of Wyoming, Laramie, WY 82071, USA
[6]Department of Earth and Environmental Sciences, University of Kentucky, Lexington, KY 40506, USA
[7]Tanzania Deep Sea Fishing Authority, Zanzibar, Tanzania
[8]Tanzanian Fisheries Research Institute, Kunduchi, 14122 Dar es Salaam, Tanzania
[9]Program in Ecology and Evolution, University of Wyoming, Laramie, WY 82071, USA

Correspondence:

* To whom correspondence should be addressed: JAR (jrick@arizona.edu) and CEW (catherine.wagner@uwyo.edu)

Document information:

Article type: Research Report
Compiled: January 13, 2025
Figures: 4
Tables: 0
Main Text References: 86
Word Count: 582 + 1991 + 1827 + 535

Abstract

Top predators have oversized impacts on food webs and ecosystem dynamics, and introducing a novel predator to a naive environment can have dramatic consequences for endemic biodiversity. Lake Tanganyika is unique among African lakes in the diversity of the pelagic top predators in the genus *Lates*, where four species are endemic to the lake. Using a combination of reduced-representation and whole genome resequencing data, and pairing these with phylogenetic and demographic modeling approaches, we find that *Lates* colonization of Lake Tanganyika was much more recent (∼1–2 Mya) than other major and diverse clades within the lake. Demographic modeling suggests that diversification among *Lates* species within the lake occurred during a time period of dramatic changes in lake levels driven by glacial-interglacial cycles, supporting a role of these fluctuations as a "species pump" for lacustrine taxa. We further find that these lake level fluctuations likely contributed to multiple bouts of admixture among *Lates* species during the mid- to late-Pleistocene (∼90–500 Kya). Together, our findings suggest a dynamic and environmentally linked evolutionary history of the *Lates* radiation with the potential for dramatic ecosystem consequences for the taxa already present in Lake Tanganyika prior to *Lates* colonization and diversification.

Significance Statement

When introduced to novel ecosystems, top predators can cause major alterations to biodiversity and food webs. Species introductions to novel habitats can also provide invading taxa with ecological opportunities that facilitate evolutionary diversification. Here, we find evidence that the radiation of endemic top predators in East Africa's Lake Tanganyika originated surprisingly recently, and that these species have experienced periods of hybridization with a widespread riverine relative throughout their history. These findings have major implications for the history of the lake and suggest that the introduction of Nile perch into Lake Victoria, which caused dramatic ecosystem and food web changes, may be a contemporary analog to the historical events in Lake Tanganyika.

2 of 16

Introduction

Predators have strong direct and indirect effects on food webs, as well as outsized influences on ecosystem dynamics (1, 2). When introduced to novel ecosystems, top predators can cause major alterations to biodiversity, trophic interactions, and community structure (3–5), both within and across ecosystems (6), and can induce evolutionary change in native species (7). While much research focuses on the anthropogenic introduction of predator species, predators can also colonize new habitats, often as a result of novel habitat formation or reorganization caused by changing environmental conditions.

Environmental changes can act as an important control on speciation and hybridization through facilitating cycles of isolation and contact among populations (8–10). Pleistocene glacial-interglacial cycles transformed high latitude landscapes, caused wide-ranging global climatic changes (11–13), and have been invoked to explain speciation patterns in mammals (e.g., 14, 15), birds (e.g., 16–18), arachnids (e.g., 19), plants (e.g., 10, 20), marine invertebrates (e.g., 21), and freshwater fishes (e.g., 22). Global glacial periods are also associated with megadrought conditions in Africa (Fig. S1 23–25), and these periods had dramatic consequences for freshwater ecosystems, many of which contracted or disappeared during drought periods (13, 26–28). In tropical Africa, both climatic and tectonic events since the early Miocene periodically altered barriers to dispersal and environmental suitability, which played large roles in shaping patterns of biodiversity in both aquatic and terrestrial taxa (29, 30), including early hominids (31, 32).

Ancient lakes are among the best records of past environmental change; these lakes not only record long histories of environmental variation and human activity in their sediments, but also harbor high levels of biodiversity and endemism. The African Great Lakes region is home to numerous species radiations and abundant biodiversity (29). Lake Tanganyika (LT)—the largest (by volume; 18,880km^3), deepest (1,470m), and oldest (∼9-12 My; 33) of the African Great Lakes—represents one of the most diverse aquatic ecosystems in the world (34) and is home to >1500 aquatic animal species, of which >600 are endemic (29, 35) and >200 belong to a radiation of endemic cichlid fishes. Evolutionary diversification in these speciose clades is often linked to divergences in niche space as ecological opportunity arises (36), and the fish radiations have diversified along many common ecological axes (29, 37).

While the majority of LT biodiversity is concentrated in the near-shore littoral zone, the open-water pelagic zone is home to several smaller endemic radiations of fishes, including a four-species radiation of predators in the genus *Lates*. The LT *Lates* clade is the most species-rich *Lates* intralacustrine radiation, and it is highly unusual for an adaptive radiation of freshwater fish in that all four species are piscivorous predators, with differentiation in habitat, life history, and size, but not trophic position (35). Freshwater *Lates* are native and widespread in African rivers and have a patchy distribution among large lakes in East Africa; however, all of the species except *L. niloticus* are endemic to single lakes. These predators have large effects on their ecosystems, most clearly demonstrated by the introduction of *L. niloticus* into Lake Victoria, which famously led to ecological collapse in the endemic pelagic cichlid community (38, 39).

Here, we use genome-wide analyses to investigate the evolutionary history of LT's *Lates* species in relation to environmental change during the lake's history. We find genomic evidence that suggests glacial-interglacial cycles shaped the tempo and extent of this radiation via periodic cycles of admixture between *L. niloticus* and LT's endemic *Lates* species, as well as periodic opportunities for allopatric diversification. Furthermore, we find evidence that *Lates* only recently arrived in Lake Tanganyika, with dramatic implications for both the rapidity of this radiation and for the ecosystem impacts of their arrival on the lake's already diverse fauna.

Results

We generated genotyping-by-sequencing (GBS) and whole genome resequencing (WGS) data for *Lates* specimens from the four LT-endemic species and three *L. niloticus* museum specimens. With WGS, maximum

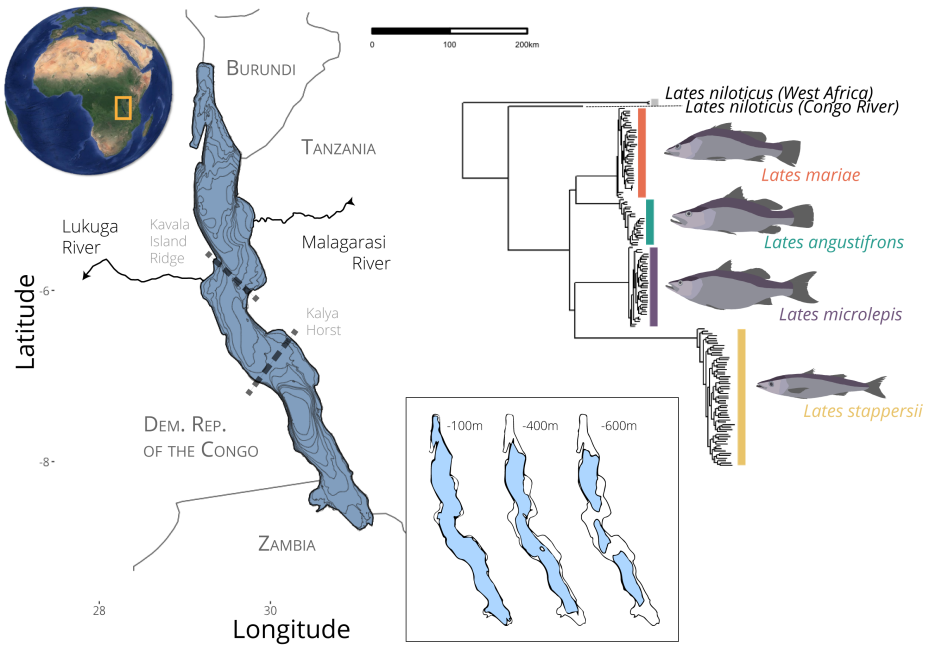


Fig. 1. Map of Lake Tanganyika, with inflow via the Malagarasi River and outflow via the Lukuga River indicated, as well as the location of the Kavala Island Ridge and Kalya Horst, separating the north and middle, and middle and south basins, respectively. The inset map shows the inferred lake extent at 100m, 400m, and 600m below the present surface of the lake based on present-day bathymetry, all levels that are likely to have occurred as a result of climatic cycling during the late Pleistocene (40). The phylogeny indicates relationships among species inferred using concatenated genotyping-by-sequencing data from (41), with the tree rooted using *L. calcarifer* as an outgroup.

likelihood concatenation (42) and gene tree (43) methods inferred a species tree with a monophyletic LT group and two sister species pairs within Lake Tanganyika (Fig. 2a). Monophyly of the LT species is consistent with the phylogenetic hypothesis recently proposed by (44), but intra-LT relationships differ. However, in both our results and those of (44), there are short internodes during LT diversification. Our analyses show high gene tree incongruence (only 13.2% and 12.4% of gene trees support the intra-LT relationships in the species tree topology; Fig. S9) that results in low support for splits at the base of the LT *Lates* radiation (Fig. S5a). With GBS, where we had sampling from multiple locations for *L. niloticus*, we additionally infer that *L. niloticus* is paraphyletic, with *L. niloticus* from the Congo basin sister to the LT clade, but West African individuals forming a distinct clade sister to the clade of Congo basin *L. niloticus* plus LT *Lates* (Fig. 1), mirroring the recent findings of (44) and agreeing with our own analyses of mitochondrial haplotypes (Fig. S10).

An analysis of topology weightings for 100-SNP windows across the genome showed that the most common topology for all 24 of the *L. calcarifer* chromosomes matched the species tree (Fig. 3). However, only 2.63% of the genome supported this relationship, while a total of 9.4% of the genome supported the four topologies classified as "alternate" topologies (i.e., those in which one sister relationship agreed with the species tree topology and the LT taxa were monophyletic), and a total of 72.6% of the genome supported the 90 possible topologies with *L. niloticus* nested within the LT clade (Fig. 3a).

Analysis of cross-coalescence among species in `MSMC2` (45) inferred an average divergence time estimate (i.e., time to 50% cross-coalescence rate) between *L. niloticus* and Lake Tanganyika *Lates* species of 1.61 Mya (range 1.41–2.10 Mya; Fig. 2c). Divergence between *L. angustifrons* and *L. mariae* dated to ∼0.49 Mya, between *L. microlepis* and *L. stappersii* dated to ∼1.21 Mya, *L. angustifrons*-*L. microlepis* to 0.53 Mya, *L. mariae*-*L. microlepis* to 0.55 Mya, *L. angustifrons*-*L. stappersii* to 1.21 Mya, and *L. mariae*-*L. stappersii* to 1.23 Mya (Fig. 2c). While the MSMC-inferred divergence times do not overlap with those from our demographic model for some speciation events (Fig. 2), this is likely because MSMC inferences do not account for gene flow among taxa, and we find much evidence for gene flow from other analyses.

4 of 16

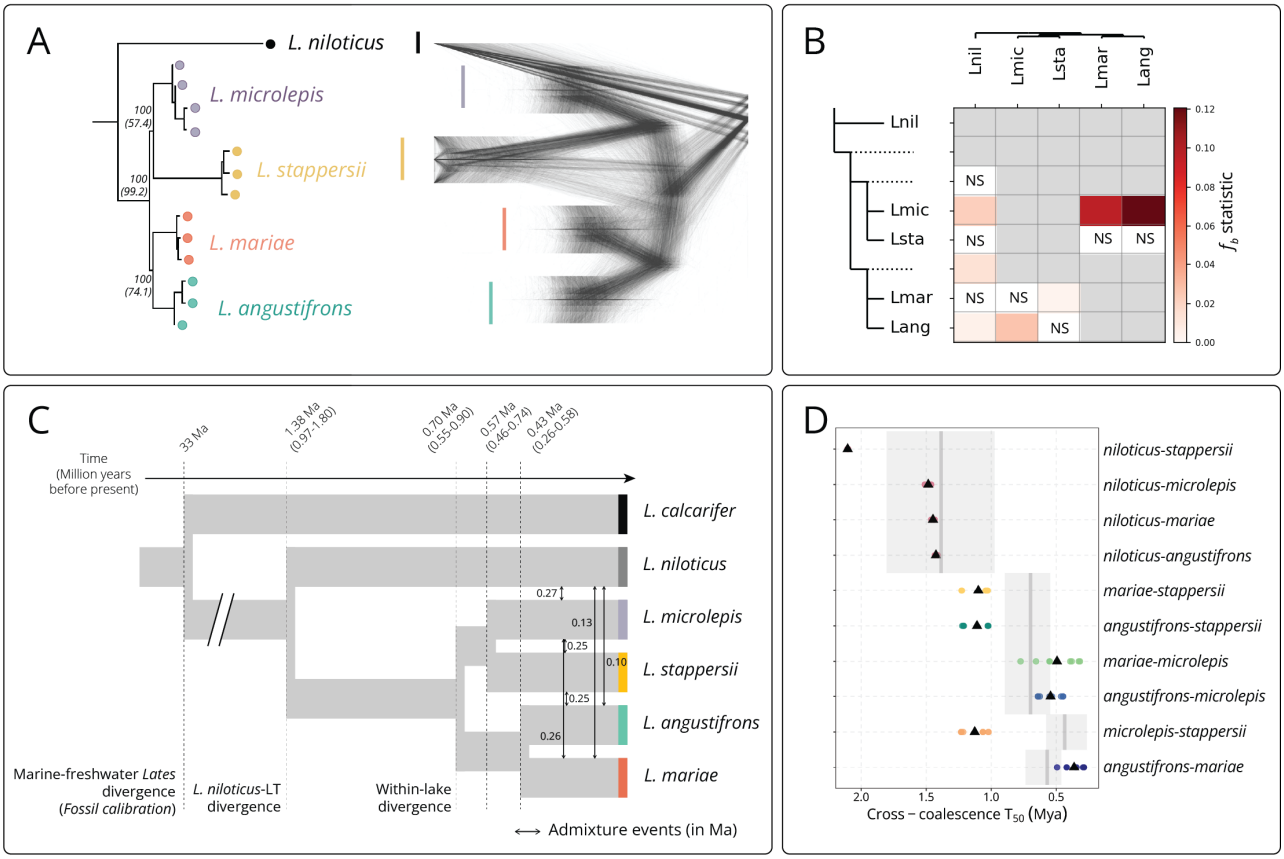


Fig. 2. Multiple lines of evidence support a monophyletic Lake Tanganyika *Lates* clade with a relatively recent origin ∼1-2 Mya and a history of admixture among species. (A) Species tree inference from whole genome sequences (left), rooted using *L. calcarifer* as the outgroup and including the *L. niloticus* sequence from the Congo River, demonstrates monophyly of the Lake Tanganyika clade (*L. microlepis, L. stappersii, L. mariae, L. angustifrons*). Confidence from bootstrap estimates (from `RAxML`) are shown for major splits, with quartet scores (from `ASTRAL`) shown in parentheses. Gene trees from 4000bp windows (right) show incongruence contributing to uncertainties in species tree inference. (B) The `Fbranch` statistics, conditioned on the species tree in (A), infer significant admixture events between many of the focal taxa. Darker red colors indicate a stronger signal of gene flow. Gray boxes indicate comparisons that could not be tested due being sister taxa, while white boxes indicate non-significant comparisons. (C) Visualization of the best model from the 32 tested `fastsimcoal2` demographic models, with the median bootstrap estimates for timing of divergence (dashed lines) and admixture (arrows) events labeled. Generation times have been converted to years using an average generation time of three years. For divergence events, we show estimates from the best model and 95% bootstrap intervals from nonparametric bootstrapping. (D) Reconstruction of divergence times using relative cross-coalescence rates for each pair of individuals in `MSMC2` provide an estimate of divergence times independent of estimates in (C). Divergence times are inferred as the point when the cross-coalescence rate reaches 0.5, T_{50}. Colored dots show the estimated T_{50} for each pair of individuals, while black triangles indicate the mean for each species pair. To provide a direct comparison to `fastsimcoal2` results, estimates (dark gray lines) and bootstrap confidence intervals (light gray boxes) are shown from `fastsimcoal2` results.

Evidence for ancient and recent admixture with *L. niloticus*. D-statistics for excess allele sharing among species (46, 47)—which can be indicative of admixture among non-sister species—indicate that all LT species are more closely related to one another than any of the four is to *L. niloticus* (Table S1), further supporting the finding that Congolese *L. niloticus* is sister to a monophyletic LT clade. The remaining significant D estimates indicate excess allele sharing between *L. niloticus* and each of *L. angustifrons*, *L. microlepis*, and *L. stappersii*, although with weaker signal in *L. stappersii* (Table S2). The derivative `Fbranch` analysis (48) of the D-statistics suggests that this excess allele sharing comes from an admixture event between *L. niloticus* and the common ancestor of *L. microlepis* and *L. stappersii*, with additional independent *L. niloticus*-*L. angustifrons* and *L.niloticus*-*L. microlepis* admixture events following intra-LT divergence into four species. Additionally, there is evidence for admixture between *L. microlepis* and each of *L. mariae* and *L. angustifrons*, and between *L. stappersii* and *L. angustifrons*. When *L. niloticus* is split into West Africa and Congo populations using the GBS data, we find some evidence for admixture between the Congo River lineage and LT species that is not present in the West African lineage (mean $D = 0.213$, F_4-ratio $= 0.139$; Fig. S3a), although alternate phylogenetic hypotheses indicate evidence for admixture

between the West African *L. niloticus* and the LT common ancestor that does not involve the Congo River lineage (Fig. S3b,c).

For each *L. niloticus*-LT *Lates* pair with an elevated D-statistic, we identified regions of the genome involved in admixture by calculating the f_d statistic (49) in 100-SNP sliding windows. The proportion of significant f_d regions was similar across chromosomes (mean f_d=0.51%, range=0.014–2.36%), although two chromosomes (ASB_LG14, ASB_LG8) had an elevated proportion of admixed windows compared to the other chromosomes (Fig. 3f). Several f_d peaks were shared across three-taxon comparisons (Fig. S4), while others were unique to specific species. As would be expected from theory (50) and recent empirical observations (51–53) of selection against admixed alleles, admixture signal was elevated in regions with high recombination (Fig. 3d; Pearson's $r(62437) = 0.230, p < 2.2x10^{-16}$), and both f_d and recombination rate increased near chromosome ends (Fig. 3c). In addition, an elevated f_d signal was associated with low topology weighting for the topology matching the `RAxML` tree (Fig. 3b; Pearson's $r(61933) = -0.039, p < 2.2x10^{-16}$), while high f_d was associated with windows where the topology weightings were high for topologies with *L. niloticus* nested within the LT clade. Admixture proportions calculated using f_d indicate that a higher proportion of the genome is involved in admixture between *L. angustifrons* and *L. niloticus* than between the other species and *L. niloticus* (Fig. 3e). Similar proportions of the sequenced genome are admixed between *L. niloticus* and *L. microlepis* and *L. stappersii*, while *L. mariae* had the smallest proportion of admixed regions. These proportions varied among chromosomes, with ASB_LG14 generally having the highest proportion of sequenced bases involved in admixture across comparisons (Fig. S5).

