Drivers and dynamics of a massive adaptive radiation in cichlid fishes

Fabrizia Ronco[1*], Michael Matschiner[1,2,3], Astrid Böhne[1,4], Anna Boila[1], Heinz H. Büscher[1], Athimed El Taher[1], Adrian Indermaur[1], Milan Malinsky[1], Virginie Ricci[1], Ansgar Kahmen[5], Sissel Jentoft[3] and Walter Salzburger[1,3*]

[1] Zoological Institute, Department of Environmental Sciences, University of Basel, Basel, Switzerland

[2] Palaeontological Institute and Museum, University of Zurich, Zurich, Switzerland

[3] Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway

[4] Centre for Molecular Biodiversity Research (ZMB), Zoological Research Museum Alexander Koenig, Bonn, Germany

[5] Botany, Department of Environmental Sciences, University of Basel, Basel, Switzerland

e-mail: fabrizia.ronco@unibas.ch, walter.salzburger@unibas.ch

Adaptive radiation is the likely source of much of the ecological and morphological diversity of life[1–4]. How adaptive radiations proceed and what determines their extent remains elusive in most cases[1,4]. Here we report the in-depth examination of the spectacular adaptive radiation of cichlid fishes in African Lake Tanganyika. Based on whole-genome phylogenetic analyses, multivariate morphological measurements of three ecologically relevant trait complexes (body shape, upper oral jaw morphology, and lower pharyngeal jaw shape), scoring of pigmentation patterns, and approximations of the ecology of virtually all ~240 cichlid species endemic to Lake Tanganyika, we show that the radiation occurred within the confines of the lake and that morphological diversification proceeded in consecutive trait-specific pulses of morphospace expansion. We provide empirical support for two theoretical predictions on how adaptive radiations proceed, the 'early-burst' scenario[1,5] (for body shape) and the stages model[1,6,7] (for all traits investigated). Through the analysis of two genomes per species and by taking advantage of the uneven distribution of species in subclades of the radiation, we further show that species richness scales positively with per individual heterozygosity, but is not correlated with transposable element content, number of gene duplications, or genome-wide levels of selection in coding sequences.

At the macroevolutionary level, the diversity of life has mainly been shaped by two antagonistic processes: evolutionary radiations increase and extinction events decrease organismal diversity over time[5,8,9]. Evolutionary radiations are referred to as adaptive radiations if new lifeforms evolve rapidly through adaptive diversification into a variety of ecological niches, which typically presupposes ecological opportunity[1–3,10]. Whether or not an adaptive radiation unfolds depends on a variety of extrinsic and intrinsic factors as well as on contingency, whereas the magnitude of an adaptive radiation is determined by the interplay between its main components, speciation (minus extinction) and adaptation to distinct ecological niches[1,2,4,11]. Despite considerable scientific interest in the phenomenon of adaptive radiation as cradle of organismal diversity[1,2,10,12,13], many predictions

regarding its drivers and dynamics remain untested, particularly in exceptionally species-rich instances. Here, we examine what some consider as the 'most outstanding example of adaptive radiation'[14], the species flock of cichlid fishes in African Lake Tanganyika. This cichlid assemblage comprises about 240 species[15], which together feature an extraordinary degree of morphological, ecological, and behavioural diversity[14–17]. We construct a species tree of Lake Tanganyika's cichlid fauna based on genome-wide data, demonstrate the adaptive nature of the radiation, reconstruct eco-morphological diversification along the species tree, and test general and cichlid-specific predictions related to adaptive radiation.

In situ radiation in Lake Tanganyika

To establish the phylogenetic context of cichlid evolution in Lake Tanganyika, we estimated the age of the radiation through divergence time analyses based on cichlid and other teleost fossils[18], and constructed time-calibrated species trees using 547 newly sequenced cichlid genomes (Extended Data Table 1). Our new phylogenetic hypotheses (Fig. 1, Extended Data Figs. 1-3) support the assignment of the Tanganyikan cichlid fauna into 16 subclades – corresponding to the taxonomic grouping of species into tribes[15] – and confirm that the Tanganyikan representatives of the tribes Coptodonini, Oreochromini, and Tylochromini belong to more ancestral and widespread lineages that have colonised the lake secondarily[12,15,19] (Supplementary Discussion). It has been under debate whether all endemic Tanganyikan cichlid tribes evolved within the confines of Lake Tanganyika or whether some of them evolved elsewhere before the formation of the lake[20–22]. Our new time calibrations establish that the most recent common ancestor of the cichlid radiation in Lake Tanganyika lived around 9.7 (±0.5) Ma (Fig. 1), which coincides with the appearance of lacustrine conditions in the Tanganyikan Rift[23]. This suggests that the radiation commenced shortly after the lake had formed and that all endemic cichlid tribes have evolved and diversified *in situ*, that is, within the temporal and geographic context of Lake Tanganyika.

Phenotypes correlate with environments

Because – in the case of adaptive radiation – diversification occurs via niche specialisation, a strong association is expected in the extant fauna between the environment occupied by a species and the specific morphological features used to exploit it[2,3]. To quantify eco-morphological diversification across the radiation, we investigated three trait complexes through landmark-based morphometric analyses. Specifically, we quantified body shape and upper oral jaw morphology using 2D-landmarks acquired from X-ray images and the shape of the lower pharyngeal jaw bone based on 3D-landmarks derived from micro-computed tomography (μCT) scans (Extended Data Fig. 4). To approximate the ecological niche of each species we used the carbon and nitrogen stable isotope composition of muscle tissue, which informs about the relative position along the benthic-pelagic axis ($\delta^{13}C$ value) and the relative trophic level ($\delta^{15}N$ value), respectively[16,24] – a pattern which we corroborate here for Lake Tanganyika (Extended Data Fig. 5, Supplementary Discussion). The major axes of shape variation for each trait complex were identified through a principal component analysis (PCA). To test for phenotype-environment correlations and to identify the ecologically most relevant components of each of these trait complexes, we performed a two-block partial least square analysis (PLS) with the stable isotope measurements, and applied a phylogenetic generalised least square analysis (pGLS) to account for phylogenetic dependence.

The quantification of variation in body shape revealed that PC1 mainly represented differences in aspect ratio, while PC2 was loaded with changes in head morphology (Fig. 2a). The changes in aspect ratio (comparable to PC1) are correlated with the $\delta^{13}C$ and $\delta^{15}N$ values (PLS: Pearson's r = 0.69, $R^2 = 0.48$, $P = 0.001$; pGLS: $R^2 = 0.12$, $P < 0.001$, $\lambda_{pGLS} = 1.007$). PC1 of upper oral jaw morphology mainly represented changes in the orientation and relative size of the premaxilla, which was also the main correlate to the stable C and N isotope composition (PLS: Pearson's r = 0.62, $R^2 = 0.38$, $P = 0.001$; pGLS: $R^2 = 0.09$, $P < 0.001$, $\lambda_{pGLS} = 1.023$), while PC2 was defined by changes in the ratio of the rostral versus the lateral part of the bone (Fig. 2b). For lower pharyngeal jaw shape we found that PC1 mainly reflected changes in the aspect ratio of the jaw bone in combination with an increased posterior thickness, while PC2 involved similar shifts in thickness, yet in this case in combination with changes in the lengths of the postero-lateral horns that act as muscle attachment structures[25] (Fig. 2c). The PLS revealed that shape changes similar to PC2 are best associated with stable isotope values (PLS: Pearson's r = 0.67, $R^2 = 0.45$, $P = 0.001$; pGLS: $R^2 = 0.16$, $P < 0.001$, $\lambda_{pGLS} = 1.018$). The PCAs further revealed that the occupied area of the morphospace and ecospace scales with the number of species in the subclades (Extended Data Fig. 5; body: Pearson's r = 0.91, df = 9, $P < 0.001$; oral jaw: Pearson's r = 0.88, df = 9, $P < 0.001$; pharyngeal jaw: Pearson's r = 0.83, df = 9, $P = 0.002$; ecospace: Pearson's r = 0.88, df = 9, $P < 0.001$) – a pattern which is not only driven by sample size (Supplementary Discussion).

Overall, the significant association between each of the three traits and the stable C and N isotope composition underpins their adaptive value (Extended Data Fig. 6). A joint consideration points out that deep-bodied cichlids with an inferior mouth and thick lower pharyngeal jaws with short horns are associated with higher stable isotope projections (high $\delta^{13}C$ and low $\delta^{15}N$ values), indicating that such fishes predominantly occur in the benthic/littoral zone and feed on plants and algae, while more elongated species with a more superior mouth and longer and thinner lower pharyngeal jaws are generally associated with lower stable isotope projections (low $\delta^{13}C$ and high $\delta^{15}N$ values), suggesting a more pelagic lifestyle and a higher position in the food chain.

Pulses of morphological diversification

Next, we investigated the temporal dynamics of how the observed eco-morphological disparity emerged over the course of the radiation. In addition to the three eco-morphological traits, we also scored male pigmentation patterns to approximate disparity along the signalling axis – another potentially important component of diversification in adaptive radiations[1,6,7,26]. For all four traits, we estimated morphospace expansion through time using ancestral state reconstructions along the time-calibrated species tree and applying a variable rate model of trait evolution[27,28]. We calculated morphological disparity as extent of occupied morphospace in time intervals of 0.15 million years in comparison to a null model that assumes Brownian motion. Likewise, evolutionary rates through time were calculated as mean evolutionary rates derived from the variable rates model sampled at the same timepoints along the phylogeny.

