nature research

https://doi.org/10.1038/s41586-020-2652-7

Supplementary information

The ecological and genomic basis of explosive adaptive radiation

In the format provided by the authors and unedited

Supplementary Information

1. Phylogenetics

1.1 Alignment

We used the PHLAWD pipeline (31) to generate alignments for cichlids with sequences available on GenBank, targeting sixteen mitochondrial loci (15,005 bp) and thirteen nuclear protein coding genes (11,589 bp). We supplemented sequences obtained using PHLAWD with sequences extracted from cichlid mitochondrial genomes using R package 'AnnotationBustR' (32) and obtained additional cytochrome oxidase I (*COI*) sequence from the Barcode of Life Database (33), for a total of 30 loci consisting of 26,594 bp for 1,015 cichlid taxa and 2 outgroup taxa, resulting in a matrix that is 20.2% complete (Supplementary Table S1, Appendix 1). These 30 loci were chosen as they have been commonly used for phylogenetic reconstruction of cichlid relationships (34), have sufficient coverage across cichlid diversity, and are not obviously associated with a specific phenotype known to be heavily under selection across cichlid diversity.

We also included meristic data from the taxonomic literature for all currently valid cichlid species described prior to 2019 (n=1712, Appendix 2). We recorded minimum and maximum values for four traits: dorsal spine count, dorsal soft ray count, anal spine count, and anal soft ray count. We note that our full alignment includes the 167 formally described species out of the currently >500 known species of Lake Victoria haplochromines, likely biasing our estimates of explosive speciation downwards for this clade.

	Species	rRNA12S	rRNA16S	ATP6	ATP8	PLAGL2	coi	coii	coiii	DLoop	cytb	enc1	glyt	gpr85	h3	myh6	nd1	nd2	nd3	nd4	nd4l	nd5	nd6	ptchd4	rag1e3	rag2	S7i1	snx33	tbr1	tmo4c4	zic1
Astronotini	2	1	2	1	1	0	2	1	1	1	2	0	0	0	1	0	1	1	1	1	1	1	1	0	1	0	0	0	0	1	0
Bathybatini	9	1	2	0	0	1	4	0	0	8	8	3	1	1	0	2	0	9	0	1	0	0	0	3	2	0	2	3	2	2	1
Boulengerochromini	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
Chaetobranchini	4	2	3	0	0	0	3	0	0	0	2	2	0	2	2	0	0	0	0	2	0	0	0	0	0	1	1	0	0	1	0
Chromidotilapini	58	13	34	0	0	2	14	0	0	2	15	37	1	2	1	4	0	44	0	3	0	0	0	39	14	2	35	39	5	2	2
Cichlasomatini	78	13	42	9	10	3	23	9	9	7	46	25	3	21	10	8	9	10	9	24	9	8	8	6	37	19	39	3	9	9	4
Cichlini	19	3	8	16	16	1	7	2	2	15	13	3	1	2	1	2	2	3	2	6	2	2	2	3	1	3	3	1	2	4	1
Ectodini	33	0	5	0	0	3	20	0	0	28	28	6	2	3	0	3	0	27	0	1	0	0	0	6	11	0	5	6	6	5	3
Eretmodini	5	1	1	0	0	2	4	0	0	4	4	2	2	2	0	2	0	4	0	1	0	0	0	2	3	0	1	2	2	1	2
Etroplini	16	3	13	3	3	3	13	3	3	3	5	4	3	3	13	4	3	4	3	4	3	3	3	4	3	1	3	3	4	13	3
Geophagini	280	8	77	12	12	5	57	5	5	10	101	38	5	36	13	2	5	34	5	38	5	3	3	10	6	41	64	7	11	31	7
Haplochromini	726	40	57	37	37	14	70	37	36	280	126	30	11	14	4	21	36	226	36	40	37	36	36	34	37	0	79	30	29	26	14
Hemichromini	11	1	6	0	0	3	7	0	0	2	7	5	1	3	1	6	0	5	0	2	0	0	0	4	3	0	2	3	5	2	3
Heroini	176	19	86	18	18	29	77	17	17	27	120	57	18	55	11	22	16	44	18	65	16	17	15	35	63	58	93	21	23	37	31
Heterochromini	1	1	1	0	0	1	1	0	0	0	1	1	1	1	1	1	0	1	0	1	0	0	0	1	1	1	1	1	1	1	1
Lamprologini	92	1	9	24	24	3	39	1	1	68	47	14	2	3	1	5	1	78	1	4	1	1	1	14	37	0	21	10	14	12	3
Limnochromini sensu lato*	32	1	6	1	1	8	7	1	1	29	27	12	7	8	0	11	1	31	1	5	1	1	1	12	13	0	8	12	12	7	8
Oreochromini	63	16	19	5	5	7	17	5	5	32	25	14	7	6	3	7	5	40	5	6	5	4	5	15	7	0	12	15	7	15	7
Pelmatochromini	4	3	3	0	0	0	2	0	0	0	1	3	0	0	1	2	0	3	0	1	0	0	0	3	0	0	3	3	1	3	0
Ptychochromini	16	3	13	2	2	3	13	2	2	2	5	3	3	3	12	4	2	4	3	3	2	2	2	4	3	1	1	3	4	13	3