Divergence and admixture timing and magnitude. Site frequency spectrum-based coalescent modeling of demographic history indicated bouts of ancient admixture between *L. niloticus* and the LT *Lates* species. Model selection using both Akaike information criteria (AIC) and distributions of likelihood values (Fig. S7) indicated that our best model involved six independent bouts of admixture among our focal taxa (Fig. 2c). We estimated the split between *L. niloticus* and the LT *Lates* clade at 1.38 Mya (95% bootstrap CI: 0.97–1.80 Mya) and between the two LT sister clades at 697 Kya (95% bootstrap CI: 545–895 Kya). The two Tanganyika clades then split with divergence times dating to 570 (460–735) Kya and 435 (264-577) Kya (Table S5, Fig. 2c).

Our best model had an overlapping likelihood distribution with the next 10 best models (Fig. S7), and therefore we examined the distribution of estimated divergence dates across these top models. Across these ten models, the mean divergence time estimate between *L. niloticus* and the LT *Lates* clade was 1.3 Mya (95%: 1.2–1.7 Mya). The other intra-LT divergence times similarly overlapped with those of our best model, suggesting that our choice of best model does not significantly bias our divergence time estimates (Fig. S8).

Based on our best model, admixture events between *L. niloticus* and the LT *Lates* occurred in two distinct bouts. Around 0.27 Mya (95% CI: 0.02–0.56 Mya), admixture occurred between *L. niloticus* and *L. microlepis*. In this same period, we also infer admixture events between *L. microlepis* and each of *L. stappersii* (0.25 Mya; 95% CI: 0.06–0.45 Mya) and *L. mariae* (0.26 Mya; 95% CI: 0.06–0.46 Mya), as well as admixture between *L. stappersii* and *L. angustifrons* (0.25 Mya; 95% CI: 0.05–0.46 Mya; Table S5). Closer to the present, we infer that admixture occurred between *L. niloticus* and *L. mariae* around 0.13 Mya (95% CI: 0.01–0.43 Mya) and between *L. niloticus* and *L. angustifrons* around 0.10 Mya (95% CI: 0.007–0.36 Mya). While confidence intervals for all of these admixture events overlap, they also are quite wide. In these admixture bouts, population size-scaled migration rates were largely unidirectional from *L. niloticus* into the LT taxa, while intra-LT admixture events had more evenly balanced migration rates (Table S5). Using two-taxon models with the GBS data, we further investigated whether there is evidence for multiple bouts of admixture between taxa and find that the best models for *L. niloticus* and each of the LT taxa had either two or three admixture events (Fig. S13), while those between LT taxa inferred either one bout of post-divergence admixture (for *L. angustifrons*-*L. microlepis*, *L. mariae*-*L. microlepis*, and *L. mariae*-*L. stappersii* or two bouts of admixture (*L. microlepis*-*L. stappersii*; Fig. S15). The best model for *L. angustifrons*-*L. stappersii* was the only species pair for which three bouts of post-divergence admixture were inferred (Fig. S15), although these three admixture events were inferred to have occurred within a

6 of 16

narrow window of time (Fig. S16).

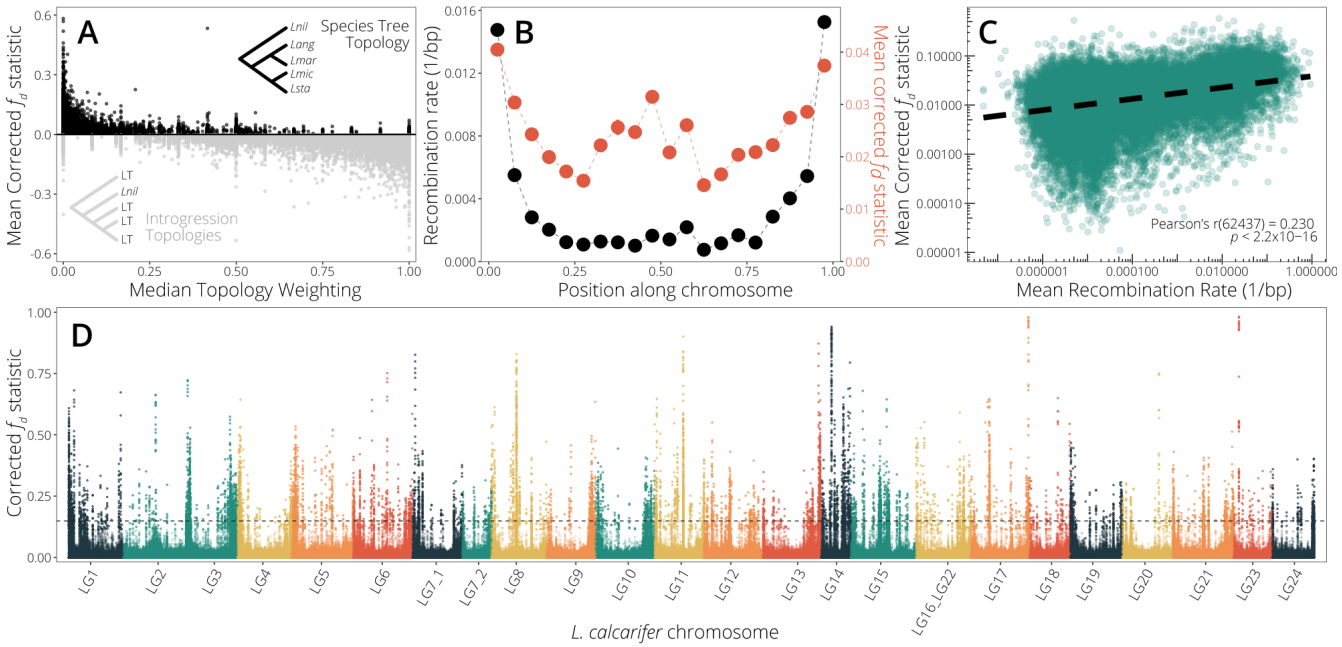


Fig. 3. The signal of admixture varies across the *Lates* genome, with admixed regions contributing to gene tree incongruence, and regions with higher recombination rates holding higher admixture proportions. (A) Windows with high f_d values were most often those with a low topology weighting for the species tree (top) and high topology weightings for the introgression topologies (bottom). (B) In both recombination rates (black, lefthand axis) and f_d (red, righthand axis), mean estimated values generally increase toward the ends of chromosomes, with a smaller peak near the middle of chromosomes, which is more pronounced in f_d than recombination rate. (C) Estimated recombination rates are correlated with both f_d, where regions with higher recombination rates tend to have higher f_d. Each point represents data binned to 5000bp regions for all three statistics. Both axes have been log-transformed for better visualization of the data. (D) Sliding window analysis of f_d across the genome demonstrates elevated values near the ends of chromosomes, as well as some well-defined peaks that indicate regions of increased admixture. The horizontal dashed line indicates the p-value = 0.01 for f_d estimates genome-wide, which was used as a significance threshold.

Discussion

Using whole genome data and multiple independent approaches to model the evolutionary history of Lake Tanganyika *Lates* species, we find that the origin of this radiation is much younger than the age of the LT basin, with colonization and diversification likely facilitated by climate-driven increases in precipitation that reconnected the lake to the Congo River system and allowed this novel predator to enter the LT ecosystem. This mid-Pleistocene origin (∼1–2 Mya) of the radiation has two major implications. First, it suggests that these top predator species evolved in the face of an already species-rich cichlid fauna, implying that their colonization of LT would likely have had dramatic ecosystem consequences. Second, it implies that all four species evolved during time periods when the lake may have undergone cycles of water level change that subdivided distinct basins, facilitating divergence and allopatric speciation. In addition, we detect multiple bouts of admixture between lacustrine *Lates* species and between the riverine *L. niloticus* and lacustrine *Lates*, further suggesting cyclic connectivity among these species.

Ecosystem implications of a young *Lates* clade. Both site frequency-based and Markovian coalescent divergence dating independently infer the origin of the Tanganyikan *Lates* clade to ∼1–2 Mya and the crown age of the lacustrine radiation to ∼ 0.70 Mya (Fig. 2). This age estimate is similar to but slightly younger than the result obtained via distinct methods and data from (44), which estimated a crown age of 0.87–2.35 Mya for the Tanganyikan *Lates* clade and the divergence from *L. niloticus* 2.01–4.59 Mya. These divergence estimates are much more recent than those for other endemic clades in the lake

(e.g., 54–57); current evidence suggests that other clades colonized the lake either near the onset of its formation 9–12 Mya or during the establishment of a truly lacustrine habitat and deepwater conditions 5–6 Mya (33), and radiated soon thereafter. If these clades colonized and radiated within LT prior to *Lates* colonization, then this implies that the *Lates* diversification happened in the face of an already species-rich fish community. Given that the vast majority of large East African lakes without *Lates* have diverse pelagic cichlid communities—and that those with endemic *Lates* generally lack diverse pelagic fish communities—it is possible that the introduction and diversification of *Lates* into LT would have had major ecosystem effects, namely through the extinction or competitive exclusion of a diverse pelagic cichlid community due to predation or competition from *Lates*.

With this insight, the recent anthropogenic introduction of *L. niloticus* into Lake Victoria acts as a modern analog for understanding the events within LT in the mid-Pleistocene following *Lates* colonization. In Lake Victoria, introduced *L. niloticus* were able to outcompete (39) and prey on native fishes (58), leading to altered community and food web structure within the lake over the course of a few generations. The introduction of *L. niloticus* into Lake Victoria interacted with changing limnological conditions (59, 60) to decimate cichlid stocks, resulting in the disappearance of about 40% of the 500+ endemic haplochromine cichlid species (38, 61), with extinction falling disproportionately among pelagic piscivorous species (39, 58, 62). While numerical abundance of cichlids has subsequently recovered among pelagic zooplanktivores and detritivores to reach pre-Nile perch numbers (58, 60), species diversity remains much reduced and pelagic piscivores show no signs of recovering (60). The disappearance of haplochromine species occurred concurrently with increases in the native cyprinid and shrimp populations (38, 63), shifting both community composition and food web structure within the lake. Similar shifts in food web structure occurred following Nile perch introduction into Lakes Kyoga and Nabugabo (63, 64), indicating that this may be a general trend for the introduction of these predators into native fish communities.

LT has few pelagic cichlid species in comparison to nearby lakes without *Lates*. If the Lake Victoria *Lates* invasion acts as an accurate modern analog to the process of ecosystem change which *Lates* introduction causes, we expect that LT would have had a diverse pelagic cichlid community prior to *Lates* invasion that would have been decimated with their arrival and diversification. Intriguingly, this history should be available through proxies contained in the sedimentary record of the lake; although currently available sediment cores do not extend far enough back in time to examine dynamics at this timescale (65), this hypothesis of dramatic ecosystem change upon *Lates* introduction is testable with long sediment cores in the future.

Cyclic environmental conditions may have facilitated both divergence and admixture between taxa. In tropical Africa, major tectonic and climatic events—such as continental rifting and glacial-interglacial cycles—have played a large role in shaping faunal biogeography for both aquatic and terrestrial clades. In this way, geology and climate together have led to periodic reorganization of corridors and barriers to dispersal in the region since the early Miocene (29, 30, and references therein). Although rifting began in the LT region 9–12 Mya (33) and the lake became a truly lacustrine environment 5–6 Mya, the lake environment has faced considerable change during this long history (66). LT's tie to the Congo River via the Lukuga River outflow is common in recent times; however it appears that this connection is only a recent phenomenon (∼1.5-2.5 Mya; 67), and thus the lake was isolated from the Congo basin for much of its history. The re-establishment of the Lukuga River influenced the divergence of chimpanzees and bonobos in the region (67–69), and our results suggest that it similarly had a strong influence on divergence in *Lates* via facilitating colonization of the Lukuga River (44) and subsequently LT around 1.2 Mya. Mitochondrial data (Fig. S13) further support a rapid divergence between *L. niloticus* from the Congo River, *L. stappersii*, *L. microlepis*, and a common ancestor of *L. mariae* and *L. angustifrons*, with mtDNA haplotypes found in *L. niloticus* from the Congo River remaining much more similar to the LT haplotypes than to those found in *L. niloticus* from the Niger or Nile River basins (Fig. S10).

Patterns of lake level fluctuations over the course of LT's long history undoubtedly influenced the evolution of its fauna, and likely were influential in the evolution of high levels of endemism within the lake (70).

8 of 16

Within LT, considerable fault-related lake floor features (e.g., the Kavala Island Ridge and Kalya Horst; Fig. 1) create barriers to movement of lacustrine taxa at lowstands. As a result, low water levels can separate previously panmictic populations for evolutionarily relevant lengths of time, and this has likely contributed to diversification in LT's littoral species (70, 71). Significant water level drops likely separated the lake's three major basins hydrologically (>400m) or physically (>500m), which could have created barriers (e.g., for pelagic species) or conduits (e.g., for littoral species) for gene flow between populations in geographically distant parts of the lake (Fig. 1; 71). Allopatric divergence in *Lates* may have been facilitated by periods of basin separation. While seismic reflection data confirm major water level fluctuations in LT (24, 66), we currently lack chronological control on the timeline of these events prior to ∼100 Kya (26). Yet water level fluctuations in nearby Lake Malawi (13) and correlations with global climate phenomena (Fig. S1; (27)) indicate that LT also experienced large water level fluctuations during the mid- to late-Pleistocene, concurrent with our inferred divergence times among the four LT endemic *Lates* species and introgression events involving *L. niloticus*. While there is uncertainty in the timing and magnitude of lake level changes, as well as in the timing of *Lates* divergence, this indicates a possible role for these fluctuations as a "species pump" (sensu 71) within the *Lates*, as has been suggested for cichlid species in LT and other lakes (72–75).

The most recent extreme drought conditions in the African Great Lakes region occurred ∼75–135 Kya, as evidenced by sediment cores from Malawi (which was reduced to a ∼125m deep saline, alkaline, well mixed lake; 26, 27) and Tanganyika (drop >400m; 28), although Tanganyika likely remained a large, hydrologically continuous water body despite this significant, basin-wide lake level decline (40, Fig. 1). The most recent admixture event that we detect, between *L. niloticus* and *L. angustifrons*, appears to have occurred toward the end of these known megadrought conditions (∼100 Kya, Fig. 2D), and thus may relate to a time of rising lake levels that reconnected the Lukuga River to LT. The group of admixture events prior to this, both within LT and between *L. niloticus* and *L. microlepis*, are dated to the late Pleistocene, during this period of extreme drought (median estimates ∼250-270 Kya). While our confidence intervals on admixture timing for these post-divergence admixture events are wide, this uncertainty may in fact suggest that there were multiple admixture events among these taxa in the period ∼250 Kya, a time period during which LT water levels were relatively low (Fig. S1; 26, 40, 76), dramatically altering the habitat availability and connectivity for populations of lacustrine fauna. Our two-taxon demographic models are consistent with the hypothesis that multiple bouts of admixture occurred between *Lates* taxa within LT (Fig. S15), but further investigations will be needed to pinpoint the timing and number of admixture events, as our current analyses are limited in the resolution of data available for inferring such events.

Variation in admixture across the genome. Our results show substantial variation in admixture proportion across the *Lates* genomes (Fig. 3). We see strong correspondence between admixture proportions (F_d) and chromosomal position, tied to variation in recombination rates across chromosomes (Fig. 3B,D). Relatedly, there is a genome-wide relationship between recombination rate and admixture proportion (Fig. 3C), demonstrating that alleles derived from introgression are selected against, and more likely to be retained in regions where high recombination rates reduce the effect of linked selection against deleterious introgressed alleles (e.g., 51, 52). These results have implications for understanding how selection and genome structure interact to influence genetic diversity across species' genomes, and how these forces therefore influence our inferences of species histories through phylogenetic approaches. In regions of the genome with high recombination, we observe the retention of more alleles derived from introgression, and thus infer topologies consistent with this history of introgression; in contrast in regions with low retention of introgressed alleles, with low recombination, we recover the species tree topology more often (Fig. 3A). These results underscore the value and increasingly, necessity, of understanding genome structure and recombination rates in order to effectively interpret species histories from genomic data.

Summary. Together, our analyses suggest a dynamic evolutionary history of the *Lates* radiation within LT, marked by rapid, recent, environmentally linked diversification. The colonization of the lake by these top predators in the face of an already species-rich community likely would have dramatically altered the

lake's community and food web, as has been observed via the anthropogenic introduction of related fish in other lakes in the region. While a clearer record of past lake conditions accessible through scientific drilling will clarify how wide-ranging the effects of *Lates* colonization and diversification were, we predict that the ecosystem consequences of this recent radiation would have been substantial, and would have left the biodiversity of LT irreversibly altered, making the *Lates* of LT a newfound example of the strong direct and indirect effects that predators can have on biodiversity, food webs, and community structure in novel ecosystems.

10 of 16

Acknowledgements

All analyses of genomic data requiring high performance computing were accomplished with an allocation from the University of Wyoming's Advanced Research Computing Center, on its Teton (https://doi.org/10.15786/M2FY47) and Beartooth (https://doi.org/10.15786/M2FY47) Computing Environments; GNU Parallel was used to parallelize computations (77). Special thanks go to Benedikt Ehrenfels for coordinating field campaigns, the late Mupape Mukuli for facilitating logistics during fieldwork, and to the crew of the MV Maman Benita for fieldwork assistance. We thank the whole team at the Tanzanian Fisheries Research Institute for their support and the Tanzanian Commission for Science and Technology (COSTECH) for the permits to collect fish from Lake Tanganyika. Live fish samples were collected under the approved University of Wyoming IACUC protocol #20160602CW00241-01. We would like to thank Chad Brock for discussions on admixture analyses and members of the extended Wagner Lab at the University of Wyoming for discussions of analyses and comments on the manuscript. Bathymetry data for Lake Tanganyika maps (Fig. 1) were shared by TCarta (Denver, CO). JAR was supported by fellowships from the Data Science Center, NASA Space Grant Consortium (NASA Grant #NNX15AI08H), and Office of Graduate Education at the University of Wyoming, and sequencing was accomplished with support from the University of Wyoming INBRE Data Science Core, which is supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under Grant #2P20GM103432. JAR was also supported by an NSF Postdoctoral Fellowship in Biology (award DBI-2109825). This work was also supported by NSF grants DEB-1556963 and DEB-2224892 to CEW.

Data sharing plan

All genomic data are archived in the NCBI Sequence Read Archive, under BioProjects PRJNA776855 (genotyping-by-sequencing) and PRJNA1207083 (whole genome sequencing, *note for review: these data are currently embargoed*). Metadata and scripts used in data processing and analysis are archived on Zenodo at [TBD], and scripts can additionally be found in the GitHub repository https://github.com/jessicarick/lates-admixture.

Declaration of conflict of interest

The authors declare no conflict of interest.