Our analyses uncovered a pattern of discrete pulses in morphospace expansion, which were followed, in most cases, by morphospace packing (Fig. 3). Importantly, the timing of these pulses differed among the traits. For body shape, we found a pulse of rapid morphospace expansion early in the radiation, alongside with the first pulse of lower pharyngeal jaw shape diversification (Fig. 3b, c); this early phase of the radiation also features the highest evolutionary rates for body shape (Fig. 3d). The pulse in upper oral jaw diversification occurred in the middle phase of the radiation.

Evolutionary rates were elevated during this period, yet even higher at a later phase that was dominated by packing of the upper oral jaw morphospace rather than its expansion (Fig. 3b-d). This suggests that – in that later phase – rapidly evolving lineages diverged into pre-occupied regions of the morphospace, ultimately resulting in convergent forms[16]. The second pulse in lower pharyngeal jaw morphospace expansion happened late in the radiation when also evolutionary rates were highest for this trait (Fig. 3b-d). Thus, the theoretical prediction that eco-morphological diversification is rapid early in an adaptive radiation and slows down through time as the available niche space becomes filled[1,5] applies only to body shape. Yet, this 'early burst' in body shape diversification was not connected to a substantial increase in lineage accumulation (Fig. 3c, d).

Interestingly, the pigmentation patterns showed a single pulse of diversification and increased evolutionary rates late in the radiation – a signature unlikely to be caused by a high turnover rate in this trait (Supplementary Discussion). This late pulse of diversification in pigmentation patterns, together with the consecutive pulses of morphospace expansion in the eco-morphological traits, is in agreement with the prediction that diversification in an adaptive radiation proceeds in discrete temporal stages – first in macrohabitat use, then by trophic specialisation, followed by a final stage of divergence along the signalling axes[1,6,7]. However, in contrast to the conventional stages model, the most recent stage of the cichlid adaptive radiation in Lake Tanganyika, which coincides with a large number of speciation events (Fig. 3c, d), is characterised by temporally overlapping pulses of diversification in both a putative signalling trait and in an ecologically relevant trait. The lower pharyngeal jaw shape is the only trait complex showing two discrete pulses of morphospace expansion – one early in the radiation and one late when niche space already became limited. This later pulse suggests that diversification in the pharyngeal jaw apparatus facilitated fine-scaled resource partitioning after body shape and upper oral jaw morphospaces had been explored, resulting in the densely packed niche space observed today (Fig. 3b).

Genomic features and species richness

Finally, we examined whether the diversity patterns arising over the course of the radiation are linked with particular genomic features. It has previously been suggested – based on five reference cichlid genomes – that the radiating African cichlid lineages are characterised by elevated transposable element counts, increased levels of gene duplications, and genome-wide accelerated coding sequence evolution[13]. Because of the phylogenetic sub-structure of Lake Tanganyika's cichlid fauna and the widely differing species numbers among tribes, our data offered the opportunity to examine genomic features for an association with per-tribe species richness within a large-scale radiation. We did not find evidence that members of species-rich tribes exhibit greater numbers of transposable elements (Fig. 4a) or more duplicated genes in their genomes (Fig. 4b), nor do they feature elevated genome-wide signatures of selection in coding sequences (Fig. 4c). However, we found that a tribe's species richness scales positively with a common measure of genetic diversity, genome-wide heterozygosity (Fig. 4d). That genetic diversity is linked to species richness has been suspected before, although the nature of this relationship as well as the determinants of genetic diversity are under debate[29,30].

Elevated levels of heterozygosity could potentially result from hybridisation[31], which by itself has been suggested as a trigger of cichlid radiations[22,32,33]. In Tanganyikan cichlids, the level of gene flow within tribes (estimated using f_4-ratio values[34]) does not correlate with a tribe's species richness (Fig. 4e; Extended Data Fig. 8). Nevertheless, much of the variation in heterozygosity as well as its correlation with species richness can be explained by the observed levels of gene flow within tribes

in combination with the reduced gene flow among them: Through coalescent simulations of genome evolution along the species tree we show that variation in migration rates, sampled from the empirical f_4-ratio estimates, can produce levels of heterozygosity that are similar to the ones observed in nature (Fig. 4f). Hence, the correlation between species richness and heterozygosity can be explained by gene flow and phylogenetic structure, which is consistent with the expectation that the effect of gene flow scales positively with the number of hybridising species and the divergence among these. In the, an order of magnitude younger, cichlid radiation in Lake Malawi, heterozygosity levels vary much less among lineages and do not scale with species richness, which – according to our findings – can be explained by the much lower levels of genetic differentiation between the hybridising species[33].

Conclusion

Based on a comprehensive dataset on cichlid fishes from African Lake Tanganyika we tested predictions related to the phenomenon of adaptive radiation. We establish that the Tanganyikan cichlid radiation unfolded within the temporal and spatial confines of the lake, giving rise to an endemic fauna consisting of ~240 species in 52 genera and 13 tribes in less than 10 Myr. Although the ancestors of the tribes initially found comparable ecological opportunity, present-day species numbers differ by two orders of magnitude among these phylogenetic sub-lineages. Our analyses of morphological, ecological, and genomic information revealed that, taken as a whole, species-rich tribes occupy larger fractions of the morphospace and ecospace and contain species that are, at the per-genome level, genetically more diverse, which appears to be linked to gene flow. We demonstrate a phenotype-environment association in three trait complexes (body shape, upper oral jaw morphology, and lower pharyngeal jaw shape) and pinpoint their most relevant adaptive components. Importantly, we show that eco-morphological diversification was not gradual over the course of the radiation. Instead, we identified trait-specific pulses of accelerated phenotypic evolution, whereby only diversification in body shape shows an 'early burst'[1,5]. The sequence of the trait-specific pulses essentially follows the pattern postulated in the stages model of adaptive radiation[1,6,7], with the extension that the most recent 'stage' of the cichlid adaptive radiation in Lake Tanganyika, which is characterised by a large number of speciation events, is defined by increased diversification in both an ecological (lower pharyngeal jaw) and a signalling (pigmentation) trait. To what extent the observed diversity and disparity patterns were shaped by past environmental fluctuations and extinction dynamics cannot be answered conclusively through the investigation of the extant fauna alone.

References

1. Gavrilets, S. & Losos, J. B. Adaptive radiation: Contrasting theory with data. *Science* **323**, 732–737 (2009).
2. Schluter, D. *The ecology of adaptive radiation*. (Oxford University Press, 2000).
3. Simpson, G. G. *The Major Features of Evolution*. (Columbia University Press, 1953).
4. Glor, R. E. Phylogenetic Insights on Adaptive Radiation. *Annu. Rev. Ecol. Evol. Syst.* **41**, 251–270 (2010).
5. Foote, M. The evolution of morphological diversity. *Annu. Rev. Ecol. Syst.* **28**, 129–152 (1997).
6. Danley, P. D. & Kocher, T. D. Speciation in rapidly diverging systems: lessons from Lake Malawi. *Mol. Ecol.* **10**, 1075–1086 (2001).
7. Streelman, J. T. & Danley, P. D. The stages of vertebrate evolutionary radiation. *Trends Ecol. Evol.* **18**, 126–131 (2003).

8. Benton, M. J. Diversification and extinction in the history of life. *Science* **268**, 52–58 (1995).
9. Sepkoski, J. J. Rates of speciation in the fossil record. *Philos. Trans. R. Soc. London. Ser. B Biol. Sci.* **353**, 315–326 (1998).
10. Berner, D. & Salzburger, W. The genomics of organismal diversification illuminated by adaptive radiations. *Trends Genet.* **31**, 491–499 (2015).
11. Wagner, C. E., Harmon, L. J. & Seehausen, O. Ecological opportunity and sexual selection together predict adaptive radiation. *Nature* **487**, 366–369 (2012).
12. Salzburger, W. Understanding explosive diversification through cichlid fish genomics. *Nat. Rev. Genet.* **19**, 705–717 (2018).
13. Brawand, D. *et al.* The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513**, 375–381 (2014).
14. Fryer, G. & Iles, T. D. *The Cichlid Fishes of the Great Lakes of Africa*. (T.F.H. Publications, 1972).
15. Ronco, F., Büscher, H. H., Indermaur, A. & Salzburger, W. The taxonomic diversity of the cichlid fish fauna of ancient Lake Tanganyika, East Africa. *J. Gt. Lakes Res.* (2019) doi:doi.org/10.1016/j.jglr.2019.05.009.
16. Muschick, M., Indermaur, A. & Salzburger, W. Convergent evolution within an adaptive radiation of cichlid fishes. *Curr. Biol.* **22**, 2362–8 (2012).
17. Salzburger, W., Van Bocxlaer, B. & Cohen, A. S. Ecology and evolution of the African Great Lakes and their faunas. *Annu. Rev. Ecol. Evol. Syst.* **45**, 519–545 (2014).
18. Matschiner, M., Böhne, A., Ronco, F. & Salzburger, W. The genomic timeline of cichlid diversification across continents. co-submitted to *Nat. Commun.*
19. Koch, M. *et al.* Evolutionary history of the endemic Lake Tanganyika cichlid fish Tylochromis polylepis: A recent intruder to a mature adaptive radiation. *J. Zool. Syst. Evol. Res.* **45**, 64–71 (2007).
20. Salzburger, W., Meyer, A., Baric, S., Verheyen, E. & Sturmbauer, C. Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. *Syst. Biol.* **51**, 113–135 (2002).
21. Schedel, F. D. B., Musilova, Z. & Schliewen, U. K. East African cichlid lineages (Teleostei: Cichlidae) might be older than their ancient host lakes: New divergence estimates for the east African cichlid radiation. *BMC Evol. Biol.* **19**, 1–25 (2019).
22. Irisarri, I. *et al.* Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. *Nat. Commun.* **9**, 1–12 (2018).
23. Cohen, A. S., Soreghan, M. J. & Scholz, C. A. Estimating the age of formation of lakes: an example from Lake Tanganyika, East African Rift system. *Geology* **21**, 511–514 (1993).
24. Post, D. M. Using stable isotopes to estimate trophic position: models, methods, and assumptions. *Ecology* **83**, 703–718 (2002).
25. Liem, K. F. Evolutionary strategies and morphological innovations: Cichlid pharyngeal jaws. *Syst. Zool.* **22**, 425–441 (1973).
26. Salzburger, W. The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. *Mol. Ecol.* **18**, 169–85 (2009).
27. Venditti, C., Meade, A. & Pagel, M. Multiple routes to mammalian diversity. *Nature* **479**, 393–396 (2011).
28. Cooney, C. R. *et al.* Mega-evolutionary dynamics of the adaptive radiation of birds. *Nature* **542**, 344–347 (2017).
29. Ellegren, H. & Galtier, N. Determinants of genetic diversity. *Nat. Rev. Genet.* **17**, 422–433 (2016).
30. Schluter, D. & Pennell, M. W. Speciation gradients and the distribution of biodiversity. *Nature* **546**, 48–55 (2017).
31. Grant, P. R. & Grant, B. R. *40 Years of Evolution: Darwin's Finches on Daphne Major Island*. (Princeton University Press, 2014).
32. Meier, J. I. *et al.* Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat. Commun.* **8**, 1–11 (2017).