Tilapini sensu lato	59	33	35	2	2	5	16	2	2	13	10	26	6	6	3	8	2	42	2	4	2	2	2	33	7	3	35	34	7	33	6
Trematocarini	9	0	0	0	0	0	0	0	0	2	2	1	0	0	0	0	0	5	0	0	0	0	0	1	2	0	1	1	1	1	0
Tylochromini	18	1	3	1	1	1	5	1	1	5	2	3	2	2	1	2	1	2	1	1	1	1	1	3	3	0	1	2	3	2	2

Supplementary Table S1. Coverage for the 30 markers used across our cichlid phylogeny across cichlid tribes. Mitochondrial genes are shaded in grey.

* includes all species in the following genera: *Baileychromis*, *Benthochromis*, *Cyphotilapia*, *Cyprichromis*, *Gnathochromis*, *Greenwoodochromis*, *Haplotaxodon*, *Limnochromis*, *Paracyprichromis*, *Perissodus*, *Plecodus*, *Reganochromis*, *Tangachromis*, *Triglachromis*, and *Xenochromis*
** includes all species in the following genera: *Chilochromis*, *Coelotilapia*, *Congolapia*, *Coptodon*, *Etia*, *Gobiocichla*, *Heterotilapia*, *Paragobiocichla*, *Pelmatolapia*, *Steatocranus*, and *Tilapia*.

1.2 Constraint tree

We generated an extensive constraint tree based on phylogenomic data from recent RADseq or UCE phylogenies (35-39). Nodes with support at 95% or higher and not contradicted by other phylogenomic datasets were introduced as constraints and converted into a constraint tree with the as.phylo() function in ape (40). Genera never previously split up by multimarker molecular phylogenies were utilized as constraints. Within genera, species groups recognized by taxonomists, particularly complexes recently split by taxonomists from one species into several species, were constrained unless these assignments were contradicted by a molecular phylogeny.

1.3 Phylogenetic reconstruction

We partitioned all coding genes by codon; non-coding elements (eg. mitochondrial control region) were assigned to individual partitions. Meristic counts for our four count variables were split into 4 partitions. We used RAxML-HPC v.8 (41) using a GTR+G model, our constraint tree, and the default hill-climbing algorithm to generate a phylogeny. We then used 'treePL' (42) in conjunction with a set of cichlid fossil calibrations (Supplementary Table S2) derived from a recent paper (43) as well as a broad set of lake age calibrations to time calibrate the phylogeny with the optimal smoothing parameter identified using cross-validation. Cichlid lake and radiation ages have historically been subject to much controversy within the field (43); we therefore selected broad constraints that represent several differing perspectives on the age of these radiations. For example, Lake Tanganyika has a constraint between 10-20 million years, and Lake Victoria has a constraint of 10,000 to 100,000 years.

	Min (MY)	Max (MY)	Taxon1	Taxon2
Family crown age	46	81	*Heterochromis multidens*	*Etroplus suratensis*
Tribe Hemichromini and Pelmatochromini	46	81	*Hemichromis fasciatus*	*Pterochromis congicus*