References

1. Carpenter SR, Kitchell JF, Hodgson JR (1985) Cascading Trophic Interactions and Lake Productivity. *BioScience* 35(10):634–639.
2. Estes JA, et al. (2011) Trophic downgrading of planet Earth. *Science* 333:301–306.
3. Alston J, et al. (2019) Reciprocity in restoration ecology: When might large carnivore reintroduction restore ecosystems? *Biological Conservation* 234:82–89.
4. Schmitz OJ, Krivan V, Ovadia O (2004) Trophic cascades: The primacy of trait-mediated indirect interactions.
5. Wainright CA, Muhlfeld CC, Elser JJ, Bourret SL, Devlin SP (2021) Species invasion progressively disrupts the trophic structure of native food webs. *Proceedings of the National Academy of Sciences* 118(45):e2102179118.
6. Koel TM, et al. (2019) Predatory fish invasion induces within and across ecosystem effects in Yellowstone National Park. *Science Advances* 5(3).
7. Mooney HA, Cleland EE (2001) The evolutionary impact of invasive species. *Proceedings of the National Academy of Sciences* 98(10):5446–5451.
8. He Z, et al. (2019) Speciation with gene flow via cycles of isolation and migration: insights from multiple mangrove taxa. *National Science Review* 6(2).
9. Linck E, Battey CJ (2019) On the relative ease of speciation with periodic gene flow. *bioRxiv* pp. 1–33.
10. Nevado B, Contreras-Ortiz N, Hughes C, Filatov DA (2018) Pleistocene glacial cycles drive isolation, gene flow and speciation in the high-elevation Andes. *New Phytologist* 219(2).
11. Tierney JE, et al. (2008) Northern hemisphere controls on tropical southeast African climate during the past 60,000 years. *Science* 322(5899):252–255.
12. DeMoncal PB (1995) Plio-Pleistocene African Climate. *Science* 270(5233):53–59.
13. Lyons RP, et al. (2015) Continuous 1.3-million-year record of East African hydroclimate, and implications for patterns of evolution and biodiversity. *Proceedings of the National Academy of Sciences* 112(51):2–7.
14. Barnosky AD, et al. (2004) Exceptional record of mid-Pleistocene vertebrates helps differentiate climatic from anthropogenic ecosystem perturbations, Technical report.
15. Aubry KB, Statham MJ, Sacks BN, Perrine JD, Wisely SM (2009) Phylogeography of the North American red fox: vicariance in Pleistocene forest refugia. *Molecular Ecology* 18(12).
16. Weir JT, Schluter D (2004) Ice sheets promote speciation in boreal birds. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 271(1551).
17. Claramunt S, Cracraft J (2015) A new time tree reveals Earth history's imprint on the evolution of modern birds. *Science Advances* 1(11).
18. Weir JT, Haddrath O, Robertson HA, Colbourne RM, Baker AJ (2016) Explosive ice age diversification of kiwi. *Proceedings of the National Academy of Sciences* 113(38):E5580–E5587.
19. Ayoub NA, Riechert SE (2004) Molecular evidence for Pleistocene glacial cycles driving diversification of a North American desert spider, Agelenopsis aperta. *Molecular Ecology* 13(11).
20. DeChaine EG, Wendling BM, Forester BR (2014) Integrating environmental, molecular, and morphological data to unravel an ice-age radiation of arctic-alpine Campanula in western North America. *Ecology and Evolution* 4(20).
21. Clarke A, Crame JA (2010) Evolutionary dynamics at high latitudes: speciation and extinction in polar marine faunas. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1558).
22. April J, Hanner RH, Dion-Côté AM, Bernatchez L (2013) Glacial cycles as an allopatric speciation pump in north-eastern American freshwater fishes. *Molecular Ecology* 22(2).
23. Tierney JE, et al. (2008) Northern Hemisphere Controls on Tropical Southeast African Climate During the Past 60,000 Years. *Science* 322(5899).
24. Russell JM, et al. (2020) ICDP workshop on the Lake Tanganyika Scientific Drilling Project: A late Miocene-present record of climate, rifting, and ecosystem evolution from the world's oldest tropical

12 of 16

lake. *Scientific Drilling* 27.
25. Verschuren D, et al. (2009) Half-precessional dynamics of monsoon rainfall near the East African Equator. *Nature* 462(7273).
26. Scholz CA, et al. (2007) East African megadroughts between 135 and 75 thousand years ago and bearing on early-moderns human origins. *Proceedings of the National Academy of Sciences* 104(42):16416–16421.
27. Cohen AS, et al. (2007) Ecological consequences of early Late Pleistocene megadroughts in tropical Africa. *Proceedings of the National Academy of Sciences* 104(42):16422–16427.
28. Burnett AP, Soreghan MJ, Scholz CA, Brown ET (2011) Tropical East African climate change and its relation to global climate: A record from Lake Tanganyika, Tropical East Africa, over the past 90+ kyr. *Palaeogeography, Palaeoclimatology, Palaeoecology* 303(1-4).
29. Salzburger W, Van Bocxlaer B, Cohen AS (2014) Ecology and Evolution of the African Great Lakes and Their Faunas. *Annual Review of Ecology, Evolution, and Systematics* 45(1):519–545.
30. Stewart KM (2001) The freshwater fish of Neogene Africa (Miocene-Pleistocene): systematics and biogeography. *Fish and Fisheries* 2(3).
31. Levin NE (2015) Environment and Climate of Early Human Evolution. *Annual Review of Earth and Planetary Sciences* 43(1):405–429.
32. Trauth MH, et al. (2010) Human evolution in a variable environment: the amplifier lakes of Eastern Africa. *Quaternary Science Reviews* 29(23-24):2981–2988.
33. Cohen AS, Soreghan MJ, Scholz CA (1993) Estimating the age of formation of lakes: an example from Lake Tanganyika, East African Rift system. *Geology* 21(6):511–514.
34. Groombridge B, Jenkins M (1998) Freshwater biodiversity: a preliminary global assessment, (United Nationals Environmental Programme - World Conservation Montitoring Centre, Cambridge, UK), Technical report.
35. Coulter GW, ed. (1991) *Lake Tanganyika and its Life*. (Oxford University Press, London).
36. Wagner CE, Harmon LJ, Seehausen O (2012) Ecological opportunity and sexual selection together predict adaptive radiation. *Nature* 487(7407):366–369.
37. Seehausen O, Wagner CE (2014) Speciation in Freshwater Fishes. *Annual Review of Ecology, Evolution, and Systematics* 45:621–651.
38. Goudswaard K, Witte F, Katunzi EF (2008) The invasion of an introduced predator, Nile perch (Lates niloticus, L.) in Lake Victoria (East Africa): Chronology and causes. *Environmental Biology of Fishes* 81(2):127–139.
39. McGee MD, et al. (2015) A pharyngeal jaw evolutionary innovation facilitated extinction in Lake Victoria cichlids. *Science* 350(6264).
40. McGlue MM, et al. (2008) Seismic records of late Pleistocene aridity in Lake Tanganyika, tropical East Africa. *Journal of Paleolimnology* 40(2):635–653.
41. Rick JA, et al. (2021) The genetic population structure of Lake Tanganyika's Lates 1 species flock, an endemic radiation of pelagic top predators 2 3. *bioRxiv*.
42. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
43. Zhang C, Rabiee M, Sayyari E, Mirarab S (2018) ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19(Suppl 6):15–30.
44. Koblmüller S, et al. (2021) African lates perches (Teleostei, Latidae, Lates): Paraphyly of Nile perch and recent colonization of Lake Tanganyika. *Molecular Phylogenetics and Evolution* 160.
45. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nature Genetics* 46(8):919–25.
46. Malinsky M, Matschiner M, Svardal H (2021) Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources* 21(2):584–595.
47. Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* 15:356.

48. Malinsky M, et al. (2018) Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology and Evolution* 2(12):1940–1955.
49. Martin SH, Davey JW, Jiggins CD (2015) Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution* 32(1):244–257.
50. Aeschbacher S, Selby JP, Willis JH, Coop G (2017) Population-genomic inference of the strength and timing of selection against gene flow. *Proceedings of the National Academy of Sciences of the United States of America* 114(27):7061–7066.
51. Schumer M, et al. (2018) Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* 660(May):656–660.
52. Martin SH, Davey JW, Salazar C, Jiggins CD (2019) Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLOS Biology* 17(2):e2006288.
53. Edelman NB, et al. (2019) Genomic architecture and introgression shape a butterfly radiation. *Science* 366:594–599.
54. Irisarri I, et al. (2018) Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. *Nature Communications* 9(1).
55. Wilson AB, Teugels GG, Meyer A (2008) Marine incursion: The freshwater herring of Lake Tanganyika are the product of a marine invasion into West Africa. *PLoS ONE* 3(4).
56. Ronco F, et al. (2021) Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature* 589(7840):76–81.
57. Milec LJM, et al. (2022) Complete mitochondrial genomes and updated divergence time of the two freshwater clupeids endemic to Lake Tanganyika (Africa) suggest intralacustrine speciation. *BMC Ecology and Evolution* 22(1):127.
58. Witte F, et al. (2007) Differential decline and recovery of haplochromine trophic groups in the Mwanza Gulf of Lake Victoria. *Aquatic Ecosystem Health & Management* 10(4).
59. Hecky RE, Mugidde R, Ramlal PS, Talbot MR, Kling GW (2010) Multiple stressors cause rapid ecosystem change in Lake Victoria. *Freshwater Biology* 55(SUPPL. 1):19–42.
60. van Zwieten P, et al. (2015) The Nile perch invasion in Lake Victoria: cause or consequence of the haplochromine decline? *Canadian Journal of Fisheries and Aquatic Sciences* 73(4):622–643.
61. Witte F, et al. (1992) The destruction of an endemic species flock : quantitative data on the decline of the haplochromine cichlids of Lake Victoria, Technical report.
62. Goldschmidt T, Witte F, Wanink J (1993) Cascading Effects of the Introduced Nile Perch on the Detritivorous/ Phytoplanktivorous Species in the Sublittoral Areas of Lake Victoria Cascading Effects of the Introduced Nile Perch on the Detritivorous/Phytoplanktivorous Species in the Sublittoral Areas of Lake Victoria TIJS GOLDSCHMIDT, Technical Report 3.
63. Ogutu-Ohwayo R (1990) Changes in the prey ingested and the variations in the Nile perch and other fish stocks of Lake Kyoga and the northern waters of Lake Victoria (Uganda). *Journal of Fish Biology* 37(1).
64. Ogutu-Ohwayo R (1993) The Effects of Predation by Nile Perch, Lates niloticus L., on the Fish of Lake Nabugabo, with Suggestions for Conservation of Endangered Endemic Cichlids. *Conservation Biology* 7(3).
65. Scholz CA, et al. (2003) Paleolimnology of Lake Tanganyika, East Africa, over the past 100 k yr. *Journal of Paleolimnology* 30:139–150.
66. Shaban SN, Scholz CA, Muirhead JD, Wood DA (2021) The stratigraphic evolution of the Lake Tanganyika Rift, East Africa: Facies distributions and paleo-environmental implications. *Palaeogeography, Palaeoclimatology, Palaeoecology* 575.
67. Flügel TJ, Eckardt Frank FD, Cotterill FP (2015) The present day drainage patterns of the Congo river system and their Neogene evolution in *Geology and Resource Potential of the Congo Basin.* (Springer Berlin Heidelberg), pp. 315–337.
68. Prüfer K, et al. (2012) The bonobo genome compared with the chimpanzee and human genomes.

14 of 16

Nature 486(7404).
69. Langergraber KE, et al. (2012) Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences* 109(39).
70. Coulter GW (1991) Zoogeography, affinities and evolution, with special regard to the fish in *Lake Tanganyika and its Life*, ed. Coulter GW. (Oxford University Press, London), pp. 275–305.
71. Rossiter A (1995) The Cichlid Fish Assemblages of Lake Tanganyika: Ecology, Behaviour and Evolution of its Species Flocks.
72. Duftner N, Koblmüller S, Sturmbauer C (2005) Evolutionary Relationships of the Limnochromini, a Tribe of Benthic Deepwater Cichlid Fish Endemic to Lake Tanganyika, East Africa. *Journal of Molecular Evolution* 60(3).
73. Sturmbauer C, Börger C, Van Steenberge M, Koblmüller S (2017) A separate lowstand lake at the northern edge of Lake Tanganyika? Evidence from phylogeographic patterns in the cichlid genus Tropheus. *Hydrobiologia* 791(1):51–68.
74. Koblmüller S, Duftner N, Katongo C, Phiri H, Sturmbauer C (2005) Ancient Divergence in Bathypelagic Lake Tanganyika Deepwater Cichlids: Mitochondrial Phylogeny of the Tribe Bathybatini. *Journal of Molecular Evolution* 60(3):297–314.
75. Ivory SJ, et al. (2016) Environmental change explains cichlid adaptive radiation at Lake Malawi over the past 1.2 million years. *Proceedings of the National Academy of Sciences of the United States of America* 113(42):11895–11900.
76. Tiercelin JJ, Mondeguer A (1991) The geology of the Tanganyika Trough in *Lake Tanganyika and its Life*, ed. Coulter GW. (Oxford University Press), p. 7–48.
77. Tange O (2011) GNU Parallel - The Command-Line Power Tool. *The USENIX Magazine* pp. 42–47.
78. Vij S, et al. (2016) Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. *PLoS Genetics* 12(4):1–35.
79. Martin M, et al. (2016) WhatsHap: fast and accurate read-based phasing. *bioRxiv*.
80. Chan AH, Jenkins PA, Song YS (2012) Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLoS Genetics* 8(12).
81. Martin SH, Van Belleghem SM (2017) Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. *Genetics* 206:429–438.
82. Brawand D, et al. (2015) The genomic substrate for adaptive radiation in African cichlid fish. *Nature* 513(7518):375–381.
83. Patterson N, et al. (2012) Ancient admixture in human history. *Genetics* 192(3):1065–1093.
84. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics* 9(10).
85. Murray AM, Attia YS (2004) A new species of Lates (Teleostei: Perciformes) from the Lower Oligocene of Egypt. *Journal of Vertebrate Paleontology* 24(2):299–308.
86. Meier JI, et al. (2017) Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications* 8(1).

Methods

Study species and sampling. Our analyses focus on four *Lates* species endemic to Lake Tanganyika, *L. angustifrons, L. mariae, L. microlepis*, *L. stappersii*, and *L. niloticus*. The *L. niloticus* specimens were obtained from tissues archived in the Cornell University Museum of Vertebrates (CUMV) and American Museum of Natural History (AMNH). A total of 152 tissue samples were sequenced using a genotyping-by-sequencing (GBS) protocol, as described in (41), and 13 of those were additionally used to generate whole genome sequences (WGS). Reads were aligned to the chromosome-level *L. calcarifer* genome (78). WGS haplotypes were phased using `WhatsHap` (79) and population recombination rates in all species combined across each chromosome were estimated using the maximum likelihood method implemented in `LDhelmet` (80).

Phylogenetic and diversification histories. We inferred phylogenetic relationships among taxa using both maximum-likelihood and gene tree methods, implemented in `RAxML` (42) and `ASTRAL-III` (43), from both GBS and WGS data. To investigate whether those regions of the genome exhibiting elevated levels of introgression contribute disproportionately to topological uncertainty, we conducted sliding window analysis of gene trees in `TWISST` (81). We estimated dates for divergence between taxa based on the relative cross-coalescence rates from the WGS data using `MSMC2`, using a generation time of 3 years and recombination rate of $3.5x10^{-9}$ (82) to convert coalescent units to years.

Admixture inferences. We used `Dsuite` (46) to calculate genome-wide D-statistics using both WGS and GBS data for all three-taxon combinations of *L. niloticus, L. angustifrons, L. mariae, L. stappersii,* and *L. microlepis*, again always keeping *L. calcarifer* as the outgroup. To determine the extent of introgression and the likely location of introgression events on the *Lates* phylogeny, we used `Dsuite` to calculate $f4$-ratio statistics (83) and, from these, calculate f-branch statistics (48) constrained to the maximum likelihood topology inferred in `RAxML`. To further investigate differential admixture signal across the genome, we calculated sliding window analyses of f_d. We calculated the z-score for each f_d value, as well as its associated p-value, and identified significantly admixed regions as those with $p < 0.01$.

Demographic modeling. To test competing hypotheses for the evolutionary history and gene flow among our focal species, we explicitly modeled demographic history based on the site frequency spectra (SFS) summary of WGS sequences using the continuous-time sequential Markovian coalescent method implemented in `fastsimcoal2` (84). We built 32 six-taxon models including the four Lake Tanganyika species, *L. niloticus*, and *L. calcarifer*. The models all had the same pattern of divergence history conditioned on the best `RAxML` species tree but differed in the number of admixture events between the taxa. In all models, we constrained the split between *L. calcarifer* and the freshwater *Lates* using the oldest known *Lates* fossil found in Africa (*L. qatraniensis,* ∼33 Mya; 85) by constraining this split to 11 million generations. We assessed model fit and performance via both likelihood distribution and AIC methods, following (86). We used a non-parametric block bootstrapping approach (86) to generate 100 bootstrap replicates of parameter estimates for our best model. We then constructed two-taxon `fastsimcoal2` models for all pairs of ingroup taxa (i.e., not including *L. calcarifer*) using the GBS data. For each pair, models with 0–3 admixture events (10 total models) were competed to infer the most likely number of admixture events between the given pair of taxa.

16 of 16

Supplementary Materials

Supplementary Methods

Sampling and DNA sequencing. Between 2001 and 2018, tissue samples (fin clips) were collected from *Lates* along the Tanzanian shore of Lake Tanganyika, as described in (1). Fish were predominantly collected opportunistically from fishermen at fishing grounds and landing sites. Fish were identified in the field by sampling personnel following United Nations Food and Agricultural Organization (FAO) identification guidelines (2) and guidance from local fishermen. The species can be difficult to distinguish morphologically, particularly in the juvenile stage, and field-based identifications were checked using genetic identification, as described in (1).

In addition to our own collections, we obtained tissues from two *Lates niloticus* collected in Cameroon from the Cornell University Museum of Vertebrates (CUMV Catalog No. 93501 and 90829) and one *L. niloticus* collected in the Democratic Republic of the Congo from the American Museum of Natural History (AMNH Catalog No. 216682). We additionally downloaded the chromosome-level reference genome assembly of one *L. calcarifer* (barramundi) individual (v3; 3) and the corresponding 100bp PE `fastq` reads from NCBI (SRR3140997). This genome was used as a reference for assembly, and the reads from this individual were included as an outgroup for phylogenetic analyses.

Fin clips were stored in 95% ethanol or RNAlater until extraction. We extracted DNA from fin clips using Qiagen DNeasy Blood and Tissue Kits (Qiagen, Inc.) following the standard protocol, with the addition of an RNAse A incubation step.

Genotyping-by-sequencing. Samples were sequenced using a genotyping-by-sequencing (GBS) protocol, as described in (1). Briefly, we prepared libraries following protocols outlined in (4), using MseI and EcoRI restriction enzymes. Digested and barcoded fragments were amplified via PCR and then pooled for sequencing in two libraries. The prepared libraries were size-selected for 200–350bp fragments using Blue Pippin (Sage Science, MA). Each library contained ∼192 individuals, and select individuals were duplicated across libraries to check for library compatibility.

We processed raw sequencing reads by first filtering common contaminants and unwanted sequences (PhiX, E. coli, and leftover barcodes, primers, and adapters from library preparation) from our data using `bowtie2` (v2.3.4.1; 5). We then matched sequence barcodes to individual fish and assembled all data to the *Lates calcarifer* reference genome (v3; 3) using `bwa mem` (v0.7.17; 6) with default settings. Following alignment, we discarded any individuals with <50,000 high-quality reads assembled to the reference genome. We then identified variable sites in the assembly using `samtools` and `bcftools` (v1.8; 7), omitting indels and keeping only high-quality biallelic variant sites (`QUAL >20` and `GQ >9`). We filtered SNPs by minor allele frequency (`-maf 0.01`) and amount of missing data (`-max-missing 0.5`) using `vcftools` (v0.1.14; 8), in addition to only calling SNPs at sites with a minimum read depth of 10.