33. Malinsky, M. *et al.* Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat. Ecol. Evol.* **2**, 1940–1955 (2018).

34. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).

Figure Legends

Fig. 1 | Time-calibrated species tree of the cichlid fish fauna of African Lake Tanganyika. The species tree was time calibrated with a relaxed-clock model and is based on a maximum-likelihood topology inferred from genome-wide nuclear SNPs. Species names are abbreviated using a six-letter code, whereby the first three letters represent the genus and the last three letters the species name (Extended Data Table 1; see Extended Data Fig. 2a for the phylogeny with full species names). Branches are coloured according to tribes and for all endemic species an illustration is shown. Representatives of riverine cichlids (grey font) are nested within the radiation. The inset shows the time-calibrated phylogeny of more ancestral cichlid lineages (estimated under the multi-species coalescent model, Extended Data Fig. 1), highlighting the phylogenetic positions of the Tanganyikan representatives of the tribes Coptodonini (*Copren*; *Coptodon rendalli*), Oreochromini (*Oretan; Oreochromis tanganicae*), Tylochromini (*Tylpol; Tylochromis polylepis*) which colonised the lake secondarily. The schematic map of the African continent shows the position of the three Great Lakes Victoria, Malawi, and Tanganyika, with a magnified section of the latter. The presumed age of Lake Tanganyika (9-12 Ma)[23] is indicated in blue along the time axes. Species trees based on alternative topologies are presented in Extended Data Fig. 2b,c, and uncalibrated nuclear and mitochondrial phylogenies on the specimen level are shown in Extended Data Fig. 3.

Fig. 2 | Morphospace and ecospace occupation of the cichlid fish fauna of Lake Tanganyika. Principal component analyses of body shape (**a**, $n = 242$ taxa; 2,197 specimens), upper oral jaw morphology (**b**, $n = 242$; 2,197 specimens) and lower pharyngeal jaw shape (**c**, $n = 239$) along with the associated shape changes. **d**, Ecospace spanned by the stable C and N isotope compositions ($\delta^{13}C$ and $\delta^{15}N$ values; $n = 236$; 1,168 specimens). The colour scale indicates the number of species in 20 by 20 bins across the morpho- and ecospace, respectively (see Extended Data Fig. 5 for PCA and stable isotopes biplots with a focus on morpho- and ecospace occupation per tribe).

Fig. 3 | Temporal dynamics of diversification in body shape (first row), upper oral jaw morphology (second row), lower pharyngeal jaw shape (third row), and pigmentation patterns (fourth row) in the adaptive radiation of cichlid fishes in Lake Tanganyika. a, Species tree (Fig. 1) with branches coloured according to the mean relative rates of trait evolution for each trait. PP = posterior probability for rate shift. **b**, Morphospace densities (number of lineages) through time for each trait. Blue lines indicate the point in time when 50% of the extant morphospace had become occupied. **c**, Comparison of slopes (blue) of morphospace expansion over time between the observed data and the Brownian motion null model of trait evolution. A difference in slopes above zero represents morphospace expansion and values below zero indicate morphospace packing relative to the null model. The shaded areas show 95% quantiles of the 500 Brownian motion simulations. Lineage accumulation through time derived from the species tree is shown in dark grey. **d**, Mean relative rates of trait evolution over time with standard deviation (blue). Lineage accumulation through time is shown in dark grey.

Fig. 4 | Association between genomic features and species richness across the cichlid tribes in Lake Tanganyika. Each genomic summary statistic was tested for a correlation with species richness per tribe (log-transformed). To account for phylogenetic structure in the data, we calculated phylogenetic independent contrasts for each variable. Data points are coloured according to tribes; large points are tribe means, shown with 95% confidence intervals, small points represent species means and are only shown for group sizes < 40.

a, Percentage of the genome identified as transposable elements (TEs) (Pearson's r = -0.31, df = 10, $P = 0.33$; tribe means are based on one genome per species; Extended Data Fig. 7a). **b**, Number of duplicated genes (Pearson's r = -0.27, df = 10, $P = 0.40$; tribe means are based on species means). **c**, Genome-wide dN/dS ratios as a measure of selection on coding sequences (Pearson's r = 0.26, df = 10, $P = 0.42$; tribe means are based on species means across a set of 15,294 genes per genome; Extended Data Fig. 7b). **d**, Percentage of heterozygous sites per genome (Pearson's r = 0.70, df = 10, $P = 0.012$; tribe means are based on species means). **e**, f_4-ratio statistics as a measure of gene flow among species within each tribe (Pearson's r = -0.35, df = 9, $P = 0.29$; tribe means are based on all species triplets within each tribe; see Extended Data Fig. 8b for a summary of the f_4-ratio statistics for all species comparisons). **f**, Mean percentage of heterozygous sites in simulations with within-tribe migration rates sampled from the observed f_4-ratio statistics (Pearson's r = 0.85, df = 10, $P < 0.001$; tribe means are based on species means across 20 simulations; Extended Data Fig. 7c).

Methods

Sampling

Sampling was conducted between 2014 and 2017 at 130 locations at Lake Tanganyika. To maximise taxon coverage, we included additional specimens from previous expeditions (4.9% of the samples) as well as from other collections (0.8%). The final dataset (301 taxa; $n = 2,723$ specimens) contained an almost complete taxon sampling of the cichlid fauna of Lake Tanganyika, as well as 18 representative cichlid species from nearby waterbodies, and 32 outgroup species. All analyses described below are based on the same set of typically 10 specimens per species, or subsets thereof (see Extended Data Table 1 and Supplementary Methods for details).

Whole genome sequencing

Genomic DNA of typically one male and one female specimen per species ($n = 547$) was extracted from fin-clips preserved in ethanol using the E.Z.N.A. Tissue DNA Kit (Omega Bio-Tek) and sheared on a Covaris E220 (60 µl with 10% duty factor, 175 W, 200 cycles for 65 sec). Individual libraries were prepared using Illumina's TruSeq DNA PCR-Free Sample Preparation kit (Low Sample Protocol) for 350 bp insert size, pooled (six per lane), and sequenced at 126 bp paired-end on an Illumina HiSeq 2500 (Extended Data Table 1 contains information on read depths).

Assessing genomic variation

After adapter removal with Trimmomatic[35] (v.0.36), reads of 528 genomes (all species belonging to the cichlid radiation in Lake Tanganyika plus additional species nested within this radiation and some selected outgroup species; Extended Data Table 1) were mapped to the Nile tilapia reference genome (RefSeq accession GCF_001858045.1[36]) using BWA-MEM[37] (v.0.7.12). Variant calling was performed with GATK's HaplotypeCaller and GenotypeGVCF tools[38] (v.3.7), applying a minimum base quality score of 30. Variant calls were filtered with BCFtools[39] (v.1.6; FS < 20, QD > 2, MQ > 20, DP > 4000, DP < 8000, ReadPosRankSum > -0.5, MQRankSum > -0.5). We applied a filter to sites in proximity to indels with a minor allele count greater than 2, depending on the size of the indel. With SNPable (http://lh3lh3.users.sourceforge.net/snpable.shtml), we determined all sites within regions of the tilapia reference genome in which read mapping could be ambiguous and masked these sites. Using VCFtools[40] (v.0.1.14) we further masked, per individual, genotypes with a read depth below 4 or a genotype quality below 20. Sites that were no longer polymorphic after the filtering steps were excluded, resulting in a dataset of 57,751,375 SNPs. Called variants were phased with the software beagle[41] (v.4.1). The phasing of *Neolamprologus cancellatus,* which appeared to be F1 hybrids, was further improved with a custom script. Further details are provided in the Supplementary Methods.