Tribe Tylochromini	34	81	*Tylochromis polylepis*	*Tylochromis intermedius*
Genus Gymnogeophagus	39.9	81	*Gymnogeophagus rhabdotus*	*Gymnogeophagus balzanii*
Crenicichlines, (incl. stem)	39.9	81	*Acarichthys heckelii*	*Crenicichla sveni*
Tribe Cichlini	5.332	81	*Cichla temensis*	*Cichla ocellaris*
Tribes Heroini + stem	39.9	81	*Acaronia nassa*	*Hypselecara temporalis*
Genus Cichlasoma and relatives	5.332	39.9	*Krobia xinguensis*	*Aequidens metae*
Tribe Cichlasomatini	30	81	*Krobia xinguensis*	*Ivanacara adoketa*
Nandopsine cichlids	5.332	39.9	*Nandopsis tetracanthus*	*Trichromis salvini*
Genus Coptodon	2.588	81	*Coptodon zillii*	*Coptodon discolor*
Tribe Oreochromini	9.3	46	*Oreochromis niloticus*	*Danakilia dinicolai*
Malawi open water clade	0.1	1.2	*Rhamphochromis longiceps*	*Diplotaxodon limnothrissa*
Malawi mbuna species flock	0.1	1.2	*Alticorpus mentale*	*Mylochromis mola*
Malawi 'hap' species flock	0.1	1.2	*Genyochromis mento*	*Labidochromis gigas*
Barombi Mbo, Stomatepia	0.0001	1	*Stomatepia pindu*	*Stomatepia mariae*
Barombi Mbo, Sarotherodon	0.0001	1	*Sarotherodon steinbachi*	*Sarotherodon caroli*
Barombi Mbo, others	0.0001	1	*Pungu maclareni*	*Konia eisentrauti*
Kinneret species flock	0.0001	2.6	*Tristramella sacra*	*Tristramella simonis*

Tanganyika	10	20	*Neolamprologus gracilis*	*Labidochromis gigas*
Xiloa species flock	0.0001	0.006	*Amphilophus amarillo*	*Amphilophus sagittae*
Apoyo species flock	0.0001	0.024	*Amphilophus zaliosus*	*Amphilophus chancho*
Victoria species flock	0.015	0.1	*Haplochromis arcanus*	*Neochromis rufocaudalis*

Supplementary Table S2. Time constraints used as input for TreePL, along with specific taxa used to identify the MRCA defining the constraint. Time minima and maxima are in units of millions of years. Note that TreePL requires identification of maxima and minima.

1.4 Verification that meristic characters possess strong phylogenetic signal

To verify that meristic characters contain suitably strong phylogenetic signal, we performed the same pipeline as above, but excluding meristic characters, producing a phylogeny of 1000 cichlids. We then used the 'phylosig' function in R package 'phytools' to calculate the phylogenetic signal of each meristic trait, using lambda and a likelihood ratio test. All values of lambda were between 0.94 and 0.999 for every trait, and the p-value of the likelihood ratio tests were all below $p<0.001$ (Supplementary Table S3), suggesting an extremely strong phylogenetic signal for all traits. This is expected as closely related species generally exhibit similar spine and ray counts for both the dorsal and anal fin regardless of ecological divergence, whereas ecologically equivalent species belonging to older cichlid clades often exhibit great disparity in these traits.

	Phylogenetic signal (lambda)	**Likelihood**	**Likelihood, lambda=0.0**	**P-value**
Dorsal fin spines, min	0.953	-1707.362	-2299.956	1.01e-259
Dorsal fin spines, max	0.949	-1731.852	-2303.491	1.30e-250
Dorsal fin rays, min	0.973	-1667.007	-2422.689	0
Dorsal fin rays, max	0.972	-1744.448	-2493.265	0
Anal fin spines, min	0.999	-938.131	-1842.754	0
Anal fin spines, max	0.999	-1107.417	-2094.739	0
Anal fin rays, min	0.970	-1550.336	-2223.014	1.57e-294
Anal fin rays, max	0.972	-1662.672	-2325.842	2.13-290

Supplementary Table S3. Phylogenetic signal for each of our four meristic variables calculated with 'phylosig' from R package 'phytools' on a phylogeny of cichlids produced solely with molecular data (n=1000 species).

1.5 Geospatial data

We utilized the GBIF database (44) with the R package 'rgbif' to access all cichlid coordinate information, as well as the country for each coordinate pair. We then used Fishbase using the R package 'rfishbase' (45) to filter out coordinate data that did not correspond to countries where the species in question was listed as native. Many cichlid radiations, particularly Victoria and Malawi, do not have reliable locality data available for many species. Therefore, for endemic lake species, we ignored coordinate data and used a set of coordinates defining the geographic expanse of the lake, then assigned species endemic to those lakes the full set of coordinates for that lake.