Whole genome resequencing. In addition to the GBS dataset, we generated a set of whole genome resequencing data for 3 *L. stappersii*, 3 *L. mariae*, 3 *L. angustifrons*, 3 *L. microlepis*, and 1 *L. niloticus* individuals. Tissues for these individuals were sent to HudsonAlpha Discovery (Huntsville, AL) for high molecular weight DNA extraction, PCR-free library preparation, multiplexing, and 150bp paired-end sequencing. For library preparation, samples were sonicated on the Covaris LE220 using a protocol designed to achieve a target insert size of 550 bp. The samples were then prepared using the KAPA HyperPrep whole genome protocol and dual indices. Nine of the samples were sequenced across 3 lanes of Illumina HiSeq X (150bp PE), and four were sequenced on 10% each of an Illumina NovaSeq S4 flowcell (150bp PE), resulting in ∼40x and ∼100x coverage at sequenced sites, respectively.

Short reads from whole genome sequencing were prepared in a similar manner to the GBS reads. First, read quality was checked using FastQC (v0.11.19; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Then, common contaminants and unwanted sequences were removed from the data using `bowtie2`, followed by assembling the remaining reads to the *L. calcarifer* reference genome using `bwa mem` with default settings. We then created an `mpileup` with all 13 individuals using `samtools` and created three separate filtered datasets using `vcftools`, appropriate for the different types of analyses: (1) a whole genome dataset, where we filtered for missing data but kept both variable and invariant sites; (2) a full SNP dataset, where we filtered for biallelic sites, <50% missing data (`-max-missing 0.5`), and minor allele frequency greater than 0.01 (`-maf 0.01`); and (3) a linkage-filtered SNP dataset, where we filtered for biallelic sites, <20% missing data, minor allele frequency greater than 0.01. We additionally filtered all datasets for sites with a minimum read depth greater than or equal to 30 (roughly 1.5x the interquartile range of read depths across sites).

Mitochondrial DNA sequencing. A subset of samples were selected for mitochondrial sequencing. We amplified the mitochondrial 16s region using primers 16sL2510 (CGC CTG TTT ATC AAA AAC AT) and 16sH3058 (CCG GTC TGA ACT CAG ATC ACG T), as well as internal sequencing primers 16s135F (GCA ATA GAV AWA GTA CCG CAA GGG) and 16s1543R (AAC TGA CCT GGA TTR CTC CGG TC) (9). The PCR products were visualized on agarose gels to check the quality and size of amplification before sequencing. Sequencing was carried out by the Cornell University Biotechnology Resource Center (Ithaca, NY). Standard chromatographic curves of forward and reverse sequences were imported into the program Geneious (https://www.geneious.com) and manually aligned and edited. Consensus sequences were then combined with 16s *Lates* sequences downloaded from NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/), and aligned in Geneious using MAFFT sequence alignment with default parameters for DNA. We imported alignments were into PopART (10) and constructed a median-joining haplotype network (11) using an epsilon value of 0 to visualize haplotype relationships.

Genetic species affinities. Because we did not have definitive identification to species for every specimen in our GBS analyses, we used hierarchical Bayesian clustering in the program `entropy` (12, 13) to determine species identity for each specimen, as detailed in (1). For the full set of Lake Tanganyika *Lates* GBS individuals filtered for 50% missing data and a minor allele frequency of 0.01, we ran `entropy` for K=4, and used the results to assign individuals to the group for which group membership (q) was greater than 0.6. We assigned each of these inferred groups to a species using individuals for which we had definitive species assignments based on morphology.

Phylogenetic relationships. We used multiple approaches to infer phylogenetic relationships among the taxa included in this study. First, we estimated a maximum-likelihood (ML) phylogenetic tree from the concatenated GBS data using the hybrid, MPI/PThreads version of `RAxML v8.1.17` (14). We concatenated all variant sites that had <50% missing data and a minor allele frequency >0.01, and used the Lewis correction for invariant sites with the ASC_GTRGAMMA model of molecular evolution (Lewis 2001) within RAxML to construct the maximum likelihood phylogeny, including *L. calcarifer* as an outgroup. We additionally estimated the maximum-likelihood phylogeny from the concatenated WGS dataset containing all SNPs with <50% missing data and minor allele frequency >0.05, again using the ASC_GTRGAMMA model and Lewis ascertainment bias correction. For each dataset, 100 bootstrap replicates were used to infer nodal support.

We estimated second phylogenetic tree using the multi-species coalescent with the program `SNAPP` (15) using the whole genome data. For `SNAPP` analyses, we started with the linkage-filtered WGS data set and thinned randomly to one SNP per 10,000bp. We tuned operators and priors in `SNAPP` to produce effective sample sizes of at least 200 for all parameters, running each run for 10 million steps, sampling every 1000 steps following a burn-in of 500,000 steps. The first 20% of samples were treated as burn-in and removed before individual runs were combined in `logcombiner`. Stationarity and chain convergence were assessed using the program `Tracer` (v1.7.1, 16).

We further generated a phylogenetic tree from individual gene trees using `ASTRAL-III` (v5.7.1, 17). `ASTRAL` finds the species tree that has the maximum number of shared induced quartet trees with the set of

2 of 33

gene trees, and is statistically consistent under the multi-species coalescent. Gene trees were inferred using `RAxML` for 4,000bp windows across the genome (118,707 windows), again using the ASC_GTRGAMMA model of molecular evolution and the Lewis correction for invariant sites. We inferred the species tree in `ASTRAL` using the bootstrap gene trees, and calculated local posterior probabilities (Sayyari & Mirarab) and quartet support for each node.

Estimating N_e and divergence timing. To estimate effective population sizes (N_e) through time and approximate dates for divergence between taxa, we used `MSMC2` (18). We generated input files for individual samples using `samtools` and custom scripts from the `msmc-tools` repository (https://github.com/stschiff/msmc-tools), filtering for a minimum read depth of 10 and minimum mapping quality of 20. In addition, a mappability mask was generated for the *L. calcarifer* genome using the `SNPable regions` program (http://lh3lh3.users.sourceforge.net/snpable.shtml) to select only regions to which short reads could be uniquely assembled using 35bp k-mers and a stringency of 50% (∼89.6 % of the genome). Variant sites were phased using `WhatsHap` (v1.0; 19) and input files generated using scripts from the `msmc-tools` repository. Input files were generated for each sample individually (for N_e estimation), for each species, and for each pair of individuals (for cross-coalescence). For individuals and species, we used a time pattern of $1*3+3*2+30*1+3*2+1*3$ and a maximum of 100 expectation maximization iterations. For cross-coalescence analyses, we reduced the time patterning to $1*3+3*2+20*1+3*2+1*3$ to increase the number of SNPs in each time segment. For runs that had problems with overfitting (11 of 78 comparisons), we further reduced the time patterning to $1*3+3*2+20*1+3*2+1*3$. Coalescent-scaled units were scaled to biological units using $\mu = 3.5x10^{-9}$ (estimate from Lake Malawi cichlids; 20) per generation and a mean generation time of 3 years (estimated from data in (21); see *Calculating generation time* section below). To estimate uncertainty in our N_e estimates, we performed 50 bootstrap replicates for each individual and each species. We calculated the divergence time between a pair of taxa as the mean T_{50}, the time at which the cross-coalescence rate reaches 0.5, for all combinations of individuals belonging to those species.

Identifying and characterizing introgression. Analysis of gene flow must take ILS into account because both processes generate gene trees that are incongruent with the species tree. The D-statistic (or ABBA-BABA statistic) is a useful and widely applied parsimony-like method for detecting gene flow despite the existence of ILS (22). In a comparison of taxa ((P1)P2)P3)P4), a positive value represents an excess of "ABBA" alleles, indicating excess allele sharing between P2 and P3; a negative value represents an excess of "BABA" alleles, indicating excess allele sharing between P1 and P3, compared to the null expectation under ILS only. We used `Dsuite` (v0.4; 23) to calculate genome-wide D-statistics for all three-taxon combinations of *L. niloticus, L. angustifrons, L. mariae, L. stappersii,* and *L. microlepis*, again always keeping *L. calcarifer* as the outgroup (H4). Significant values of D can represent either the wrong tree or admixture between those taxa indicated to have an excess of shared alleles; these two hypotheses can be distinguished by testing all possible combinations of taxa.

To determine the extent of introgression and the likely location of introgression events on the *Lates* phylogeny, we used `Dsuite` to calculate $f4$-ratio statistics (24) and, from these, calculate f-branch statistics (20). The f-branch metric, or $f_b(C)$, is a summary of f_4-ratio scores that capture excess allele sharing between a species C and a branch b of a given tree. In this way, the f-branch statistic deals with the possibility of correlated f_4 statistics related to a single gene flow event and also is able to infer the directionality of introgression. For f-branch calculations, we calculated statistics constrained to (1) the maximum likelihood topology from `RAxML` in our analyses, and (2) the species tree from (25). We calculated these statistics for both the GBS and WGS variant-only SNP datasets (both filtered for <50% missing data and minor allele frequency >0.01), separating the two *L. niloticus* individuals into separate "West Africa" and "Congo" lineages in the GBS dataset, based on collection locations and the recent suggestion that these two lineages of *L. niloticus* may be paraphyletic with respect to the Lake Tanganyika radiation (25). We additionally calculated f-branch statistics for each chromosome separately for the WGS dataset constrained to the maximum likelihood topology from our `RAxML` analyses to quantify broad-scale differences in rates of

introgression across the genome.

To further investigate differential patterns of admixture signal across the genome, we calculated sliding window analyses of f_d. The f_d statistic is related to D, in that it tests for excess allele sharing among taxa, but is designed for use in smaller regions of the genome and is used to identify regions of excess allele sharing contributing to the overall D-statistic observed (26). To quantify the proportion of the genome involved in admixture, we calculated f_d using `Dsuite Dinvestigate` for 100-SNP sliding windows with 20-SNP step size, iterating over all trios identified in `Dsuite` as having elevated D-statistics. The f_d statistic gives meaningless values if D is negative, and thus f_d values for windows with negative D were converted to zero, as recommended by the authors. We calculated the z-score for each f_d value, as well as its associated p-value, and identified significantly admixed regions as those with $p < 0.01$. We identified continuous admixed blocks as those where sequential windows exhibited an elevated signal of admixture, and calculated the length of these blocks. We then compared the distribution of admixed blocks among species comparisons and among chromosomes, and calculated the mean admixed block size and mean proportion of the sequenced genome that exhibited a significantly elevated admixture signal.

To investigate whether those regions of the genome exhibiting elevated levels of introgression contribute disproportionately to topological uncertainty, we conducted sliding window analysis of gene trees in `TWISST` (27). We converted our WGS dataset from `Dsuite` to a `geno` file using `parseVCF.py` script from Simon Martin's `genomics_general` repository (release v0.4) after imputing genotypes using `beagle` (4.0, version r1399; 28). We then used `RAxML` to create locus trees for 100-SNP sliding windows for the five species, and analyzed these trees using `TWISST`, within which we identified *L. calcarifer* as the outgroup and specified the "complete" method of weighting, which calculates the exact topology weightings by considering all subtrees.

To compare f_d windows and topology weightings to recombination rates across the genome, we estimated population recombination rates across each chromosome using the maximum likelihood method implemented in `LDhelmet` (v1.10; 29). We converted phased variants from `MSMC` analyses into haplotype fastas for each chromosome using `bcftools consensus`. We estimated recombination rate in 100-SNP windows, used the recommended grid of precomputed pairwise likelihoods, used Padé coefficients in asymptotic sampling, and ran the reversible jump MCMC for 1,000,000 iterations following 100,000 burn-in steps. We repeated the MCMC sampling at block sizes of 20, 30, and 50 for each chromosome, and averaged the resulting distributions of recombination rate estimates.

Intervals with any f_d below the $p = 0.01$ threshold were identified as having elevated signatures of admixture, and SNPs within these intervals were removed for subsequent phylogenetic inference in `SNAPP`. These sites ($\sim$0.02%) were removed using `bedtools intersect` (v2.27.1, 30) and `PLINK` (v1.90, 31), and the remaining sites were further filtered for no missing data, linkage disequilibrium (sites in 1000bp windows with pairwise $r^2 > 0.5$ were removed, removing $\sim$97% of remaining sites), and thinned to one SNP per 5000bp. We again tuned operators and priors in `SNAPP` to produce effective sample sizes of at least 200 for all parameters, running each run for 10 million steps, sampling every 1000 steps following a burn-in of 500,000 steps. The first 20% of samples were treated as burn-in and removed before individual runs were combined in `logcombiner`. Stationarity and chain convergence were again assessed using `Tracer`.

Demographic modeling. To test competing hypotheses for the evolutionary history and gene flow among our focal species, we explicitly modeled demographic history using `fastsimcoal2` (v2603, 32). Analyses in `fastsimcoal2` use a continuous-time sequential Markovian coalescent approximation to estimate demographic parameters from the site frequency spectrum (SFS). We used the WGS dataset including all sites with genotypes for all individuals (no missing data) with read depth >30 and <1000, and generated the SFS using easySFS (https://github.com/isaacovercast/easySFS) with no down projection. We built both six-taxon models including the four Lake Tanganyika species, *L. niloticus*, and *L. calcarifer*, and two-taxon models for all combinations of ingroup taxa (i.e., *L. angustifrons*, *L. mariae*, *L. microlepis*, *L. stappersii*, and *L. niloticus*).

4 of 33

Six taxon models. We tested a total of 32 six-taxon demographic models, all having the same pattern of divergence history conditioned on the best `RAxML` tree but differing from one another by the number of admixture events between the taxa (see Supplemental Methods; Fig. S2). For each model, `fastsimcoal2` was used to estimate the expected multi-SFS generated using 400 independent runs, each involving 200,000 coalescent simulations. We performed parameter estimation using the maximum composite likelihood from the SFS (`-M`), a minimum SFS count threshold of 10 (`-C 10`), and 50 expectation maximum cycles (`-L 50`). For each set of 400 runs, the run that led to the highest maximum likelihood was selected as the best. In all models, we constrained the split between *L. calcarifer* and the freshwater *Lates* using the oldest known *Lates* fossil found in Africa (*L. qatraniensis* from the Jebel Qatrani Formation of the Fayum Depression in Egypt, ∼33 Mya; 33) by constraining this split to 11 million generations.

We assessed model fit and performance via two approaches, following (34). First, we compared each model's best run using the Akaike information criterion (AIC), adjusted to penalize parameters by a factor of 200 due to the large likelihoods generated in six-taxon models (35). Second, to determine whether models are truly different or if they differ simply due to stochasticity in the likelihood approximation, we ran each model with its best parameter values 300 times using 1,000,000 coalescent simulations. The likelihoods differ between these runs because `fastsimcoal` approximates the likelihood using simulations, rather than calculating the likelihood exactly. If the ranges of likelihoods of two models overlap, it suggests that they provide an equally good fit to the observed data. This additional method of model selection circumvents some of the problems with AIC, as AIC can overestimate support for the most likely model in situations where SNPs are not independent, and we did not filter our data explicitly for linkage.

To obtain confidence intervals around parameter estimates of the best fit model identified with AIC and the likelihood distributions, we used a non-parametric block bootstrapping approach, following (34). Briefly, 100 bootstrap SNP datasets were created by randomly sampling 10,000-SNP blocks from the original dataset with replacement to form datasets of roughly equivalent size to the original, followed by conversion to an SFS using `easySFS`. For each bootstrap SFS, `fastsimcoal2` was then used to estimate the expected multi-SFS from 200 independent runs using the same settings as the original best fit model, with 100,000 coalescent simulations per run and using the best fit model's parameter estimates as starting values. The best run from each of these 100 bootstrap datasets was then selected and parameter estimates compiled to generate distributions and 95% bootstrap confidence intervals for each parameter in the model.

Alternate fossil calibrations. For the best model in our six-taxon `fastsimcoal2` model set, we tested the influence of our fossil calibration on the dates inferred for divergence times between *Lates* taxa by re-running this model with an alternate fossil calibration scheme, following (25). For the alternate fossil calibration, we constrained the split between *L. calcarifer* and the freshwater *Lates* taxa to 5.3 million generations, based on the oldest *L. niloticus* fossil known (25, 36), and kept all other parameterizations the same. We again used `fastsimcoal2` to estimate the expected multi-SFS generated from 400 independent runs, each consisting of 200,000 coalescent simulations per estimate, generated by a maximum of 50 expectation-maximization cycles. Parameter estimation was performed using maximum composite likelihood from the SFS (`-M`) and a minimum SFS count threshold of 10 was used for likelihood estimation (`-C 10`). From the set of 400 runs, the run that led to the highest maximum likelihood was selected as the best. For the best run, we again obtained confidence intervals around parameter estimates via a non-parametric block bootstrapping approach with 100 bootstrap replicates.

Two taxon models. To further investigate the admixture events identified in our best six-taxon `fastsimcoal` model, we parameterized and competed a series of two-taxon models for each two-species pair of ingroup taxa (i.e., *L. niloticus*, *L. angustifrons*, *L. mariae*, *L. microlepis*, *L. stappersii*). These two-taxon models involved divergence between the two species, followed by no admixture events ("noadmix"), one admixture event ("1admix"), two admixture events ("2admix"), or three admixture events ("3admix"). In addition, we tested hypotheses of divergence with gene flow between taxa, where continuous gene flow occurred from divergence to some point in the past, following which either no admixture events, one admixture event, or two admixture

events occurred. For each of the evolutionary hypotheses involving admixture events, we had one model with unidirectional gene flow (i.e., only from one species into the other) and one model with bidirectional gene flow, where the migration rates were allowed to vary between the directions (thus allowing for bidirectional but uneven rates of gene flow). All of these models included invariant sites and the mutation rate was fixed at $3.5x10^{-9}$ to allow for absolute estimates of parameters. Files containing model specifications (i.e., `tpl` and `est` files) can be found on GitHub at http://github.com/jessicarick/lates-admixture.

We estimated the site frequency spectrum for each two-species pair from a combination of the genotyping-by-sequencing and whole genome sequencing data. To prepare these data, we used `bcftools mpileup` and `bcftools call` to generate a `VCF` containing all called sites for all individuals from all species from individual `BAM` files. We then subsetted this `VCF` for each pair of species, retaining only the sequences belonging to the two species. We then filtered the sites in each subsetted `VCF` using `vcftools`, choosing to retain those sites with a minimum quality >20, minimum read depth >10, and a maximum of 30% missing data. We used `easySFS -preview` to calculate the best number of individuals to retain for down-projection from each `VCF`, choosing the projection values that would maximize both the number of sites and the number of individuals retained. Down-projection values can be found in Table SX.

We estimated the expected two-dimensional SFS from 300 independent runs, each consisting of 200,000 coalescent simulations per run, generated by a maximum of 50 expectation-maximization cycles. Parameter estimation was performed using maximum composite likelihood from the SFS and a minimum SFS count threshold of 10 was used for likelihood estimation. For each set of 400 runs, the run that led to the highest maximum likelihood was selected as the best. As with the six-taxon models, we used a combination of AIC and likelihood distributions (re-estimated 300 times using 100,000 coalescent simulations) to evaluate model fit and performance, selecting the best model for each species pair as that with the lowest AIC and highest median likelihood.