De novo genome assemblies

De novo genome assemblies were generated from the raw-read data for each individual following an approach described previously[42,43], using CeleraAssembler[44] (v.8.3) and FLASH[45] (v.1.2.11). Eight genomes repeatedly failed to assemble and were therefore excluded from further analyses (specimen vouchers: A188, IRF6, IZC5, JWE7, JWG1, JWG2, LJD3, and LJE8). Assembly quality was assessed with QUAST[46] (v.4.5) and completeness was determined with BUSCO[47] (v.3). Assembly statistics summarised with MultiQC[48] (v.1.7) are available on Dryad.

Determining the age of the radiation

To determine the age of the cichlid radiation in Lake Tanganyika, we applied phylogenomic molecular-clock analyses for representatives of all cichlid subfamilies and the most divergent tribes, together with non-cichlid outgroups (44 species; Extended Data Fig. 1). Following Matschiner et al.[18] we identified and filtered ortholog sequences from genome assemblies and compiled 'strict' and 'permissive' datasets that contained alignments for 510 and 1,161 genes and had total alignment lengths of 542,922 and 1,353,747 bp, respectively. We first analysed the topology of the species with the multi-species coalescent model implemented in ASTRAL[49] (v.5.6.3), based on gene trees that we estimated for both datasets with BEAST2[50] (v.2.5.0).

As undetected past introgression can influence divergence-time estimates in molecular clock analyses, we further tested for signals of introgression in the form of asymmetric species relationship in gene trees and excluded five species (*Fundulus heteroclitus*, *'Tilapia' brevimanus*, *Pelmatolapia mariae*, *Tilapia sparrmanii,* and *Steatocranus* sp. "ultraslender") potentially affected by introgression from all subsequent molecular-clock analyses. We then estimated divergence times among the most divergent cichlid tribes and the age of the radiation with the multi-species coalescent model in StarBEAST2[51] (v.0.15.5), using the 'strict' set of gene alignments (Extended Data Fig. 1). Further details are provided in the Supplementary Methods.

Phylogenetic inference

To infer a complete phylogeny of the cichlid radiation in Lake Tanganyika (the Tanganyikan representatives of the more ancestral tribes Coptodonini, Oreochromini, and Tylochromini were excluded) from genome-wide SNPs we applied additional filters, retaining only SNPs with < 40% missing data and between-SNP distances of at least 100 bp. The remaining 3,630,997 SNPs were used to infer a maximum-likelihood phylogeny with RAxML[52] (v.8.2.4; Fig. 1, Extended Data Fig. 2a, 3a). The species-tree topology was further estimated under the multi-species coalescent model from a set of local phylogenies with ASTRAL (Extended Data Fig. 2b); these local phylogenies were inferred with IQ-TREE[53] (v.1.7-beta7) from alignments for 1,272 genomic regions determined to be particularly suitable for phylogenetic analysis (see Supplementary Methods). We also applied the multi-species coalescent model implemented in SNAPP[54] (v.1.4.2) to the dataset of genome-wide SNPs (Extended Data Fig. 2c). Species-level phylogenies resulting from these different approaches were used as topological constraints in subsequent relaxed-clock analyses of divergence times (see below). In addition, we estimated the mitochondrial phylogeny based on maximum-likelihood with RAxML. Further details are provided in the Supplementary Methods.

Divergence time estimates within the radiation

For relaxed-clock analyses, the 1,272 alignments were further filtered by applying stricter thresholds on the proportion of missing data and the strength of recombination signals. Ten remaining alignments with a length greater than 2,500 bp and less than 130 hemiplasies (total length: 30,738 bp; completeness: 95.8%), were then used jointly to estimate divergence times with the uncorrelated-lognormal relaxed-clock model implemented in BEAST2. To account for phylogenetic uncertainty in downstream phylogenetic comparative analyses, we performed three separate sets of relaxed clock analyses, in which the topology was either fixed to the species-level phylogeny inferred with RAxML (Fig. 1, Extended Fig. 2a), the species tree inferred with ASTRAL

(Extended Fig. 2b), or the Bayesian species tree inferred with SNAPP (Extended Fig. 2c). Further details are provided in the Supplementary Methods.

Morphometrics

To quantify body shape and upper oral jaw morphology we applied a landmark-based geometric morphometric approach to digital X-ray images (for the full set of 10 specimens per species whenever possible; n = 2,197). We selected 21 landmarks, of which 17 were distributed across the skeleton and four defined the premaxilla (Extended Data Fig. 4a). Landmark coordinates were digitised using FIJI[55] (v2.0.0-rc-68/1.521i). To extract overall body shape information, we excluded landmark 16, which marks the lateral end of the premaxilla, hence minimizing the impact of the orientation of the upper oral jaw. We then applied a Procrustes superimposition to remove the effect of size, orientation, and translational position of the coordinates.

For upper oral jaw morphology, we used a subset of four landmarks. A crucial feature of the oral jaw morphology is the orientation of the mouth relative to the body axes. However, this component of the upper oral jaw morphology would be lost in a classical geometric morphometric analysis, in which only pure shape information is retained. To overcome this, we extracted the premaxilla-specific landmarks (1, 2, 16, and 21) *after* Procrustes superimposition of the entire set of landmarks and subsequently re-centred the landmarks to align the specimens without rotation. Thus, the resulting landmark coordinates do not represent the pure shape of the premaxilla but additionally contain information on its orientation and size in relation to body axes and body size, respectively.

To quantify lower pharyngeal jaw bone shape in 3D, a landmark-based geometric morphometric approach was applied on μCT-scans of the head region of five specimens per species (n = 1,168). To capture all potential functionally important structures of the lower pharyngeal jaw bone, we selected a set of 27 landmarks (10 true landmarks and 17 sliding semi-landmarks) well distributed across the left side of the bone (Extended Data, Fig. 4b). Landmark coordinates were acquired using TINA[56] (v.6.0). To retain the lateral symmetric properties of the shape data during superimposition, we reconstructed the right side of the lower pharyngeal jaw bone by mirroring the landmark coordinates across the plane of bilateral symmetry fitted through all landmarks theoretically lying on this plane. We then superimposed the resulting 42 landmarks while sliding the semi-landmarks along the curves by minimizing Procrustes distances and retained the symmetric component only.

To identify the major axes of shape variation across the multivariate datasets we performed a PCA for each trait. We also calculated morphospace size per tribe as the square root of the convex hull area spanned by species means of the PC1- and PC2-scores. We then tested for a correlation between morphospace size and estimated species richness of a tribe[15] (log-transformed to obtain normal distribution). To account for phylogenetic non-independence, we calculated phylogenetic independent contrasts with the R package ape[57] (v.5.2) using the species tree (Fig. 1) pruned to the tribe level. We then calculated Pearson's correlation coefficients for independent contrasts using the function *cor.table* of the R package picante[58] (v.1.8).

All landmark coordinates for geometric morphometric analyses were processed and analysed in R[59] (v.3.5.2) using the packages geomorph[60] (v.3.0.7) and Morpho[61] (v.2.6). Further details are provided in the Supplementary Methods.

Stable isotope analysis

To approximate ecology for each species, we measured the stable carbon (C) and nitrogen (N) isotope composition of all available specimens from Lake Tanganyika (n = 2,259). We analysed a small (0.5 – 1 mg) dried muscle sample of each specimen with a Flash 2000 elemental analyser coupled to a Delta Plus XP continuous-flow isotope ratio mass spectrometer (IRMS) via a Conflo IV interface (Thermo Fisher Scientific,

Bremen, Germany). Carbon and nitrogen isotope data were normalised to the VPDB (Vienna Pee Dee Belemnite) and Air-N_2 scales, respectively, using laboratory standards which were calibrated against international standards. Values are reported in standard per-mil notation (‰), and long-term analytical precision was 0.2‰ for $\delta^{13}C$ values and 0.1‰ for $\delta^{15}N$ values. To test for a correlation of ecospace size with species richness of the tribes, we applied the same approach as described above to the $\delta^{13}C$ and $\delta^{15}N$ values.

To confirm interpretability of the $\delta^{13}C$ and $\delta^{15}N$ values, we additionally collected and analysed baseline samples covering several trophic levels from the northern and the southern basin of Lake Tanganyika (Supplementary Methods, Supplementary Discussion).

Phenotype-environment association

For each trait (body shape, upper oral jaw, lower pharyngeal jaw) we performed a two-block PLS analysis based on species means of the Procrustes aligned landmark coordinates and the stable C and N isotope compositions using the function *two.b.pls* in Geomorph. To account for phylogenetic dependence of the data we applied a pGLS as implemented in the R package caper[62] (v.1.0.1) across the two sets of PLS scores (each morphological axis and the stable isotope projection) using the time-calibrated species tree based on the maximum-likelihood topology. The strength of phylogenetic signal in the data was accounted for by optimising the branch length transformation parameter lambda using a maximum-likelihood approach.

Scoring pigmentation patterns

To quantify a putative signalling trait in cichlids, we scored the pigmentation patterns in typically five male specimens per species (n = 1,028), on the basis of standardised images taken in the field after capture of the specimens (see Supplementary Methods). Following the strategy described in Seehausen et al.[63], the presence/absence of 20 pigmentation features was recorded, whereby we extended number of scored features to include additional body and fin pigmentation patterns (Extended Data Fig. 4c). We then applied a logistic PCA implemented in the R package logisticPCA[64] (v.0.2) and used the PC1-scores as univariate proxy for differentiation along the signalling axes for further analyses.