For species without GBIF coordinate data, we first examined the taxonomic literature for coordinates, and used those where available (Appendix 3). For species with occurrence data but no specific coordinates, we examined each potential locality using Google Earth and recorded coordinates for locations matched approximately to the collection location, eg. the closest river site next to the city listed in as the collection location (Appendix 3). We utilized the Freshwater Ecoregions of the World (46) database to assign our fish as present or absent across freshwater ecoregions (Appendix 4), recording the ecoregion codes occupied by its coordinate pairs. Lake endemics were always assigned to their appropriate lake ecoregion.

1.5 Diversification variables

We generated a list of eleven variables (Supplementary Table S4) associated with diversification in previous studies of adaptive radiation: the presence of male ornamentation (19, Appendix 2), polygamous mating system vs biparental care (19, Appendix 2), water depth (19, 47), range size, predator presence (18), elevation (48), annual temperature and rainfall (49), latitude (19), and body size. As only binary variables can be analyzed across our full suite of models, we have endeavoured to produce reasonable categorizations of each, though we recognize that many are best thought of as a continuum. For variables calculated from GIS coordinates, we took the median value of each environmental variable for every occurrence point for that species.

	Scored as zero	Scored as one	Sources
Biotic variables			
Polygamy	Males and females stay together after mating to perform parental care.	Males mate with multiple females. The majority of information derives from personal expertise of SRB and OS derived both from cichlid species personally bred in	66-68

		aquaria or observed in the field supplemented by data compiled in Wagner et al. 2012, and finally by additional data from the aquarium and ichthyological literature (66-68). Close relatives were assumed to possess the mating characteristics of their congeners or species group with available data.	
Male ornaments	No obvious sexually dimorphic traits involved in display	Males exhibit sexually dimorphic traits (eg. haplochromine eggspots), displayed to either females or conspecific males	66-68
Small body size	maximum standard length greater than or equal to mean minus one standard deviation (5 cm)	maximum standard length less than mean minus one standard deviation (5 cm)	Taxonomic literature, Fishbase
Large body size	maximum standard length less than or equal to mean minus one standard deviation (21.7 cm)	maximum standard length greater than mean plus one standard deviation (21.7 cm)	Taxonomic literature, Fishbase
Predator presence	No large visual predatory fishes present (eg. *Hydrocynus*, *Hoplias*)	Species' range includes large-bodied visual predatory fishes capable of consuming adult cichlids	GIS, FEOW
Abiotic variables			
High elevation	Elevation less than 1,000 m asl	Elevation greater than 1,000 m	GIS, SRTM
High latitude	Latitude within tropics (23.5N-23.5S)	Latitude outside tropics	GIS
High rainfall habitat	BIO12 (annual precipitation) less than or equal to 1,680 mm	BIO12 (annual precipitation) greater than 1680 mm	GIS
Low rainfall habitat	BIO12 (annual precipitation) greater than or equal to 500 mm	BIO12 (annual precipitation) less than 500 mm	GIS
Endemism/small range size	Species occurs across multiple ecoregions	All occurrences within one freshwater ecoregion	GIS, FEOW
Large depth gradient	Species occurs within ecoregion, or if lake endemic, within specific lake, with water depths less than or equal to 50m	Species occurs within ecoregion/lake with water depths >50 m	GIS, RBD, GLD

Supplementary Table S4. Biotic and abiotic predictors for speciation rate.

1.6 Speciation rate

We calculated tip diversification (DR) for each species using the 'DRstatistic' function in the 'FiSSE' package (17). Previously, tip diversification was thought to represent diversification (speciation minus extinction), but more recent work has suggested it is primarily associated with speciation rate (14). We then tested each binary predictor using FiSSE, using a tolerance of 0.1 and an M-K rate type (Supplementary Table S5).

	P-value	Lambda0	Lambda1	Number of changes
Depth	0.702	0.228	3.719	204
Predators *	0.011	6.031	0.232	50
Polygamy	0.970	0.221	4.760	51
Male ornamentation	0.960	0.227	5.058	26
Endemism	0.713	0.205	3.767	244
Large body size	0.386	3.159	0.996	127
Small body size	0.396	3.070	0.537	55
High altitude	0.930	1.283	6.890	56
High latitude (subtropical)	0.673	0.210	3.022	28
Rainforest	0.129	4.107	0.208	133
Arid	0.455	2.929	0.354	18

Supplementary Table S5: Parameter estimates and p-values for FiSSE models run on each of eleven binary variables across our full cichlid tree (n=1712). Highlighted cells indicate traits with an uncorrected p-value < 0.05.