Calculating generation time. As there are not reliable generation time estimates for all of the species we are studying, we calculated the mean generation time using data from FishBase (21) and the calculations outlined there, which are based on the von Bertalanffy Growth Function (37):

$$L_t = L_\infty(1 - e^{-K(t-t_0)}) \quad [1]$$

Values of K (the growth coefficent), M (the rate of natural mortality), and L_∞ (the asymptotic length) were pulled from FishBase for each species, as available. As suggested in (21), we used t_{opt}, the time at which a fish reaches L_{opt}, as an approximation of generation time. For species without a published estimate for M, we calculated M from T (the annual mean water temperature) as $M = 10^{0.566-0.718*log(L_\infty)} + 0.02T$. The L_{opt} can then be estimated from K and M as $L_{opt} = L_\infty(\frac{3}{3+\frac{M}{K}})$ (38). The value for t_0 was calculated using the following equation (39):

$$log(-t_0) = -0.3922 - 0.2752log(L_\infty) - 1.038log(K) \quad [2]$$

We then calculated the mean t_{opt} for the six species (mean $t_{opt} = 2.96$), and rounded this estimate to 3 for our generation time in `fastsimcoal2` and `MSMC2` analyses. The longest generation time was 3.58 (*L. niloticus*), while the shortest was 2.38 (*L. stappersii*). Calculated generation times were relatively similar to the ages at sexual maturity reported in the literature, where available (21).

Supplementary Results

Genotyping-by-sequencing (GBS) resulted in short-read DNA sequence data for 150 individual *Lates* from a broad geographic area along the Tanzanian shore of Lake Tanganyika (Fig. 1), as well as three *L. niloticus* museum specimens (collected in Cameroon and from the Congo River), all of which were assembled to the chromosomal-level *Lates calcarifer* draft genome (V3, 3). The Nile perch (*L. niloticus*) is widespread across the major lakes and river basins of tropical sub-Saharan Africa, and is hypothesized to be the colonizer of

6 of 33

early LT that then diverged into the four LT-endemic *Lates*. We obtained 258,651,838 raw reads from the sequenced GBS libraries, which were filtered and assembled to the *Lates calcarifer* draft genome (Version 3; 3), resulting in an average of 1.0 million reads assembled per individual and 161 individuals retained after filtering for low coverage (< 20,000 reads). These individuals include 158 Lake Tanganyika *Lates* (31 *L. angustifrons*, 37 *L. mariae*, 18 *L. microlepis* and 72 *L. stappersii*), 2 *L. niloticus* from Cameroon, and 1 *L. calcarifer* from southeast Asia, whose FASTQ reads were downloaded from the NCBI Sequence Read Archive (SRA run SRR3140997). Whole genome sequencing (WGS) data were obtained for a subset of 3 individuals from each Tanganyika-endemic species and 1 *L. niloticus* from the Congo River. In addition, we included one *L. calcarifer* individual as an outgroup in phylogenetic analyses for both GBS and WGS datasets (3). The median genome coverage at sequenced sites was 14x per individual for GBS and 63.6x per individual for WGS, and assembly statistics did not differ systematically among taxa.

In contrast to the other phylogenetic reconstruction methods, `SNAPP` converged on a tree in which *L. niloticus* nested within the LT species, making the LT-endemic clade paraphyletic; however, species tree methods that do not explicitly account for admixture are known to be sensitive to and biased by admixture among species (40). After removing putatively admixed regions, we find that the `SNAPP` phylogeny matches the topology from `RAxML`, and that we once again infer a monophyletic LT clade.

References

1. Rick JA, et al. (2021) The genetic population structure of Lake Tanganyika's Lates 1 species flock, an endemic radiation of pelagic top predators 2 3. *bioRxiv*.
2. Eccles DH (1992) FAO Species identification sheets for fishery purposes. Field guide to the freshwater fishes of Tanzania, (Food and Agriculture Organization of the United Nations, Rome), Technical report.
3. Vij S, et al. (2016) Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. *PLoS Genetics* 12(4):1–35.
4. Parchman TL, et al. (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology* 21(12):2991–3005.
5. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9(4):357–359.
6. Li H, Durbin R (2009) Fast and accurate short-read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
7. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
8. Danecek P, et al. (2011) The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.
9. Li C, Ricardo BR, Leo Smith W, Ortí G (2011) Monophyly and interrelationships of Snook and Barramundi (Centropomidae sensu Greenwood) and five new markers for fish phylogenetics. *Molecular Phylogenetics and Evolution* 60(3):463–471.
10. Leigh JW, Bryant D (2015) POPART: full-feature software for haplotype network construction. *Methods in Ecology and Evolution* 6(9):1110–1116.
11. Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* 16(1):37–48.
12. Gompert Z, et al. (2014) Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology* 23(18):4555–4573.
13. Shastry V, et al. (year?) Model-based genotype and ancestry estimation for potential hybrids with mixed-ploidy.
14. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
15. Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, Roychoudhury A (2012) Inferring species trees

directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution* 29(8):1917–1932.
16. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA (2018) Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67(5).
17. Zhang C, Rabiee M, Sayyari E, Mirarab S (2018) ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19(Suppl 6):15–30.
18. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nature Genetics* 46(8):919–25.
19. Martin M, et al. (2016) WhatsHap: fast and accurate read-based phasing. *bioRxiv*.
20. Malinsky M, et al. (2018) Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology and Evolution* 2(12):1940–1955.
21. Froese R, Pauly D (2021) FishBase.
22. Green RE, et al. (2010) A draft sequence of the neandertal genome. *Science* 328(5979):710–722.
23. Malinsky M, Matschiner M, Svardal H (2021) Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources* 21(2):584–595.
24. Patterson N, et al. (2012) Ancient admixture in human history. *Genetics* 192(3):1065–1093.
25. Koblmüller S, et al. (2021) African lates perches (Teleostei, Latidae, Lates): Paraphyly of Nile perch and recent colonization of Lake Tanganyika. *Molecular Phylogenetics and Evolution* 160.
26. Martin SH, Davey JW, Jiggins CD (2015) Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution* 32(1):244–257.
27. Martin SH, Van Belleghem SM (2017) Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. *Genetics* 206:429–438.
28. Browning SR, Browning BL (2007) Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics* 81(5):1084–1097.
29. Chan AH, Jenkins PA, Song YS (2012) Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLoS Genetics* 8(12).
30. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6).
31. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81(3).
32. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics* 9(10).
33. Murray AM, Attia YS (2004) A new species of Lates (Teleostei: Perciformes) from the Lower Oligocene of Egypt. *Journal of Vertebrate Paleontology* 24(2):299–308.
34. Meier JI, et al. (2017) Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications* 8(1).
35. Johri P, et al. (2021) The Impact of Purifying and Background Selection on the Inference of Population History: Problems and Prospects. *Molecular Biology and Evolution* 38(7).
36. Otero O, Gayet M (2001) Palaeoichthyofaunas from the Lower Oligocene and Miocene of the Arabian Plate: palaeoecological and palaeobiogeographical implications. *Palaeogeography, Palaeoclimatology, Palaeoecology* 165(1-2):141–169.
37. Von Bertalanffy L (1938) A quantitative theory of organic growth (inquiries on growth laws. II). *Human Biology* 10(2).
38. Beverton RJ (1992) Patterns of reproductive strategy parameters in some marine teleost fishes. *Journal of Fish Biology* 41.
39. Pauly D (1979) Gill size and temperature as governing factors in fish growth: a generalization of von Bertalanffy´s growth formula.
40. Leaché AD, Harris RB, Rannala B, Yang Z (2014) The Influence of Gene Flow on Species Tree

 8 of 33

Estimation: A Simulation Study. *Systematic Biology* 63(1).
41. Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* 15:356.
42. Soraggi S, Wiuf C, Albrechtsen A (2018) Powerful Inference with the D-Statistic on Low-Coverage Whole-Genome Data. *G3 Genes/Genomes/Genetics* 8(2):551–566.
43. Ivory SJ, et al. (2016) Environmental change explains cichlid adaptive radiation at Lake Malawi over the past 1.2 million years. *Proceedings of the National Academy of Sciences of the United States of America* 113(42):11895–11900.
44. Lisiecki LE, Raymo ME (2005) A Pliocene-Pleistocene stack of 57 globally distributed benthic δ 18O records. *Paleoceanography* 20(1).
45. Tierney JE, et al. (2008) Northern hemisphere controls on tropical southeast African climate during the past 60,000 years. *Science* 322(5899):252–255.
46. Scholz CA, et al. (2007) East African megadroughts between 135 and 75 thousand years ago and bearing on early-moderns human origins. *Proceedings of the National Academy of Sciences* 104(42):16416–16421.

Supplementary Tables and Figures

Table S1. Whole genome D-statistics for all possible three-taxon combinations, using *L. calcarifer* as the outgroup. Values were calculated in `ANGSD` (41, 42) using the whole genome sequencing data and are sorted from highest to lowest z-score, where a positive D and z-score indicate excess allele sharing P2 $<->$ P3 and a negative D and z-score indicate excess allele sharing P1 $<->$ P3. Columns indicate the three taxa in the comparison (P1, P2, P3), the outgroup taxon (P4), the D-statistic (D), the block jackknife estimate of D (JK-D), the variance of the block jackknife estimate of D (V(JK-D)), the z-score associated with D (Z), the p-value associated with the z-score (pvalue), the number of ABBA sites (nABBA), the number of BABA sites (nBABA), and the number of blocks used in the calculation of D for each comparison.

P1	P2	P3	P4	D	JK-D	V(JK-D)	Z	pvalue	nABBA	nBABA	nBlocks
Lmar	Lmic	Lsta	Lcal	0.0866	0.0866	0.000002	70.09	0	394819.771	331908.969	5878
Lmar	Lmic	Lnil	Lcal	0.0425	0.0425	0.000001	41.35	0	311563.154	286149.783	5878
Lmar	Lsta	Lmic	Lcal	0.0241	0.0241	0.000002	17.75	0	394819.771	376244.369	5878
Lang	Lmic	Lsta	Lcal	0.0057	0.0057	0.000002	4.29	0.000018	340242.189	336390.058	5878
Lmar	Lsta	Lnil	Lcal	0.0031	0.0031	0.000002	2.49	0.0129	314364.87	312397.34	5878
Lmar	Lmic	Lang	Lcal	-0.0170	-0.0170	0.000001	-14.74	0	375933.156	388953.499	5878
Lmic	Lsta	Lnil	Lcal	-0.0375	-0.0375	0.000001	-31.31	0	292217.736	314959.675	5878
Lang	Lsta	Lmic	Lcal	-0.0751	-0.0751	0.000002	-49.41	0	340242.189	395463.921	5878
Lmic	Lsta	Lmar	Lcal	-0.0626	-0.0626	0.000001	-54.34	0	331908.969	376244.369	5878
Lmic	Lsta	Lang	Lcal	-0.0807	-0.0807	0.000001	-67.13	0	336390.058	395463.921	5878
Lmar	Lsta	Lang	Lcal	-0.0927	-0.0927	0.000002	-74.54	0	352858.195	424974.049	5878
Lang	Lmic	Lnil	Lcal	-0.0851	-0.0851	0.000001	-76.00	0	254073.084	301341.484	5878
Lang	Lmar	Lsta	Lcal	-0.0910	-0.0910	0.000001	-84.06	0	294003.421	352858.195	5878
Lang	Lsta	Lnil	Lcal	-0.1153	-0.1153	0.000002	-88.76	0	258719.328	326142.244	5878
Lang	Lmar	Lmic	Lcal	-0.1136	-0.1136	0.000001	-101.93	0	299249.637	375933.156	5878
Lang	Lmic	Lmar	Lcal	-0.1303	-0.1303	0.000001	-109.69	0	299249.637	388953.499	5878
Lang	Lmar	Lnil	Lcal	-0.1295	-0.1295	0.000001	-121.26	0	242830.608	315083.583	5877
Lang	Lsta	Lmar	Lcal	-0.1822	-0.1822	0.000002	-146.06	0	294003.421	424974.049	5878
Lmar	Lnil	Lsta	Lcal	-0.3811	-0.3811	0.000004	-198.70	0	314364.87	701597.593	5878
Lsta	Lnil	Lang	Lcal	-0.3657	-0.3657	0.000003	-214.64	0	326142.244	702140.923	5878
Lmar	Lnil	Lmic	Lcal	-0.3967	-0.3967	0.000003	-221.53	0	311563.154	721216.779	5878
Lmar	Lnil	Lang	Lcal	-0.4165	-0.4165	0.000003	-230.76	0	315083.583	764835.701	5878
Lsta	Lnil	Lmic	Lcal	-0.4036	-0.4036	0.000003	-233.02	0	314959.675	741204.149	5878
Lsta	Lnil	Lmar	Lcal	-0.3838	-0.3838	0.000003	-233.20	0	312397.34	701597.593	5878
Lmic	Lnil	Lsta	Lcal	-0.4345	-0.4345	0.000003	-237.38	0	292217.736	741204.149	5878
Lmic	Lnil	Lang	Lcal	-0.4199	-0.4199	0.000003	-247.47	0	301341.484	737583.101	5878
Lang	Lnil	Lsta	Lcal	-0.4615	-0.4615	0.000003	-248.25	0	258719.328	702140.923	5878
Lmic	Lnil	Lmar	Lcal	-0.4319	-0.4319	0.000003	-262.79	0	286149.783	721216.779	5878
Lang	Lnil	Lmic	Lcal	-0.4876	-0.4876	0.000003	-278.25	0	254073.084	737583.101	5878
Lang	Lnil	Lmar	Lcal	-0.5180	-0.5180	0.000003	-308.56	0	242830.608	764835.701	5878

10 of 33

Table S2. D-statistics for the five-taxon *Lates* phylogeny, using *L. calcarifer* as the outgroup, with D-statistic comparisons constrained to the species tree topology from `RAxML` (i.e., comparisons were arrange so that ((P1,P2),P3) arrangements match the species tree). Values were calculated using `Dsuite` (23) using the whole genome sequencing data. Columns indicate the three taxa in the comparison (P1, P2, P3), the D-statistic (D), the z-score associated with D (Z-score), the p-value associated with the z-score (p-value), the estimated admixture fraction (f_4 ratio), the number of concordant BBAA sites (BBAA), the number of ABBA sites (ABBA), and the number of BABA sites (BABA).

P1	P2	P3	D	Z-score	p-value	f4-ratio	BBAA	ABBA	BABA
Lmar	Lang	Lmic	0.0950	42.45	0	0.0808	243260	227060	187677
Lmar	Lang	Lnil	0.129	47.55	0	0.0249	560392	176052	135761
Lmar	Lang	Lsta	0.0690	29.17	0	0.0295	262327	212312	184907
Lmic	Lang	Lnil	0.0585	25.73	0	0.0112	528634	161172	143358
Lsta	Lang	Lmic	0.0194	3.08	0.00103	0.0195	196767	230648	221878
Lsta	Lang	Lnil	0.0834	22.56	0	0.0167	503696	172850	146242
Lmar	Lmic	Lnil	0.0674	32.99	0	0.0139	524763	177672	155231
Lmar	Lsta	Lmic	0.0630	11.53	0	0.0623	200451	253028	223046
Lmar	Lsta	Lnil	0.0374	11.26	0	0.00812	512025	180738	167717
Lsta	Lmic	Lnil	0.0282	9.89	0	0.00579	540725	167472	158282

Table S3. Summary of significant f_d windows for each trio tested in `Dsuite Dinvestigate`. Windows were defined as significiant when the p-value from their z-score was <0.01, and adjacent signficant windows combined into continuous blocks. For each trio and each chromosome, we calculated the average length of these continuous blocks, the total length of the blocks, and the average f_d among significant windows. Trio names match those in Fig. S4.

Trio	Chromosome	Mean window length (bp)	Total window length (bp)	Mean f_d
Lmar_Lang_Lnil (P1 = *L. mariae*, P2 = *L. angustifrons*, P3 = *L. niloticus*)				
	ASB_LG1	12748.000	1682736	0.261
	ASB_LG10	7722.157	540551	0.209
	ASB_LG11	7930.984	491721	0.224
	ASB_LG12	12146.246	692336	0.207
	ASB_LG13	7090.755	723257	0.252
	ASB_LG14	12390.291	978833	0.253
	ASB_LG15	7026.598	857245	0.211
	ASB_LG16_LG22	7154.357	400644	0.219
	ASB_LG17	10170.188	650892	0.225
	ASB_LG18	10919.348	251145	0.215
	ASB_LG19	7424.562	541993	0.191
	ASB_LG2	18110.000	289760	0.207
	ASB_LG20	10092.952	211952	0.201
	ASB_LG21	9483.885	739743	0.196
	ASB_LG23	15691.684	298142	0.224
	ASB_LG24	9147.871	283584	0.191
	ASB_LG3	6867.301	982024	0.211
	ASB_LG4	7615.314	898607	0.212
	ASB_LG5	10532.303	937375	0.238
	ASB_LG6	9424.784	914204	0.202
	ASB_LG7_1	8141.926	439664	0.219
	ASB_LG7_2	17014.095	357296	0.193
	ASB_LG8	6044.024	513742	0.305
	ASB_LG9	7218.311	324824	0.203
Lmar_Lmic_Lnil (P1 = *L. mariae*, P2 = *L. microlepis*, P3 = *L. niloticus*)				
	ASB_LG1	10602.056	954185	0.273
	ASB_LG10	6964.704	376094	0.206
	ASB_LG11	9320.593	549915	0.243
	ASB_LG12	8272.946	463285	0.207
	ASB_LG13	6395.967	390154	0.236
	ASB_LG14	9823.655	540301	0.218
	ASB_LG15	6008.804	306449	0.190
	ASB_LG16_LG22	7155.405	264750	0.205
	ASB_LG17	7712.636	169678	0.266
	ASB_LG18	16545.636	182002	0.193
	ASB_LG19	5652.641	220453	0.208
	ASB_LG2	10385.615	135013	0.242
	ASB_LG20	7670.600	76706	0.202
	ASB_LG21	7328.377	447031	0.211
	ASB_LG23	9679.083	116149	0.194
	ASB_LG24	8015.760	200394	0.177
	ASB_LG3	7173.289	695809	0.232
	ASB_LG4	7968.444	502012	0.209
	ASB_LG5	11360.357	636180	0.215
	ASB_LG6	8557.483	513449	0.219
	ASB_LG7_1	10637.973	393605	0.230
	ASB_LG7_2	5526.294	93947	0.180
	ASB_LG8	7751.783	465107	0.205
	ASB_LG9	9067.205	398957	0.187
Lmar_Lsta_Lnil (P1 = *L. mariae*, P2 = *L. stappersii*, P3 = *L. niloticus*)				
	ASB_LG1	8989.052	692157	0.203
	ASB_LG10	6683.566	354229	0.212
	ASB_LG11	10366.632	393932	0.262
	ASB_LG12	10304.877	587378	0.204
	ASB_LG13	7115.220	419798	0.208
	ASB_LG14	9392.660	441455	0.248
	ASB_LG15	7016.562	624474	0.210
	ASB_LG16_LG22	9282.711	417722	0.220