Trait evolution modelling and disparity estimates

To investigate the temporal dynamics of morphological diversification over the course of the radiation we essentially followed the strategy of Cooney et al.[28] (which is based on measurements on extant taxa and assumes constant niche-space and no or constant extinction over the course of the radiation), using the PLS-scores of body shape, upper oral jaw morphology, and lower pharyngeal jaw shape and the PC1-scores of pigmentation patterns as well as the time-calibrated maximum-likelihood species tree topology. For each trait we assessed the phylogenetic signal in the data by calculating Pagel's Lambda and Blomberg's K with the R package phytools[65] (v.0.6-60). We then tested the fit of four models of trait evolution for each of the four traits. We applied a white noise model, a Brownian motion model, a single-optimum Ornstein-Uhlenbeck model, and an 'early burst' model of trait evolution using the function *fitContinuous* of the R package geiger[66] (v.2.0.6.1). Additionally, we fitted a variable rates model (a Brownian motion model which allows for rate shift on branches and nodes) using the software BayesTrait (http://www.evolution.rdg.ac.uk/, v.3) with uniform prior distributions adjusted to our dataset (alpha: -1 – 1, sigma: 0 – 0.001 for morphometric traits; alpha: 0 – 10, sigma: 0 – 10 for pigmentation pattern) and applying single-chain Markov-chain Monte Carlo runs with one billion iterations. We sampled parameters every 100,000th iteration, after a pre-set burnin of 10,000,000 iterations. We then tested for each trait for convergence of the chain using a Cramer-von-Mises statistic implemented in the R package coda[67] (v.0.19-3). The models were compared by calculating their log-likelihood and Akaike information criterion (AIC) difference (Extended Data Table 2). Based on differences in AIC, the variable rates model was best supported for all traits but body shape, which showed a strong signal of an early burst of trait evolution (Extended Table 2, note that the variable rates model has the highest log-

likelihood for body shape as well). We nevertheless focused on the variable rates model for further analyses of all traits to be able to compare temporal patterns of trait evolution among the traits.

To estimate morphospace expansion through time we used a maximum-likelihood ancestral state reconstruction implemented in phytools. To account for differences in the rate of trait evolution along the phylogeny, we reconstructed ancestral states using the mean rate-transformed tree derived from the variable rates model. We then projected the ancestral states onto the original species tree and calculated the morphospace extent (i.e. the range of trait values) in time intervals of 0.15 million years (note that this is an arbitrary value; however, differently sized time intervals had no effect on the interpretation of the results). For each time point we extracted the branches existing at that time and predicted the trait value linearly between nodes. We then compared the resulting morphospace expansion over time relative to a null model of trait evolution. We therefore simulated 500 datasets (PLS and PC1 scores) under Brownian motion given the original species tree with parameters derived from the Brownian motion model fit to the original data. For each simulated dataset we produced morphospace-expansion curves using the same approach as described above. We then compared the slopes of our observed data with each of the null models by calculating the difference of slopes through time (Fig. 3) using linear models fitted for each time interval with the two subsequent time intervals. (Note that for body shape we also estimated morphospace expansion through time using the early burst model for ancestral state reconstruction, which resulted in a very similar pattern of trait diversification.)

Unlike other metrics of disparity (e.g. variance or mean pairwise distances) morphospace extent is not sensitive to the density distribution of measurements within the morphospace and captures its full range[68]. Hence, comparing the extent of morphospace between observed data and the null model directly unveils the contribution of morphospace expansion relative to the null model; and because the increase in lineages over time is identical in the observed and the simulated data, this comparison also provides an estimate for morphospace packing.

To summarise evolutionary rates we calculated the mean rate of trait evolution inferred by the variable rates model in the same 0.15 million years intervals along the phylogeny.

To account for phylogenetic uncertainty in the tree topology we repeated the analyses of trait evolution using the time-calibrated trees based on tree topologies estimated with ASTRAL and SNAPP (Extended Data Fig. 2b,c; Supplementary Methods; Supplementary Discussion). Furthermore, to also account for uncertainty in branch lengths, we repeated the analysis on 100 trees from the Bayesian posterior distribution for each of the three trees.

Further details can be found in the Supplementary Methods.

Characterisation of repeat content

For the repeat content analysis, we randomly selected one *de novo* genome assembly per species of the radiation ($n = 245$). We performed a *de novo* identification of repeat families using RepeatModeler[69] (v.1.0.11). We then combined the RepeatModeler output library with the available cichlid-specific libraries[70] (Dfam and RepBase; v.27.01.2017; 258 ancestral and ubiquitous sequences, 161 cichlid-specific repeats, and 6 lineage-specific sequences; 65,118, 273,530, and 6,667 bp in total, respectively) and used the software RepeatMasker[70] (v.4.0.7) (-xsmall -s -e ncbi -lib combined_libraries.fa) to identify and soft-mask interspersed repeats and low complexity DNA sequences in each assembly. The reported summary statistics were obtained using RepeatMasker's 'buildSummary.pl' script (Fig. 4a, Extended Data Fig. 7a, results per genome are provided on Dryad).

Gene duplication estimates

Per genome, gene duplication events were identified with the structural variant identification pipeline smoove (population calling method; https://github.com/brentp/smoove, docker image cloned 20/12/2018), which builds upon lumpy[71], svtyper[72], and svtools (https://github.com/hall-lab/svtools). Variants were called per sample (n = 488 genomes, 246 taxa of the Tanganyika radiation) from the initial mapping files against the tilapia reference genome with the function *call*. The union of sites across all samples was obtained with the function *merge,* then all samples were genotyped at those sites with the function *genotype*, and depth information was added with --duphold. Genotypes were combined with the function *paste* and annotated with *annotate* and the reference genome annotation file. The obtained VCF file was filtered with BCFtools to keep only duplications longer than 1 kb and of high quality (MSHQ > 3 or MSHQ == -1, FMT/DHFFC[0] > 1.3, QUAL > 100). The resulting file was loaded into R (v.3.6.0) with vcfR[73] (v.1.8.0) and filtered to keep only duplications with less than 20% missing genotypes. Next, we removed duplication events with a length outside 1.5 times the interquartile range above the upper quartile of all duplication length, resulting in a final dataset of 476 duplications (Fig. 4b).

Analyses of selection on coding sequence

To predict genes within the *de novo* genome assemblies, we used AUGUSTUS[74] (v.3.2.3) with default parameters and 'zebrafish' as --species parameter (n = 485 genomes, 245 taxa). For each prediction we inferred orthology to Nile tilapia genes (GCF_001858045.1_ASM185804v2) with GMAP (GMAP-GSNAP[75]; v.2017-08-15) applying a minimum trimmed coverage of 0.5 and a minimum identity of 0.8. We excluded specimens with less than 18,000 tilapia orthologous genes detected (n = 471 genomes, 243 taxa). Next, we kept only those tilapia protein coding sequences that had at least one of their exons present in at least 80% of the assemblies (260,335 exons were retained, representing 34,793 protein coding sequences). Based on the tilapia reference genome annotation file, we reconstructed for each assembly the orthologous coding sequences. Missing exon sequences were set to 'N's. We then kept a single protein coding sequence per gene (the one being present in the maximum number of species with the highest percentage of sequence length), resulting in 15,294 protein coding sequences. Per gene, a multiple sequence alignment was then produced using MACSE[76] (v.2.01). We calculated for each specimen and each gene the number of synonymous (S) and non-synonymous (N) substitutions by pairwise comparison to the ortholog tilapia sequence using *codeml* with runmode –2 within PAML[77] (v.4.9e). To obtain an estimate of the genome-wide sequence evolution rate that is independent of filtering thresholds, we calculated the genome-wide dN/dS ratio for each specimen based on the sum of dS and dN across all genes (Fig. 4c, Extended Data Fig. 7b).

Signals of past introgression

We used the f_4-ratio statistic[34] to assess genomic evidence for interspecific gene exchange. We calculated the f_4-ratio for all combinations of trios of species on the filtered VCF files using the software Dsuite[78] (v.0.2 r20), with *T. sparrmanii* as outgroup species (we excluded *N. cancellatus* as all specimens of this species appeared to be F1 hybrids; Supplementary Methods). The f_4-ratio statistic estimates the 'admixture proportion', i.e. the proportion of the genome affected by gene flow. The results presented in this manuscript (Fig. 4e, Extended Data Fig. 8a) are based on the 'tree' output of the Dsuite function *Dtrios*, with each trio arranged according to the species tree based on the maximum-likelihood topology. The per-tribe analyses (Fig. 4e) were based only on comparisons where all species within a trio belong to the same tribe (n = 243 taxa).

In addition to the f_4-ratio we also identified signals of past introgression among species using a phylogenetic approach by testing for asymmetry in the relationships of species trios in 1,272 local maximum-likelihood trees generated using IQ-TREE (Supplementary Methods; Extended Data Fig. 8b).

Heterozygosity

We calculated the number of heterozygous sites per genome (n = 488 genomes, 246 taxa from the Tanganyika radiation) from the VCF files using the BCFtools function *stats* and then quantified the percentage of heterozygous sites among the number of callable sites per genome (see above) (Fig. 4d).

To explore if the observed levels of heterozygosity per tribe can be explained by the levels of gene flow within tribes we performed coalescent simulations with msprime[79] (v.0.7.4). We simulated genome evolution of all species of the radiation following the time-calibrated species tree (maximum-likelihood topology), assuming a generation time of 3 years[80] and a constant effective population size of 20,000 individuals. Species divergences were implemented as mass migration events and introgression within tribes as migration between species pairs with rates set according to their introgression (f_4-ratio) signals inferred with Dsuite. To convert the f_4-ratio values into migration rates, we applied a scaling factor of 5×10^{-6}, which results in a close correspondence in magnitude of the simulated introgression signals to those observed empirically (Fig. 4, Extended Data Fig. 7c). In each of twenty separate simulations, we randomly sampled one pairwise f_4-ratio value for each pair of species (there are many f_4-ratios per species pair – one for each possible third species added to the test trio; the maximum values per pair are shown in Extended Data Figure 8a). The simulated data consisted of one chromosome of 100 kb (mutation rate: 3.5×10^{-9} per bp per generation[33], recombination rate: 2.2×10^{-8} per bp per generation; see Supplementary Methods). Levels of heterozygosity were calculated for all simulated datasets as described for the empirical data.