12 of 33

ASB_LG17	12800.300	384009	0.248
ASB_LG18	18495.826	425404	0.248
ASB_LG19	6516.980	319332	0.207
ASB_LG2	13693.550	273871	0.211
ASB_LG20	12550.429	175706	0.184
ASB_LG21	8905.571	498712	0.203
ASB_LG23	14813.077	385140	0.285
ASB_LG24	6694.294	113803	0.173
ASB_LG3	6466.747	588474	0.222
ASB_LG4	7155.568	264756	0.193
ASB_LG5	11819.015	803693	0.233
ASB_LG6	7739.574	526291	0.222
ASB_LG7_1	10062.080	251552	0.255
ASB_LG7_2	7708.556	69377	0.180
ASB_LG8	7750.173	403009	0.247
ASB_LG9	7121.727	235017	0.250
Lmic_Lang_Lnil (P1 = *L. microlepis*, P2 = *L. angustifrons*, P3 = *L. niloticus*)			
ASB_LG1	13587.135	706531	0.220
ASB_LG10	10497.972	377927	0.216
ASB_LG11	10118.737	576768	0.245
ASB_LG12	10750.345	311760	0.205
ASB_LG13	6090.019	322771	0.212
ASB_LG14	10959.124	975362	0.264
ASB_LG15	8398.627	697086	0.219
ASB_LG16_LG22	10392.818	457284	0.269
ASB_LG17	9674.847	570816	0.268
ASB_LG18	10745.905	225664	0.204
ASB_LG19	5254.278	189154	0.184
ASB_LG2	23765.706	404017	0.198
ASB_LG20	9592.500	115110	0.311
ASB_LG21	7245.174	166639	0.199
ASB_LG23	18618.000	148944	0.231
ASB_LG24	9019.600	270588	0.188
ASB_LG3	5166.691	418502	0.198
ASB_LG4	7991.962	639357	0.212
ASB_LG5	10912.235	742032	0.238
ASB_LG6	14786.921	561903	0.216
ASB_LG7_1	6866.111	247180	0.212
ASB_LG7_2	17425.962	453075	0.197
ASB_LG8	6142.012	491361	0.341
ASB_LG9	7114.094	227651	0.206
Lsta_Lang_Lnil (P1 = *L. stappersii*, P2 = *L. angustifrons*, P3 = *L. niloticus*)			
ASB_LG1	13539.048	1421600	0.250
ASB_LG10	8609.897	335786	0.213
ASB_LG11	8406.366	781792	0.254
ASB_LG12	10625.920	265648	0.233
ASB_LG13	8409.646	832555	0.263
ASB_LG14	10966.094	1162406	0.290
ASB_LG15	7405.429	673894	0.206
ASB_LG16_LG22	8747.685	472375	0.247
ASB_LG17	11066.682	940668	0.257
ASB_LG18	13284.692	172701	0.213
ASB_LG19	6561.962	341222	0.207
ASB_LG2	17733.950	354679	0.203
ASB_LG20	8746.960	218674	0.241
ASB_LG21	13030.607	794867	0.198
ASB_LG23	28962.909	318592	0.201
ASB_LG24	10953.455	481952	0.225
ASB_LG3	8400.810	1327328	0.221
ASB_LG4	9315.488	1490478	0.242
ASB_LG5	13913.612	1363534	0.253
ASB_LG6	8216.816	402624	0.211
ASB_LG7_1	9382.482	525419	0.221
ASB_LG7_2	15838.091	348438	0.214
ASB_LG8	7151.209	650760	0.324

ASB_LG9	12263.800	981104	0.226
Lsta_Lmic_Lnil (P1 = *L. stappersii*, P2 = *L. microlepis*, P3 = *L. niloticus*)			
ASB_LG1	11223.522	751976	0.273
ASB_LG10	8833.462	229670	0.223
ASB_LG11	7509.710	698403	0.254
ASB_LG12	7368.783	169482	0.215
ASB_LG13	8667.307	650048	0.257
ASB_LG14	9350.294	635820	0.277
ASB_LG15	8157.568	358933	0.200
ASB_LG16_LG22	7046.250	253665	0.224
ASB_LG17	8143.488	333883	0.224
ASB_LG18	2635.667	7907	0.205
ASB_LG19	6040.216	308051	0.220
ASB_LG2	9253.462	120295	0.251
ASB_LG20	11048.375	176774	0.192
ASB_LG21	12483.489	561757	0.208
ASB_LG23	32487.000	64974	0.180
ASB_LG24	9804.129	303928	0.206
ASB_LG3	11847.350	1184735	0.232
ASB_LG4	9642.375	848529	0.224
ASB_LG5	13517.032	838056	0.227
ASB_LG6	6964.238	146249	0.234
ASB_LG7_1	9588.583	345189	0.216
ASB_LG7_2	10237.727	225230	0.186
ASB_LG8	7367.000	360983	0.231
ASB_LG9	11455.852	618616	0.200

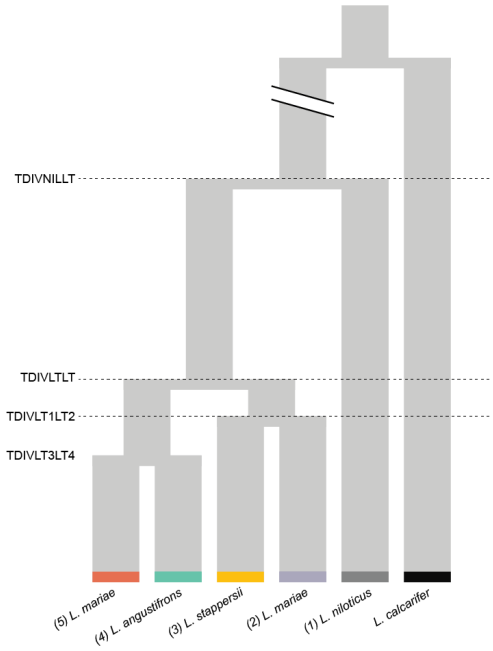
14 of 33

Table S4. Average divergence time estimates from cross-coalescence rates estimated using `MSMC2`. Reported divergence times are the average left and right time boundaries for the segment where the relative cross-coalescence rate for pairwise comparisons between individuals of the given species reached 50%. Estimates have been converted to years using a mutation rate of $3.5x10^{-}9$ and generation time of three years.

Species Pair	Left Time Boundary (yr)	Right Time Boundary (yr)
LT-LT	**901,772**	**1,076,889**
Lang-Lmar	636,979	757,300
Lang-Lmic	641,040	762,628
Lmic-Lmar	707,685	841,615
Lang-Lsta	1,122,649	1,334,336
Lmic-Lsta	1,146,387	1,391,770
Lmar-Lsta	1,155,890	1,373,685
Lnil-LT	**1,108,439**	**1,337,488**
Lnil-Lang	978,686	1,200,767
Lnil-Lmar	988,401	1,210,916
Lnil-Lmic	1,080,626	1,287,231
Lnil-Lsta	1,386,043	1,651,037

Table S5. Parameter estimates from nonparametric bootstrapping of the best model in `fastsimcoal2`, with parameters as indicated on model diagram to the right. Estimates are generated from the maximum likelihood estimates from 200 bootstrap datasets. Estimates of times are in units of generations, not years, and population sizes are in units of $2N$. The average generation time for this group of taxa is three years. Migration rate parameters are labeled with the two populations involved (i.e., MIG12 is admixture between (1) *L. niloticus* and (2) *L. mariae*, and MIG21 is the migration rate in the opposite direction). Timing for admixture events matches the same labeling pattern (i.e., TADM12 is the timing of admixture events MIG12 and MIG21).

Parameter	Mean	Median	2.5% Quantile	97.5% Quantile
Migration Rates				
MIG12	8.240000e-04	1.850000e-05	1.000000e-07	6.594000e-03
MIG14	3.617000e-04	1.810000e-05	1.000000e-07	1.911600e-03
MIG15	6.928000e-04	1.450000e-05	1.000000e-07	4.555400e-03
MIG21	1.713380e-02	1.206000e-04	1.000000e-07	1.177663e-01
MIG23	1.325300e-03	8.140000e-05	2.000000e-07	1.153420e-02
MIG25	1.474700e-03	2.120000e-05	0.000000e+00	1.547330e-02
MIG32	7.788000e-04	4.330000e-05	1.000000e-07	6.560200e-03
MIG34	1.217200e-03	1.870000e-05	1.000000e-07	8.497000e-03
MIG41	5.438450e-02	2.569500e-02	1.000000e-07	1.444328e-01
MIG43	9.051000e-04	7.570000e-05	1.000000e-07	8.156200e-03
MIG51	4.599470e-02	6.652300e-03	4.000000e-07	1.391530e-01
MIG52	9.979000e-04	3.680000e-05	1.000000e-07	7.642800e-03
Effective Population Sizes				
POP0	3.137472e+05	3.117270e+05	2.131975e+05	4.143505e+05
POP1	9.084603e+04	8.352500e+04	4.543850e+04	1.700480e+05
POP2	2.851361e+05	2.850700e+05	2.066840e+05	3.595100e+05
POP3	3.340721e+05	3.270370e+05	2.453715e+05	4.589210e+05
POP4	2.540657e+05	2.523180e+05	1.405875e+05	3.713435e+05
POP5	3.364123e+05	3.204830e+05	2.086800e+05	5.199625e+05
Admixture Events (generations ago)				
TADM12	9.104819e+04	8.901600e+04	6.654500e+03	1.870345e+05
TADM13	1.078896e+06	1.083528e+06	3.071305e+05	1.709435e+06
TADM14	4.646818e+04	3.167300e+04	2.331000e+03	1.205125e+05
TADM15	5.638590e+04	4.407900e+04	4.141500e+03	1.444000e+05
TADM23	8.563990e+04	8.441200e+04	2.111250e+04	1.502070e+05
TADM25	8.713465e+04	8.628600e+04	1.895400e+04	1.524090e+05
Divergence Events (generations ago)				
TADM34	8.035712e+04	8.208100e+04	1.492550e+04	1.564315e+05
TDIVLT1LT2	1.943398e+05	1.899580e+05	1.532045e+05	2.451165e+05
TDIVLT3LT4	1.468291e+05	1.449590e+05	8.805700e+04	1.924345e+05
TDIVLTLT	2.309192e+05	2.321800e+05	1.819790e+05	2.984175e+05
TDIVNILLT	4.603517e+05	4.615850e+05	3.248185e+05	6.015390e+05



16 of 33

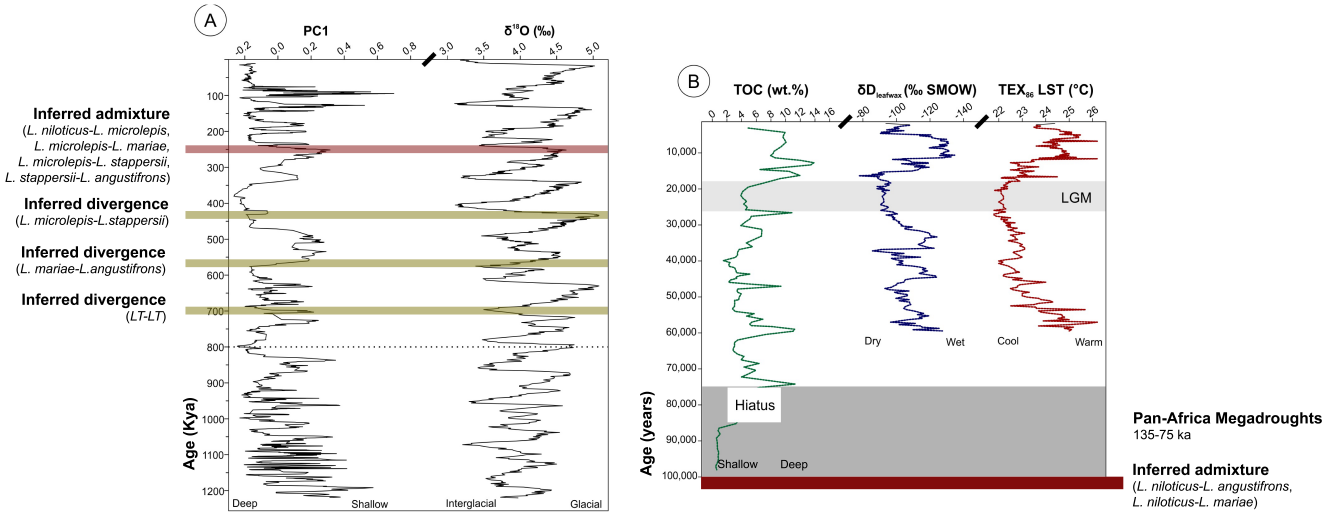


Fig. S1. Sedimentary records of Quaternary environmental change for eastern Africa. (A) Lake level reconstruction derived from paleoecological data (pollen, ostracodes, mollusks, charcoal) from central Lake Malawi (43). PC, Principal component. Amplitude and frequency of inferred lake level oscillations shift at approximately 800 Kya (dotted line), near the same time that global benthic oxygen isotope data suggest that glacial-interglacial cycles transitioned from a 41 kyr to a 100 kyr periodicity (e.g., 44). (B) Climate and lake level proxy reconstructions from deepwater sediments from Lake Tanganyika. Water levels declined several hundered meters during the Last Glacial Maximum (LGM; 45) and the Pan-African Megadrought intervals (∼135-75 ka; 46); both episodes (grey shading) generated hiatuses observable on seismic profiles, owing to subaerial exposure of the lake floor. These regressions lead to persistent hydrological closure of Lake Tanganyika, disconnecting the lake from the Congo Basin as water level elevation dropped below the elevation at the Lukuga River outlet on the western margin (Fig. 1).

Fig. S2. Diagrams of models tested in `fastsimcoal2` analyses. All topologies were constrained to the `RAxML` species tree topology, with the split between *L. calcarifer* and the freshwater *Lates* constrained using the oldest known *Lates* fossil found in Africa (*L. qatraniensis* from the Jebel Qatrani Formation of the Fayum Depression in Egypt, ∼33 MYA; Murray & Attia 2004). Lines indicate admixture events parameterized in each model. All model specification files can be found at http://github.com/jessicarick/lates-fastsimcoal-scripts.

18 of 33

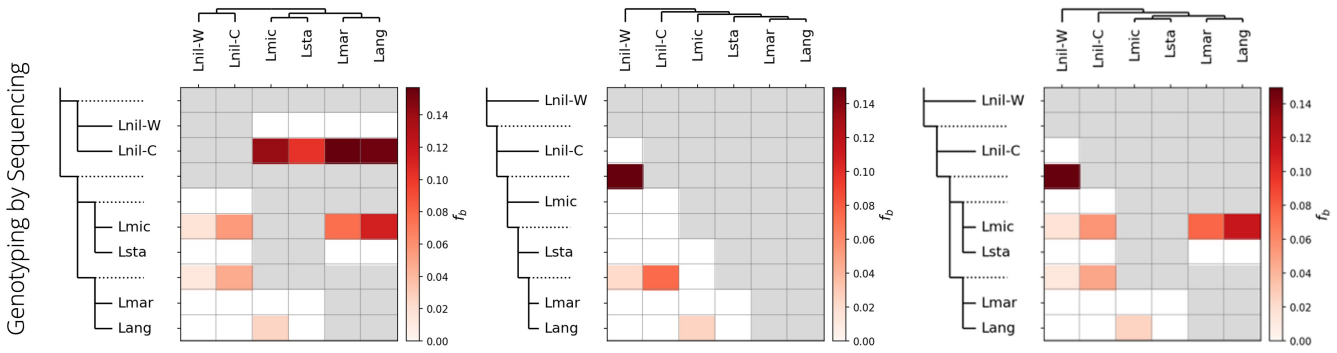


Fig. S3. A comparison of Fbranch results for the GBS dataset, constrained to three alternate phylogenetic hypotheses for the relationship among species: (a) *L. niloticus* as a monophyletic group, with two sister groups within Lake Tanganyika; (b) a paraphyletic *L. niloticus*, with pectinate toplogy within Lake Tanganyika (sensu 25); and (c) a paraphyletic *L. niloticus*, with two sister groups within Lake Tanganyika (as seen in our GBS phylogeny, Fig 1). Abbreviations: Lang, *Lates angustifrons*; Lmar, *L. mariae*; Lmic, *L. microlepis*; Lsta, *L. stappersii*; Lnil-W, *L. niloticus* from West Africa; Lnil-C, *L. niloticus* from the Congo River basin.

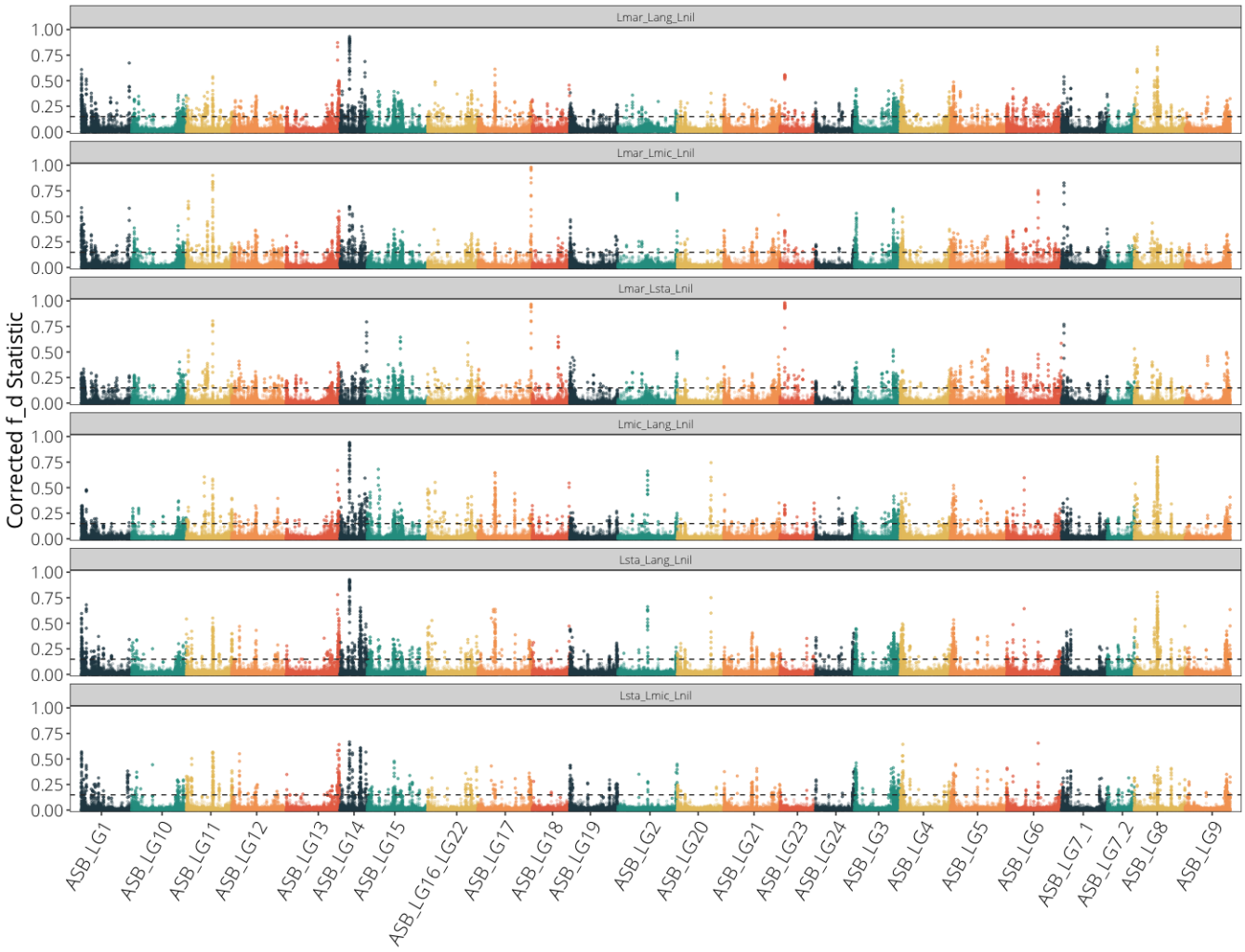


Fig. S4. Signals of admixture (f_d) in 100bp windows across the genome, as calculated in `Dsuite` (23), for each trio comprised of two Lake Tanganyika *Lates* species and *L. niloticus*. Dashed lines indicate $z_{fd} = 2.33$, corresponding to $p = 0.01$, which was used as the threshold for significantly elevated admixture signal. Values of f_d for windows with $D < 0$ have been corrected to 0, as recommended.