To account for between-tribe gene flow we further performed simulations in which migration between tribes was also sampled from the empirical f_4-ratio distribution. For simplicity in setting up the simulation model, we assume that gene flow between tribes is ongoing until present day, which is clearly an overestimate (see Supplementary Discussion). Nevertheless, the results of these simulations support our hypothesized scenario, confirming that much of the variation in heterozygosity as well as its correlation with species richness can be explained by the observed levels of gene flow.

Correlation of genome-wide statistics with species richness

We tested for a correlation between tribe means (based on species means) of each genomic summary statistics (TE counts, number of gene duplications, genome-wide dN/dS ratio, per-genome heterozygosity, and f_4-ratio, as well as the heterozygosity and f_4-ratio statistics derived from simulated genome evolution) and species richness of the tribes, applying the same approach as described above for tests of correlation between morpho- and ecospace size and species richness.

Method References

35. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
36. Conte, M. A., Gammerdinger, W. J., Bartie, K. L., Penman, D. J. & Kocher, T. D. A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. *BMC Genomics* **18**, 1–19 (2017).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
38. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
39. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
40. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

41. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

42. Böhne, A. *et al.* Repeated evolution versus common ancestry: Sex chromosome evolution in the haplochromine Pseudocrenilabrus philander. *Genome Biol. Evol.* **11**, 439–458 (2019).

43. Malmstrøm, M., Matschiner, M., Tørresen, O. K., Jakobsen, K. S. & Jentoft, S. Data descriptor: Whole genome sequencing data and de novo draft assemblies for 66 teleost species. *Sci. Data* **4**, 1–13 (2017).

44. Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287**, 2196–2204 (2000).

45. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

46. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

47. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

48. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

49. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 15–62 (2018).

50. Bouckaert, R. R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).

51. Ogilvie, H. A., Bouckaert, R. R. & Drummond, A. J. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol. Biol. Evol.* **34**, 2101–2114 (2017).

52. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

53. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

54. Bryant, D., Bouckaert, R. R., Felsenstein, J., Rosenberg, N. A. & RoyChoudhury, A. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* **29**, 1917–1932 (2012).

55. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

56. Schunke, A. C., Bromiley, P. A., Tautz, D. & Thacker, N. A. TINA manual landmarking tool: Software for the precise digitization of 3D landmarks. *Front. Zool.* **9**, 1–5 (2012).

57. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

58. Kembel, S. W. *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464 (2010).

59. R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing* (2018).

60. Adams, D. C. & Otárola-Castillo, E. Geomorph: An R package for the collection and analysis of geometric morphometric shape data. *Methods Ecol. Evol.* **4**, 393–399 (2013).

61. Schlager, S. Morpho and Rvcg – Shape Analysis in R. in *Statistical Shape and Deformation Analysis* (eds. Zheng, G., Li, S. & Szekely, G.) 217–256 (Academic Press, 2017).

62. Orme, D. The caper package : comparative analysis of phylogenetics and evolution in R. 1–36 (2018).

63. Seehausen, O., Mayhew, P. J. & Van Alphen, J. J. M. Evolution of colour patterns in East African cichlid fish. *J. Evol. Biol.* **12**, 514–534 (1999).

64. Landgraf, A. J. & Lee, Y. Dimensionality reduction for binary data through the projection of natural

parameters. 1–40 (2015).

65. Revell, L. J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

66. Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E. & Challenger, W. GEIGER: Investigating evolutionary radiations. *Bioinformatics* **24**, 129–131 (2008).

67. Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6**, 7–11 (2005).

68. Ciampaglio, C. N., Kemp, M. & McShea, D. W. Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. *Paleobiology* **27**, 695–715 (2001).

69. Smit, A. & Hubley, R. RepeatModeler Open-1.0. *2008-2015; http://www.repeatmasker.org*.

70. Smit, A. & Hubley, R. RepeatMasker Open-4.0. *2013-2015; http://www.repeatmasker.org*.

71. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* **15**, 1–19 (2014).

72. Chiang, C. *et al.* SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).

73. Knaus, B. J. & Grünwald, N. J. VCFR: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* **17**, 44–53 (2017).

74. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **9**, 7–62 (2006).

75. Wu, T. & Watanabe, C. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).

76. Ranwez, V., Douzery, E., Cambon, C., Chantret, N. & Delsuc, F. MACSE v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* **35**, 1–4 (2018).

77. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

78. Malinsky, M., Matschiner, M. & Svardal, H. Dsuite - fast D-statistics and related admixture evidence from VCF files. *bioRxiv* 634477 (2020) doi:10.1101/634477.

79. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Comput. Biol.* **12**, e1004842 (2016).

80. Malinsky, M. *et al.* Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science.* **350**, 1493–1498 (2015).

Acknowledgements

We thank the University of Burundi, the Ministère de l'Eau, de l'Environnement, de l'Amenagement du Territoire et de l'Urbanisme, Republic of Burundi, the Centre de Recherche en Hydrobiologie (CRH), Uvira, DR Congo, the Tanzania Commission for Science and Technology (COSTECH), the Tanzania Fisheries Research Institute (TAFIRI), the Tanzania National Parks Authority (TANAPA), the Tanzania Wildlife Research Institute (TAWIRI), the Lake Tanganyika Research Unit, Department of Fisheries, Republic of Zambia, and the Zambian Department for Immigration for research permits; G. Banyankimbona (University of Burundi), H. Mwima and G. Hakizimana (Lake Tanganyika Authority, Bujumbura, Burundi), N. Muderhwa and P. Masilya (CRH, Uvira, DR Congo), I. Kimirei (TAFIRI, Kigoma, Tanzania), M. Mukuli Wa-Teba (Tanzania), G. Moshi (Mahale Mountains National Park), A. Mwakatobe (TAWIRI), C. Katongo (University of Zambia, Lusaka), and T. Banda and L. Makasa (Department of Fisheries, Mpulungu, Zambia) for assistance with obtaining research permits; the boat crews of the 'Chomba' (D. Mwanakulya, J. Sichilima, H. D. Sichilima Jr., and G. Katai) and the 'Maji Makubwa II' (G. Kazumbe and family) for navigation, guidance and

company; the boat drivers M. Katumba (Mahale National Park) and T. Musisha (Zambia); the car drivers A. Irakoze (Burundi) and J. Leonard (Tanzania); M. Schreyen-Brichard (Bujumbura, Burundi), M. Mukuli Wa-Teba (Dar Es Salaam, Tanzania), G. Kazumbe (Kigoma, Tanzania), I. Kimirei (TAFIRI), D. and R. Schlatter (Marine & Motors Engineering, Tanzania), M. K. Dominico (Unity Travel, Tanzania), H. Sichilima Sr. (Chomba Trans, Zambia), C. Zytkow (Conservation Lake Tanganyika, Zambia), and P. Lassen and V. Huwiler (Kalambo Lodge, Zambia) for logistic support; G. Banyankimbona, N. Boileau, B. Egger, Y. Fermon, G. Kazumbe, G. Katai, R. Lusoma, K. Smailus, L. Widmer, and numerous fishermen at Lake Tanganyika for help during sampling; V. Huwiler (Kalambo Lodge, Zambia), 'Charity' (Nkupi Lodge, Zambia), O. Mangwangwa (Kasanga, Tanzania), and the Zytkow family (Ndole Bay Lodge, Zambia) for lodging; people of innumerable villages at the shores of Lake Tanganyika for providing workspace, shelter for night-camps, and access to village infrastructure; M. Barluenga, H. Gante, Z. Musilová, F. Schedel, J. Snoeks, M. Stiassny, H. Tanaka, and G. Turner for providing additional samples and/or specimens; M. Sánchez, A. Schweizer, and A. Wegmann (University of Zurich) for assistance with the μCT-scanning of large specimens; C. Moes (Novartis Institute for Biomedical Research, Basel, Switzerland) for help with radiographs; V. Evrard for help with stable isotopes; I. Nissen and E. Burcklen for assistance with DNA shearing; M. Conte and T.D. Kocher for sharing the RepeatMasker annotations for Nile tilapia; C. Klingenberg and M. Sánchez for discussions on the morphometric approach; A. Tooming-Klunderud and team at the Norwegian Sequencing Centre (NSC), Oslo, and C. Beisel and team at the Genomics Facility Basel (GFB) at the ETH Zurich Department of Biosystems Science and Engineering (D-BSSE), Basel, for assistance with next-generation sequencing; M. Jacquot, E. Pujades, and T. Sengstag (sciCORE) for the setup and assistance with the collection database system (LabKey); and J. Himes and A. Viertler for fish illustrations in Fig. 1 and Extended Data Fig. 4, respectively.

Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing centre at University of Basel (with support by the SIB/Swiss Institute of Bioinformatics) and the Abel computer cluster, University of Oslo.

This work was funded by the European Research Council (ERC, Consolidator Grant Nr. 617585 'CICHLID~X' jointly hosted by the University of Basel and the University of Oslo) and the Swiss National Science Foundation (SNSF, grants 156405 and 176039) to W.S. A.Böh. was supported by the SNSF (Ambizione grant 161462).

Author Contributions

F.R., A.I., and W.S. designed this study (with input from H.H.B., A.K., and S.J.). F.R., A.I., H.H.B, and W.S. collected the specimens in the field. F.R. and A.Böh. extracted DNA and prepared the libraries for sequencing. S.J. coordinated sequencing. M.Mat. performed the mapping, variant calling, phylogenetic analyses, and coalescent simulations. M.Mal. contributed to the variant calling pipelines and performed the f_4-ratio statistics. A.Böh. assembled the genomes and quantified gene duplications, A.E. conducted the dN/dS analyses, and V.R. analysed transposable elements. A.Boi. assessed stable isotope compositions, H.H.B. radiographed the specimens, and W.S. scored pigmentation patterns. F.R. performed μCT-scanning, geometric morphometric analyses, and all analyses incorporating morphological and ecological data. F.R. and W.S. wrote the manuscript with contributions and/or feedback from all authors. All authors read and approved the final version of the manuscript.

Additional Information

The authors declare no **Competing interests**.

Supplementary Information is available for this paper.

Correspondence should be addressed to F.R. (fabrizia.ronco@unibas.ch) and W.S. (walter.salzburger@unibas.ch).

Data availability

All newly sequenced genomes for this study and their raw reads are available from NCBI under the BioProject accession number PRJNA550295 (https://www.ncbi.nlm.nih.gov/bioproject/). The VCF file, tree files, summary statistics of the assembled genomes, and phenotypic datasets generated and analysed during this study are available as downloadable files on Dryad (https://datadryad.org/stash/share/13fM-BDssqlCELXWdSgpnkeawCuOFvo-tA9o-vEiZ_k).

Code availability

Code used to analyse the data is available on GitHub (https://github.com/cichlidx/ronco_et_al), except for analyses where single commands from publicly available software were used and where all settings are fully reported in the Methods and/or Supplementary Methods sections.

Extended Data

Extended Data Fig. 1 | Age of the adaptive radiation of cichlid fishes in African Lake Tanganyika. Time-calibrated species tree of species representing divergent tribes and subfamilies within cichlids as well as closely-related non-cichlid outgroups, generated with the multi-species coalescent model in StarBEAST2. Nodes marked with a black dot were constrained according to species-tree analyses with ASTRAL. Node bars indicate 95% highest-posterior density age intervals. Outgroup divergence times are not drawn to scale. Insets visualise the prior distribution applied for the age of African cichlids according to Matschiner et al.[18], as well as posterior age estimates for Oreochromini and the cichlid adaptive radiation in Lake Tanganyika.

Extended Data Fig. 2 | Alternative time-calibrated species trees of the cichlid adaptive radiation in Lake Tanganyika. Time-calibrated species tree based on **a,** the maximum-likelihood topology estimated with RAxML (Fig. 1), **b**, the topology estimated with ASTRAL, and **c**, the topology estimated with SNAPP. Time calibration used a relaxed-clock model in BEAST2, applied to a selected set of alignments with a total length of 30,738 bp.

Extended Data Fig. 3 | Individual-level phylogenies for the cichlid adaptive radiation in Lake Tanganyika. a, Maximum-likelihood tree inferred from nuclear SNPs. Node labels indicate the proportion of data subsets supporting a clade (equal to 1 for all nodes without labels). **b**, Mitochondrial phylogeny inferred from mitochondrial genomes using maximum-likelihood. Node labels represent bootstrap values (100 for all nodes without labels).

Extended Data Fig. 4 | Phenotyping of the specimens. a, Two-dimensional landmarks placed on X-ray images of the specimens. To quantify overall body shape we excluded landmark 16 (to minimise the effect of the orientation of the oral jaw). To analyse upper oral jaw morphology we used landmarks 1, 2, 16, and 21. **b**, Three-dimensional landmarks used to analyse lower pharyngeal jaw shape on μCT-scans of the heads. True landmarks are indicated in red, sliding semi-landmarks are indicated in blue. **c**, Body regions scored for presence/absence of pigmentation patterns.

Extended Data Fig. 5 | Morphospace and ecospace of the cichlid adaptive radiation in Lake Tanganyika. Scatter plots for each focal tribe (indicated with colours, see Fig. 1 for colour key) against the total morpho- and ecospace (grey). Species ranges are indicated with convex hulls. **a-c**, PC1 and PC2 of body shape, upper oral jaw morphology, and lower pharyngeal jaw shape, respectively. For shape changes associated with the respective PC-axis see Fig. 2 and for full species names see Extended Data Table 1. **d**, Stable C and N isotope compositions ($\delta^{15}N$ and $\delta^{13}C$ values). The additional plot shows $\delta^{15}N$ and $\delta^{13}C$ values of a baseline dataset which confirms the interpretability of the stable isotope N and C composition in Lake Tanganyika (see Supplementary Methods and Discussion). The last plot for each trait shows the size of the morpho- and

ecospace per tribe in relation to species numbers (body shape: Pearson's r = 0.91, df = 9, $P < 0.001$; upper oral jaw morphology: Pearson's r = 0.88, df = 9, $P < 0.001$; lower pharyngeal jaw shape: Pearson's r = 0.83, df = 9, $P = 0.002$; stable isotopes: Pearson's r = 0.88, df = 9, $P < 0.001$). Morphospace size was calculated as the square root of the convex hull area spanned by species means.

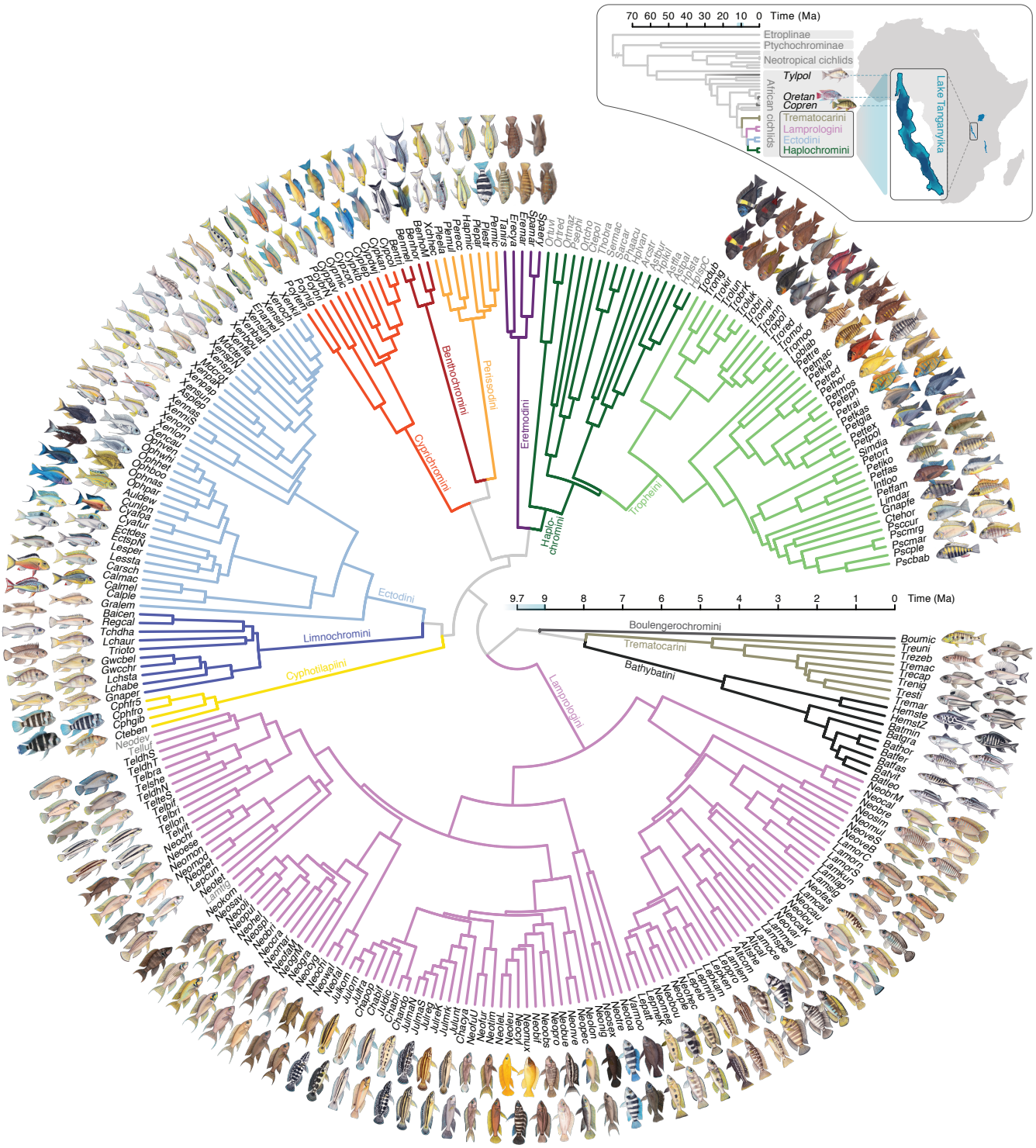
Extended Data Fig. 6 | PLS fit for each multivariate trait against the stable C and N isotope compositions ($\delta^{15}N$ and $\delta^{13}C$ values). Associated shape changes and loadings of the respective stable isotope projection are indicated next to the axes. Data points represent species means and are coloured according to tribe. **a**, Body shape; **b**, Upper oral jaw morphology; **c**, Lower pharyngeal jaw shape.