20 of 33

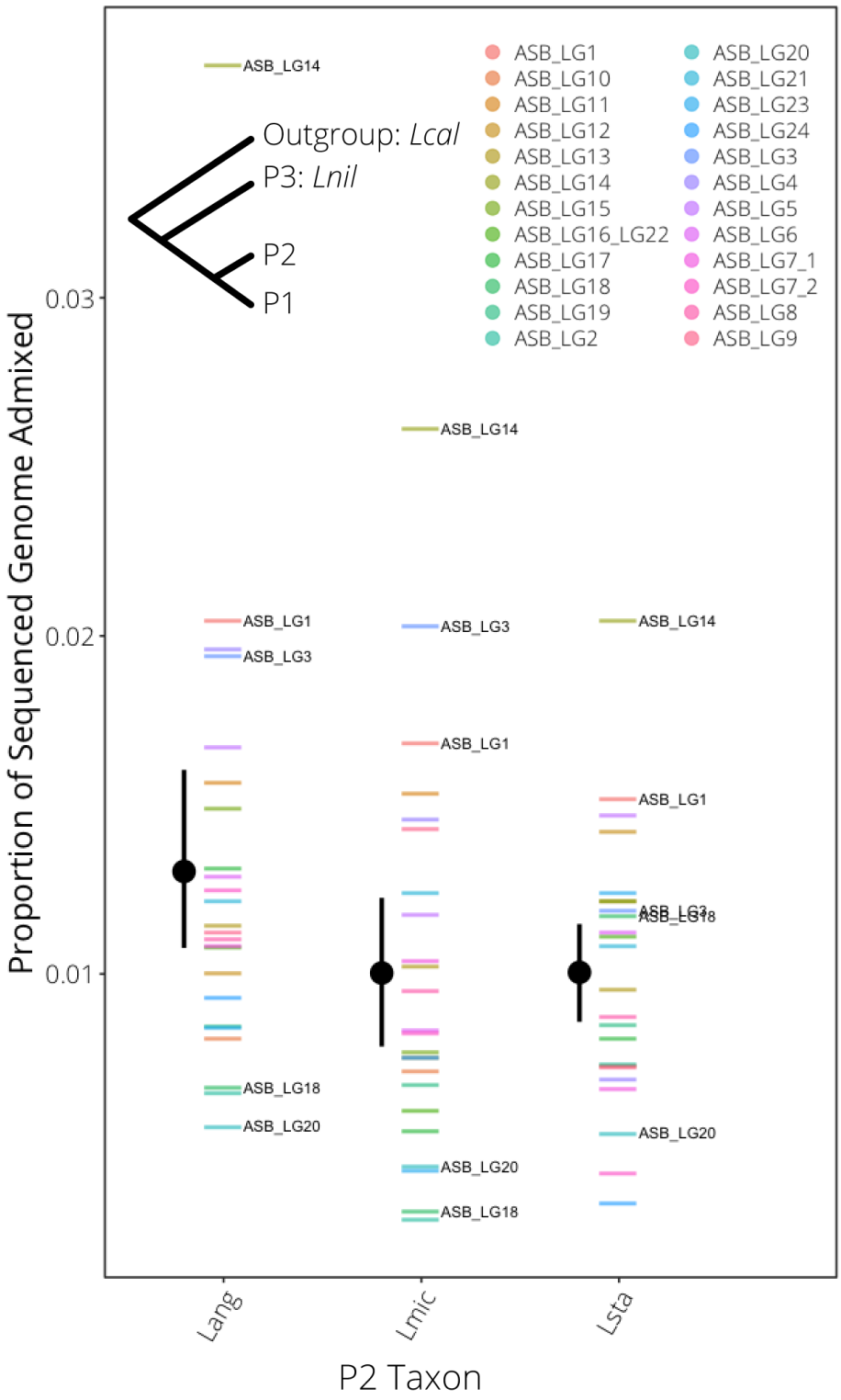


Fig. S5. Proportion of genome with an elevated signal of admixture ($z_{fd} > 2.33$, corresponding to $p < 0.01$) for each choice of P2 taxon, averaged across LT species as P1 taxa. Lines show the proportion for each chromosome, while black points and bars indicate the mean and 95% confidence interval for the overall genome-wide proportion for the given P2 taxon. To calculate the proportion, windows were combined into blocks if adjacent windows had elevated f_d signals, and the sum total length of significant blocks for a chromosome was divided by the total sequenced length of the chromosome. Estimates of (f_d) were calculated in 100bp windows across the genome in `Dsuite` (23), for each trio comprised of two Lake Tanganyika *Lates* species and *L. niloticus* that showed an elevated D-statistic in genome-wide analyses. Values of f_d for windows with $D < 0$ have been corrected to 0, as recommended. A few of the chromosomes with highest and lowest proportions have been annotated.

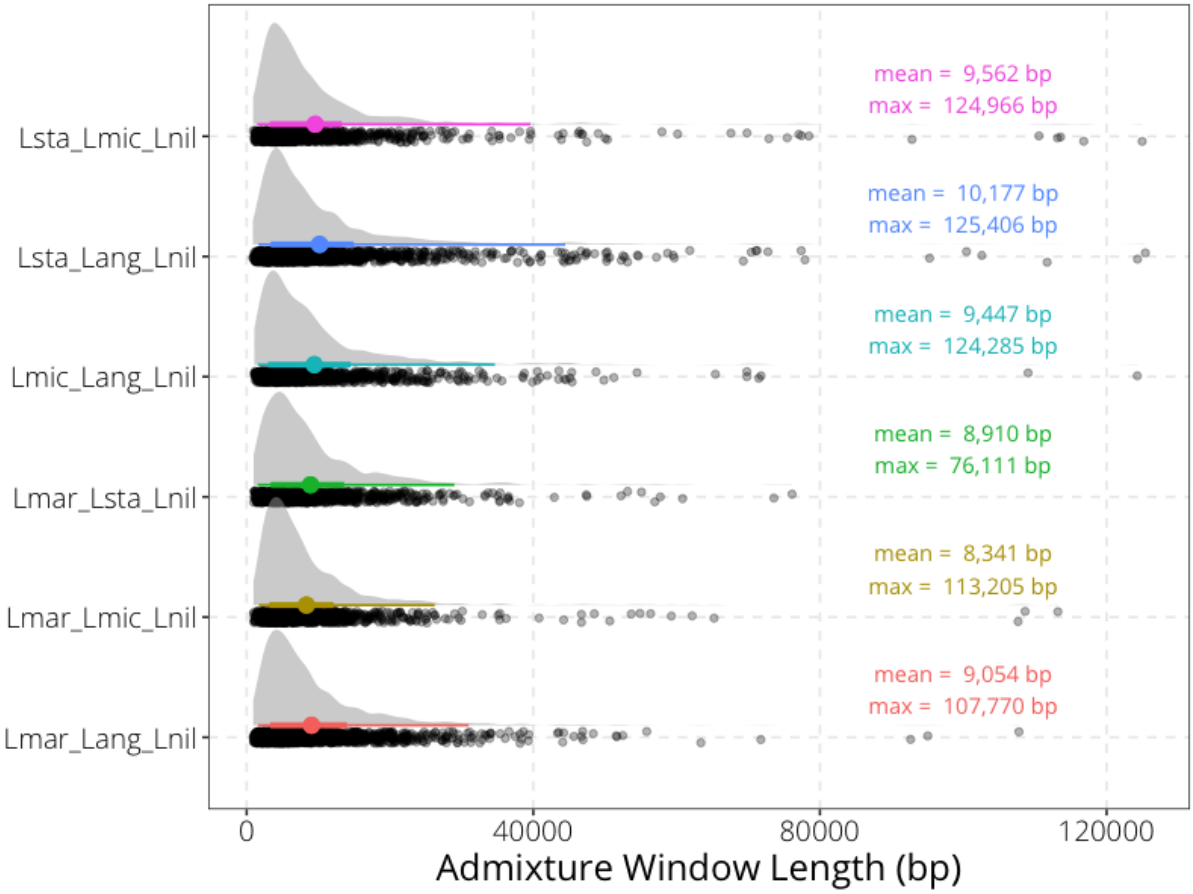


Fig. S6. Distribution of block sizes for windows with a significantly elevated signal of admixture ($z_{f_d} > 2.33$, corresponding to $p < 0.01$) for each three-taxon comparisons. Points show the overall median block size, while bars indicate the 2.5 and 97.5% quantiles. To calculate block sizes, windows were combined into blocks if adjacent windows had elevated f_d signals. Estimates of (f_d) were calculated in 100bp windows across the genome in `Dsuite` (23), for each trio comprised of two Lake Tanganyika *Lates* species and *L. niloticus*. Values of f_d for windows with $D < 0$ have been corrected to 0, as recommended. Abbreviations: Lang, *Lates angustifrons*; Lmar, *L. mariae*; Lmic, *L. microlepis*; Lsta, *L. stappersii*; Lnil, *L. niloticus*.

22 of 33

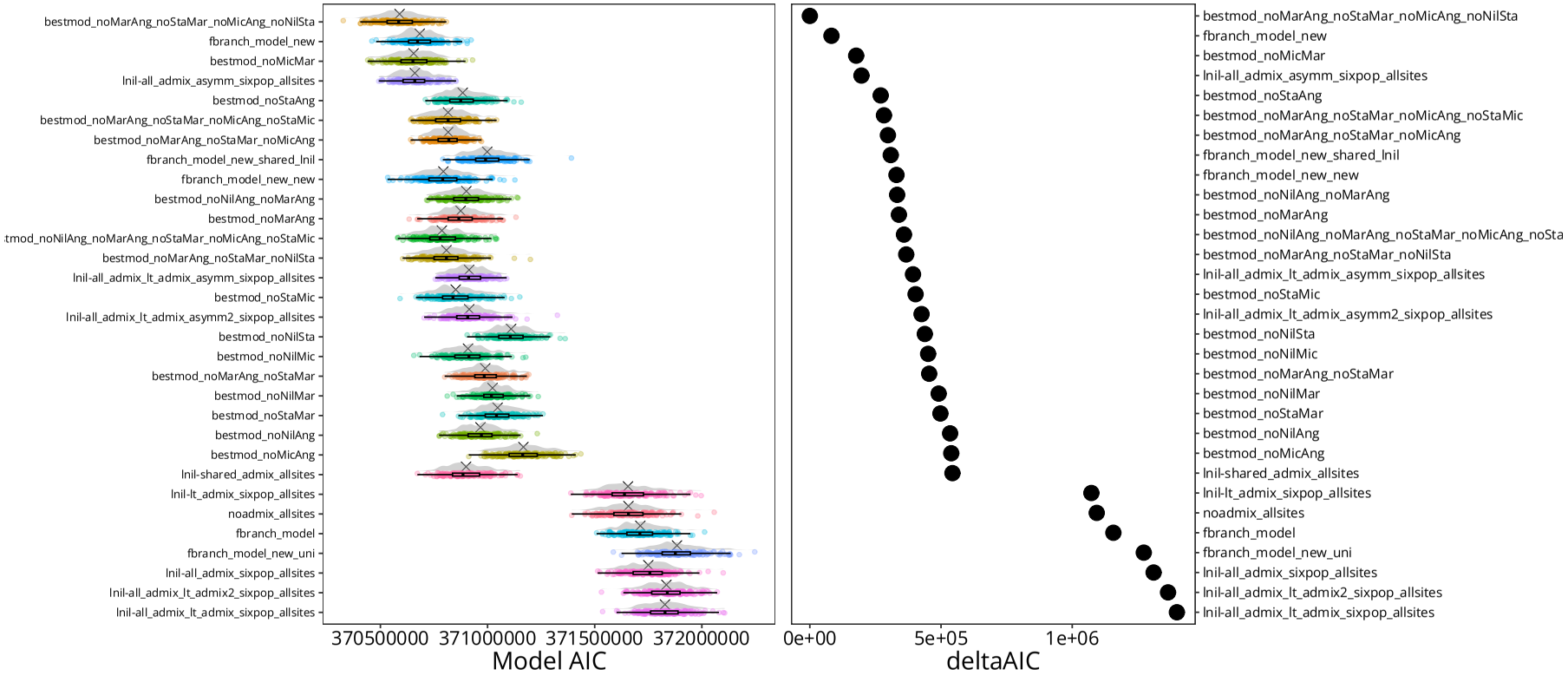


Fig. S7. A comparison of model likelihoods (left) and AIC (right) for the six-taxon models tested. Model parameters are detailed in Table S5. Models are ordered based on AIC estimates; the best models were chosen based on lowest AIC and highest likelihood distributions. For AIC calculations, parameters were penalized by 200. The vertical dashed line indicates the lower 95% quantile for the top model's distribution of likelihood values. Note that model numbers match the diagram in Fig. S2).

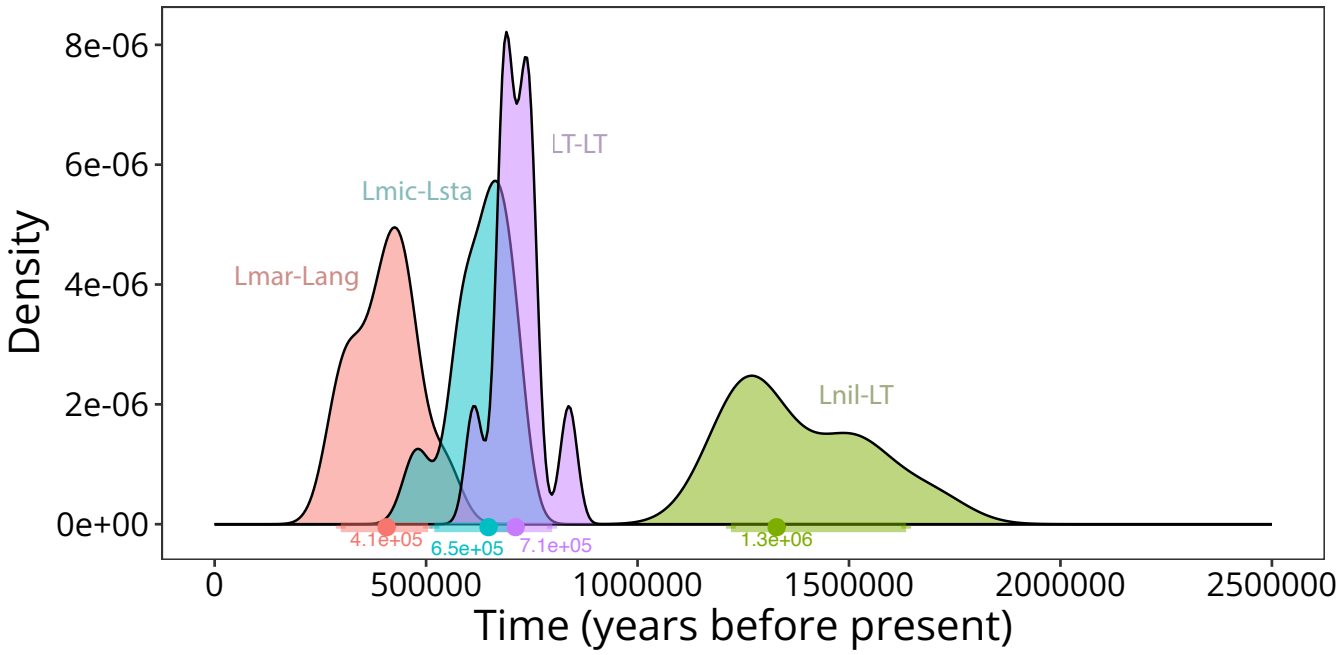


Fig. S8. A comparison of divergence time estimates from the top ten best `fastsimcoal2` models, demonstrating that our choice of top model does not significantly bias our divergence time results. The distribution of divergence estimates for the *L. niloticus*-LT clade split has a median estimate of 1.3 Mya (range 1.2-1.7 Mya), for the LT-LT split the median estimate is 712 Kya (range 613-837 Kya), the median estimate for the *L. mariae-L. angustifrons* divergence time is 647 Kya (range 478-714 Kya), and the median for the *L. microlepis-L. stappersii* split is 407 Kya (range 287-542 Kya).

24 of 33

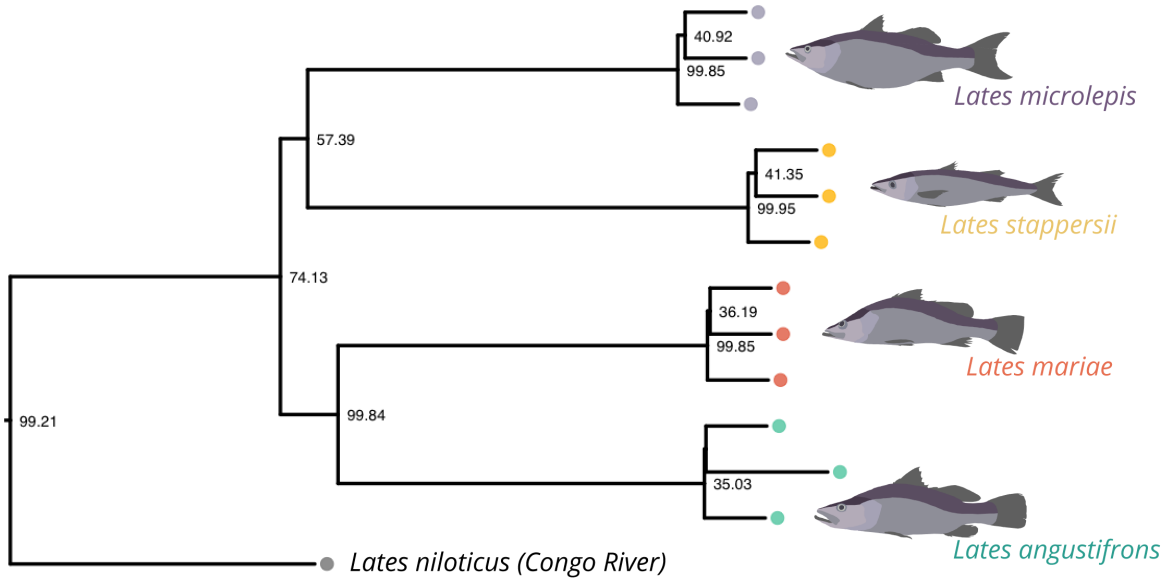


Fig. S9. Phylogeny inferred from quartet-based multispecies coalescent species tree reconstruction in `ASTRAL-III` using whole genome data. Node annotations indicate normalized quartet scores (i.e., the percentage of quartets in the 4,000bp gene trees that agree with a split). The tree has been rooted using *L. calcarifer* as an outgroup.

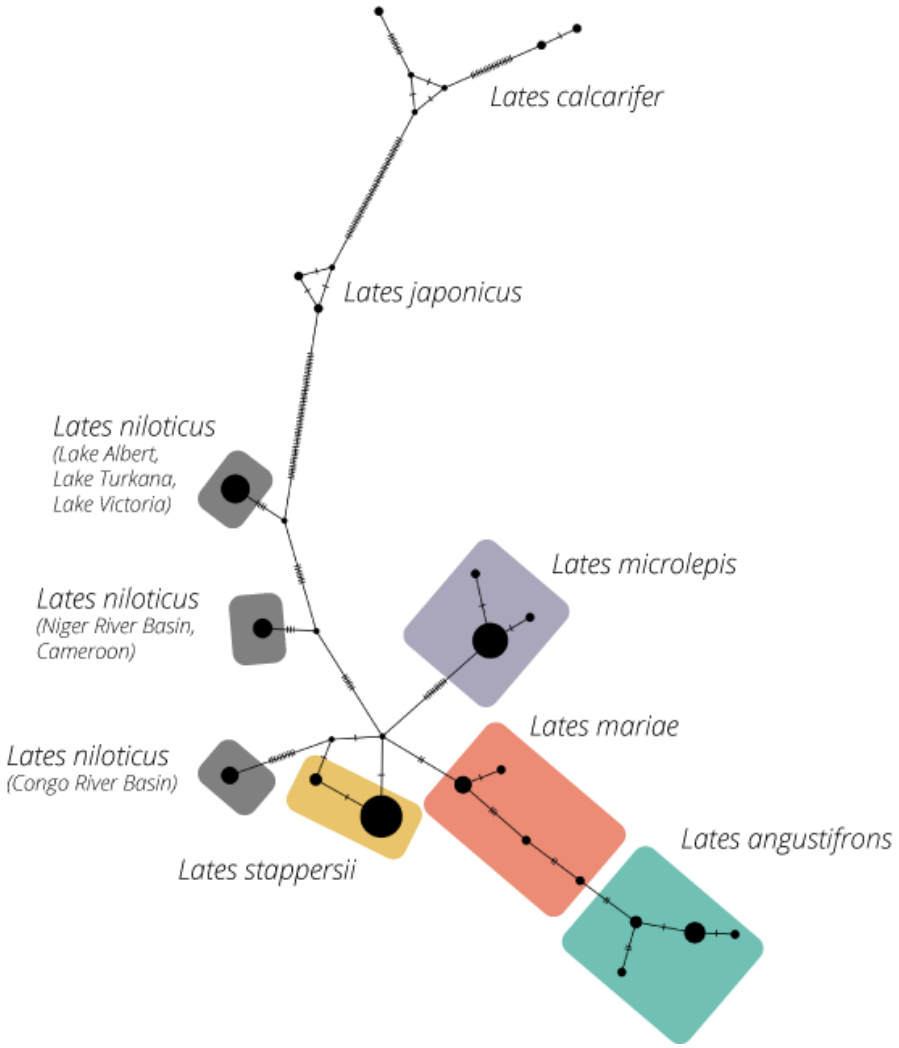


Fig. S10. Median-joining haplotype network inferred from n=87 individual *Lates* fishes' 16s mitochondrial sequencing data generated in this paper and combined with sequences from (9, 25). Each circle represents a distinct haplotype, with tick marks indicating mutations between haplotypes and circle size representing the number of individuals in the data set with the given haplotype.

26 of 33

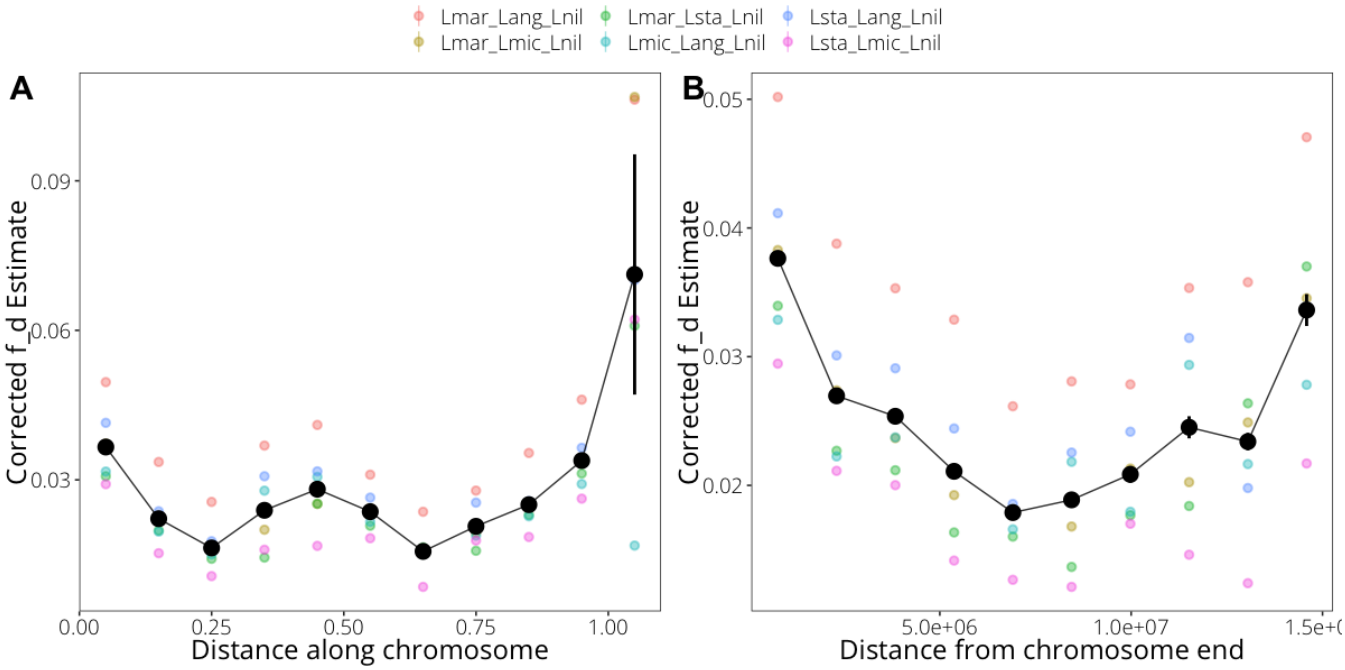


Fig. S11. Summary of average f_d estimates (A) by position along the chromosome, and (B) by distance to the end of the chromosome. Black points show overall means, and error bars represent standard deviation of estimates. Colored points show means for each trio used in comparisons.

Fig. S12. Distribution of parameter estimates from nonparametric bootstrapping of 200 datasets in `fastsimcoal2` for our best model, with each parameter estimate plotted against the difference in observed and expected likelihoods for its containing model. Dashed lines indicate the 95% bootstrap confidence intervals for each parameter. Details of summary statistics for each parameter can be found in Table S5. Solid lines indicate median of the parameter distributions, which are the estimates plotted in Fig. 2.

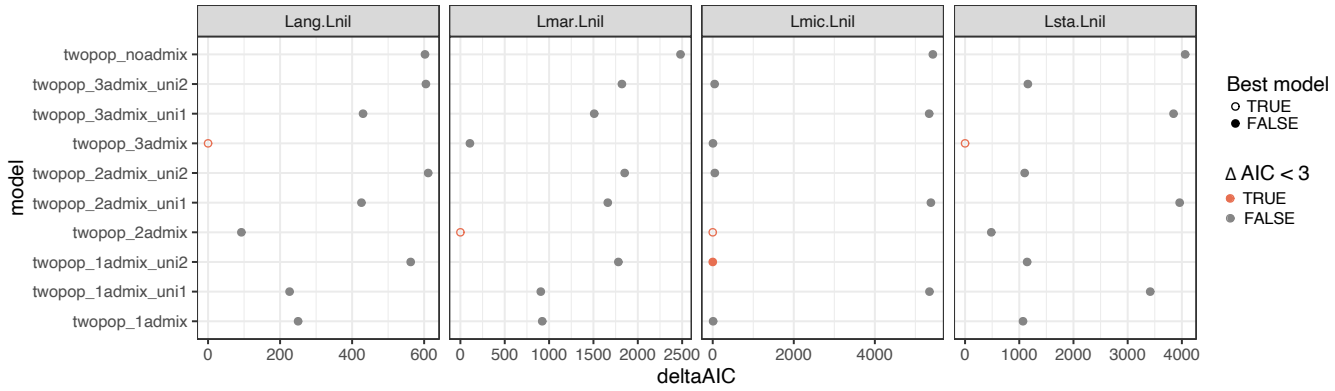


Fig. S13. A comparison of AIC values for the two-taxon `fastsimcoal2` models tested for comparisons between *L. niloticus* (Lnil) and each of the Lake Tanganyika taxa (Lang, *L. angustifrons*; Lmar, *L. mariae*; Lmic, *L. microlepis*; Lsta, *L. stappersii*). The best models were chosen based on lowest AIC and highest likelihood values. For AIC calculations, parameters were penalized by 25. Models in red have an AIC value within 3 points of the top model's AIC. For each species pair, we tested models with no admixture ("noadmix"), a single admixture event ("1admix"), two admixture events ("2admix"), or three admixture events ("3admix"), and within each model involving admixture, we also tested models with gene flow occurring unidirectionally in one direction or the other ("uni1" and "uni2").

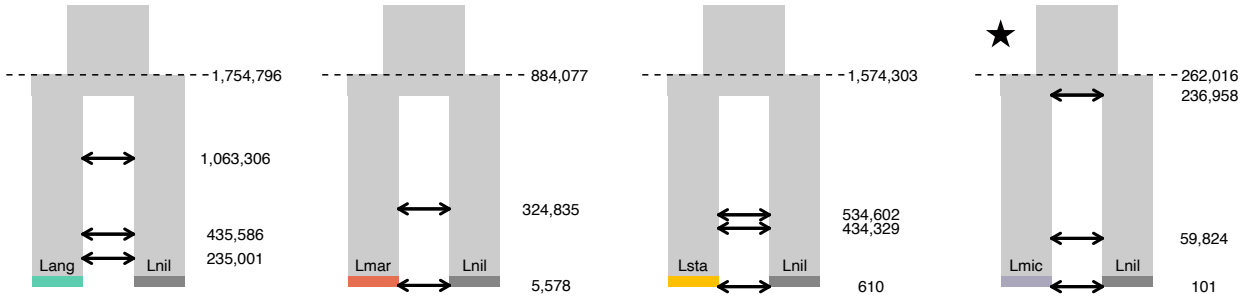


Fig. S14. Visual representations of the inferred best models for each two taxon `fastsimcoal2` comparison involving *L. niloticus* and the LT *Lates* species, including parameters inferred for the given model. Stars indicate species pairs for which there was at least one alternative model where $\Delta AIC < 3$ (Fig. S13). Divergence times are indicated for the split between the two taxa, as well as inferred timings of admixture events.

30 of 33

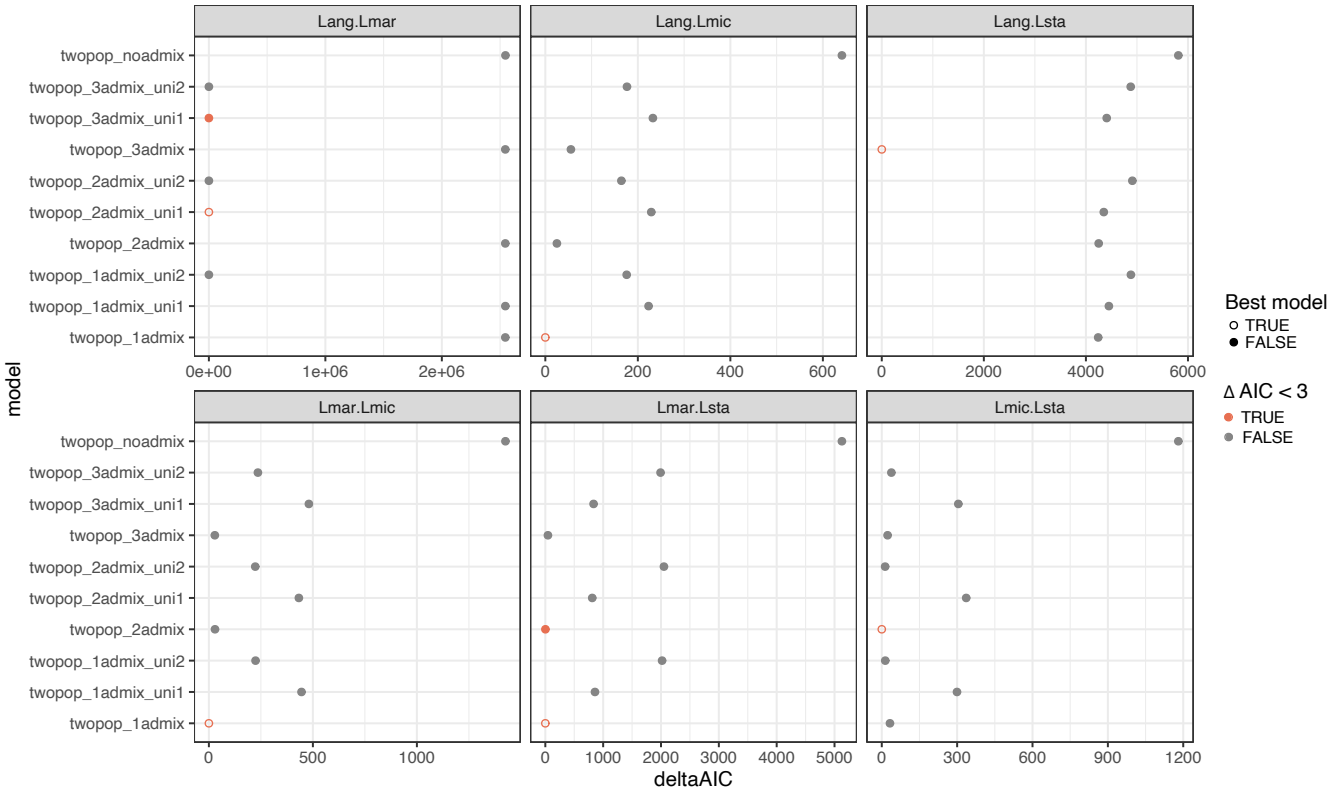


Fig. S15. A comparison of model likelihoods AIC values for the two-taxon `fastsimcoal2` models tested for comparisons within Lake Tanganyika taxa (Lang, *L. angustifrons*; Lmar, *L. mariae*; Lmic, *L. microlepis*; Lsta, *L. stappersii*). The best models were chosen based on lowest AIC and highest likelihood values. For AIC calculations, parameters were penalized by 25; models with a delta-AIC value less than 3 are indicated in red. For each species pair, we tested models with no admixture ("noadmix"), a single admixture event ("1admix"), two admixture events ("2admix"), or three admixture events ("3admix"), and within each model involving admixture, we also tested models with gene flow occurring unidirectionally in one direction or the other ("uni1" and "uni2").

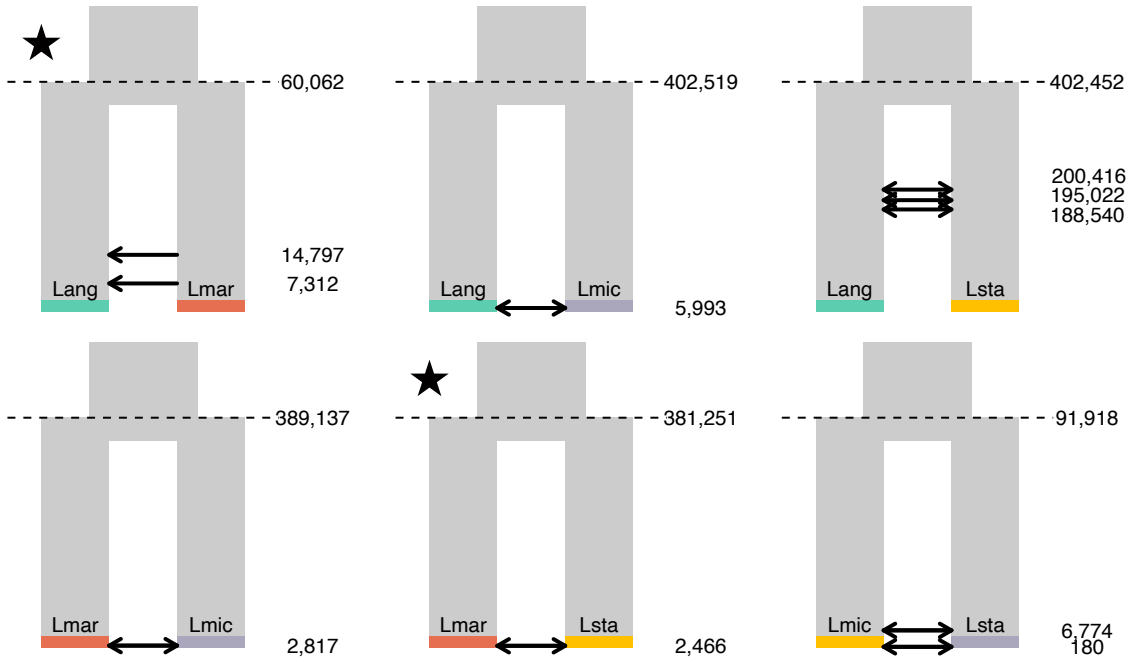


Fig. S16. Visual representations of the inferred best models for each two taxon `fastsimcoal2` comparison between LT *Lates* taxa, including parameters inferred for the given model. Stars indicate species pairs for which there was at least one alternative model where $\Delta AIC < 3$ (Fig. S15). Divergence times are indicated for the split between the two taxa, as well as inferred timings of admixture events.

32 of 33

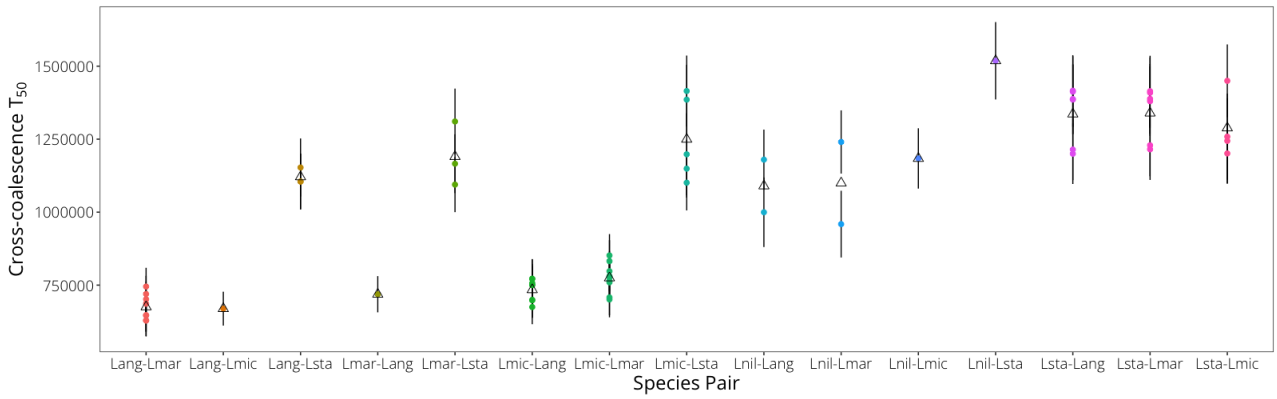


Fig. S17. Divergence time estimates from relative cross-coalescence rates inferred using `MSMC2`, with individual sequence pairs categorized by species pair. Divergence times were estimated as the time (in years) at which the relative cross-coalescence rate reached 0.50. Each colored point indicates the estimate for a single pair of individuals, and black triangles show the mean for pairs in the given species comparison. Abbreviations: Lang, *Lates angustifrons*; Lmar, *L. mariae*; Lmic, *L. microlepis*; Lsta, *L. stappersii*; Lnil, *L. niloticus*.