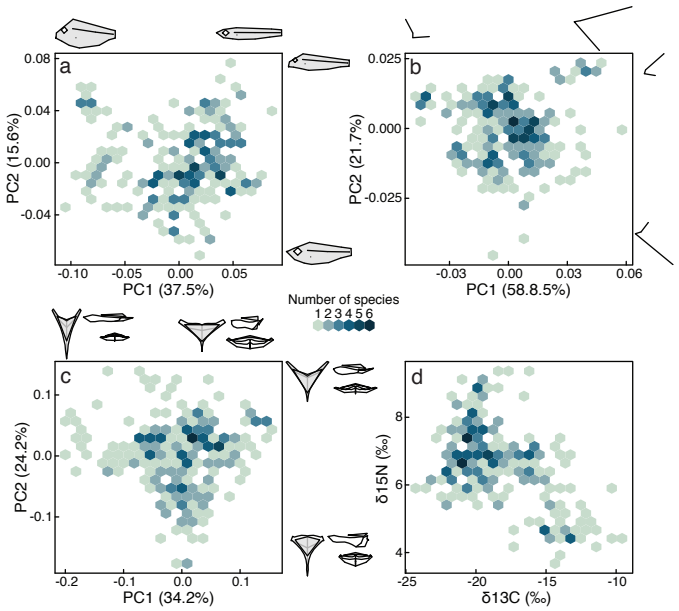
Extended Data Fig. 7 | Genome-wide statistical analyses. a, Proportion of the different classes of transposable elements (TE) among all TE for each tribe (one genome per species). **b**, Species means of dN (left) and dS (right) values over alignment length for each tribe. The boxes' centre lines show median, box limits show first and third quartiles, and whiskers show the 1.5× interquartile ranges. **c**, f_4-ratio statistics among species within each tribe in simulated data (tribe means are based on the mean across 20 simulations of each species triplet). Data points are coloured according to tribes; large points are tribe means, shown with 95% confidence intervals, small points represent species means and are only shown for group sizes < 40 species. To test for a correlation with species richness per tribe (log-transformed), we first calculated phylogenetic independent contrasts for each variable.

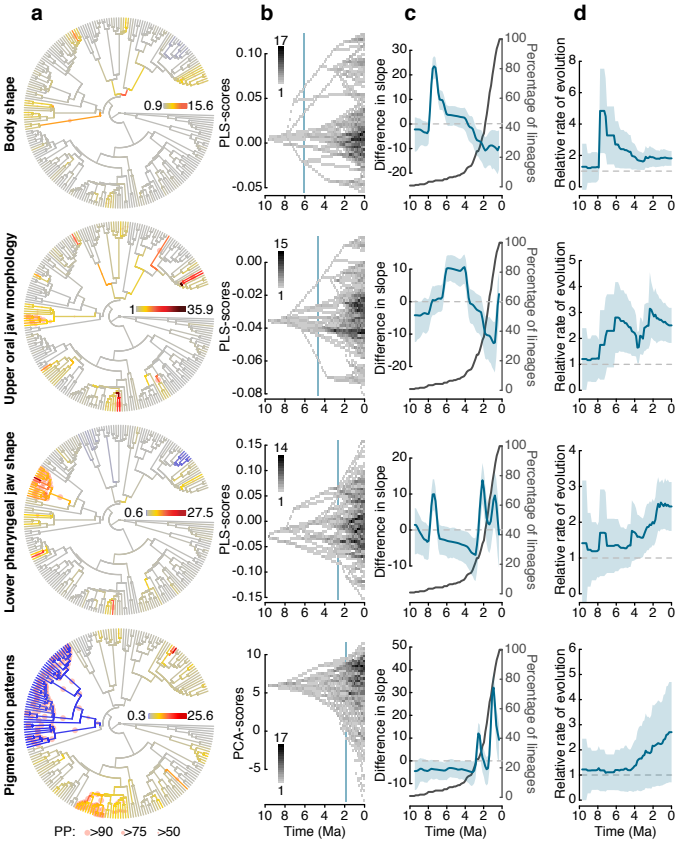
Extended Data Fig. 8 | Signals of introgression among Lake Tanganyika cichlid species. a, Maximum values of the f_4-ratio statistics between all pairs of species, derived from calculations across all combinations of species trios with *Tilapia sparrmanii* fixed as the outgroup. The f_4-ratio estimates the proportion of the genome affected by gene flow, all presented values are statistically significant ($P < 5\times10^{-5}$ after Benjamini-Hochberg correction for multiple testing). **b**, D_{tree}-statistics (hue) with corresponding P-value (log-transformed; saturation) based on a phylogenetic approach testing for asymmetry in the relationships of species trios in 1,272 local maximum-likelihood trees (see Supplementary Methods). The two different approaches uncovered little gene flow among the tribes (see Supplementary Discussion).

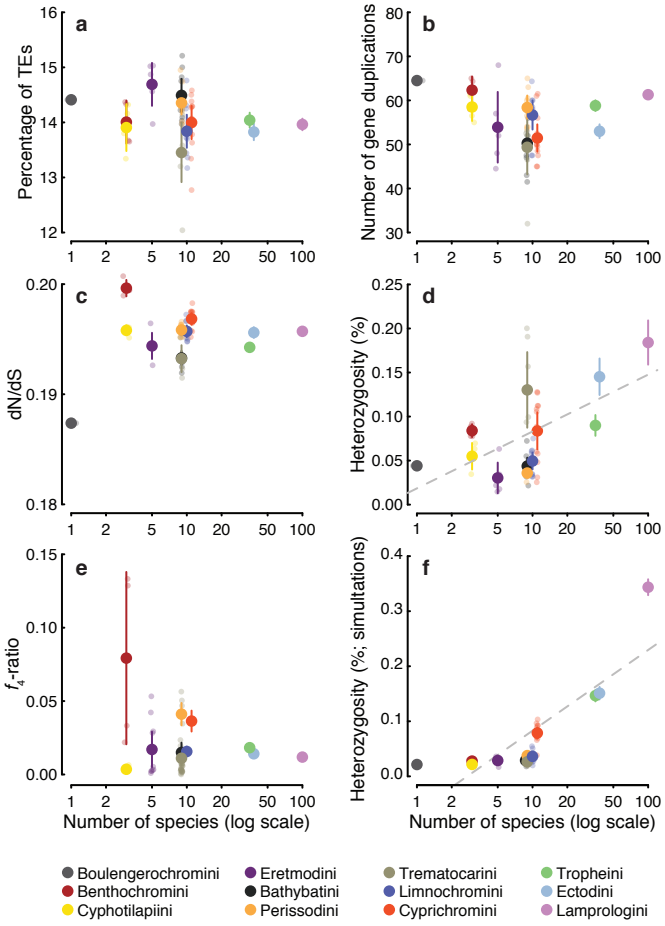
Extended Data Table 1 | Sample size information per species. For each analysis the total sample size is given whereas the number in brackets indicates the number of specimens used uniquely for the respective analysis. All genomes and raw sequences are available at NCBI under the BioProject accession number PRJNA550295. A full list of individual specimen vouchers including details on sampling location is provided as Supplementary Table 1. AMNH = American Museum of Natural History (New York, USA); MRAC = Royal Museum for Central Africa (Tervuren, Belgium); HHB = Private collection of one of the authors, H.H.B.

Extended Data Table 2 | Models of trait evolution. a, Comparison of model fits for different models of trait evolution and phylogenetic signal for each trait complex using three time-calibrated species trees with alternative topologies. **b**, Overview of the model fits and phylogenetic signal inferred using 100 trees sampled from the posterior distributions of the time calibrations for each of the three alternative tree topologies.


70 60 50 40 30 20 10 0 Time (Ma)
Etroplinae
Ptychochrominae
Neotropical cichlids
Tylpol
African cichlids
Oretan
Copren
Trematocarini
Lamprologini
Ectodini
Haplochromini
Lake Tanganyika
Benthochromini
Perissodini
Cyprichromini
Eretmodini
Tropheini
Haplo-chromini
Ectodini
Limnochromini
Cyphotilapiini
9.7 9 8 7 6 5 4 3 2 1 0 Time (Ma)
Boulengerochromini
Trematocarini
Bathybatini
Lamprologini


a
b
c
d
PC2 (15.6%)
PC1 (37.5%)
PC2 (21.7%)
PC1 (58.8%)
PC2 (24.2%)
PC1 (34.2%)
δ15N (‰)
δ13C (‰)
Number of species
1 2 3 4 5 6


a
b
c
d
Body shape
Upper oral jaw morphology
Lower pharyngeal jaw shape
Pigmentation patterns
0.9
15.6
1
35.9
0.6
27.5
0.3
25.6
PP: >90 >75 >50
PLS-scores
PCA-scores
Difference in slope
Percentage of lineages
Relative rate of evolution
Time (Ma)


a
Percentage of TEs
16
15
14
13
12
b
Number of gene duplications
80
70
60
50
40
30
c
dN/dS
0.20
0.19
0.18
d
Heterozygosity (%)
0.25
0.20
0.15
0.10
0.05
0.00
e
f_4-ratio
0.15
0.10
0.05
0.00
f
Heterozygosity (%; simultations)
0.4
0.3
0.2
0.1
0.0
1 2 5 10 20 50 100
Number of species (log scale)
Boulengerochromini
Benthochromini
Cyphotilapiini
Eretmodini
Bathybatini
Perissodini
Trematocarini
Limnochromini
Cyprichromini
Tropheini
Ectodini
Lamprologini