1.7 Bayesian regression

We used STAN (16) via the 'brms' package (50) to generate a phylogenetically corrected Bayesian regression model of speciation rate predicted from our eleven binary biotic and abiotic

variables (Supplementary Table S6). Speciation rate was modeled with a gamma family. We used the R package ‘ape’ (Paradis et al. 2004) to generate a variance-covariance matrix from the phylogeny, then included it as a random factor in our model. We utilized a normal distribution centered on zero with standard deviation of 1 for the slope and intercept priors, and a gamma prior for the random effect term associated with phylogenetic covariance. The model was run for 4,000 generations using 4 chains. We note that the Hamiltonian Monte Carlo utilized by STAN requires many fewer generations than traditional samplers, though each generation is more computationally intensive. Convergence of the four chains was assessed using the scale reduction factor (Rhat) statistic (51).

	Estimate	Est. Error	95% CI, lower bound	95% CI, upper bound	Rhat	ESS
Intercept	-1.58	0.38	-2.33	-0.83	1	3,018
Depth	0.13	0.09	-0.04	0.29	1	6,756
Predators *	-0.9	0.13	-1.15	-0.64	1	5,347
Polygamy	0.04	0.14	-0.24	0.33	1	5,979
Male ornamentation	0.14	0.18	-0.22	0.48	1	4,576
Endemism	0.13	0.08	-0.03	0.29	1	6,839
Large body size	-0.08	0.09	-0.26	0.11	1	7,095
Small body size	0.01	0.15	-0.27	0.3	1	8,364
High altitude	0.03	0.12	-0.2	0.27	1	6,519
High latitude (subtropical)	-0.06	0.19	-0.44	0.3	1	5,203
Rainforest	-0.06	0.09	-0.24	0.12	1	7,604
Arid	-0.6	0.24	-1.05	-0.12	1	7,718

Supplementary Table S6: Summary of the posterior distribution of population-level effects for our STAN speciation rate regression for all described (n=1712) cichlid species.

1.8 Additional discussion on cichlid speciation and extinction.

The hypothesis that predators negatively influence cichlid diversity and diversification was put forward over fifty years ago (18) and previously received modest support in analyses of African lake cichlids (19). We do not recover the effect of latitude on speciation rate shown in a previous analysis on African lake cichlids, likely because the previous analysis (19) was restricted to Africa and the Middle East, where aridity is highly positively correlated with latitude, something that is not the case in the Americas. We detect only a mild effect of water depth, likely due to the lack of much speciation in some large rivers with large depth gradients such as the Amazon and the Congo, probably explained by very poor visibility and the dominance of electrosensitive teleost lineages with sensory systems better adapted to these deep river habitats.

Especially telling are some within-drainage system comparisons: The Lake Victoria region, which originally lacked large apex predators (native *Lates* and *Hydrocynus* occurred only in Lake Albert), was colonized by many sexually dimorphic cichlid lineages, including four different oreochromine species (*Oreochromis esculentus*, *O. variabilis*, *O. leucostictus*, *O. niloticus eduardianus*) and three distantly related haplochromine lineages (*Astatoreochromis*, *Pseudocrenilabrus*, *Astatotilapia*/*Thoracochromis*), sexually dimorphic too. Despite all lineages experiencing the same combination of biotic and abiotic factors (52), just one haplochromine lineage (a lineage of hybrid origin between Congolese *Astatotilapia* and Upper Nile *Thoracochromis,* 20) underwent adaptive radiations in every lake in the region ('Lake Victoria Region Superflock'), and generated about 500 endemic species in Lake Victoria alone within the past 15,000 years.

1.9 Hidden-state speciation and extinction:

We also used the 'hisse' (hidden state speciation and extinction) R package (21) to test the effect of the above defined binary traits on state-dependent diversification. This method employs a hidden Markov model that assumes a hidden unobserved state that may be responsible for heterogeneity in rates of speciation, extinction, and transition rates amongst states while also allowing for the creation of null models to test for character-independent diversification. We constructed a total of 24 models(Supplementary Table S7) in which the number of free parameters to estimate varied. Of our models, four were equivalent to BiSSE models, lacking hidden states, seventeen were HiSSE models in which hidden states were incorporated into the model, and three were character-independent diversification (CID) models. Due to known difficulties with estimating transition rates between states in state-dependent speciation and extinction models, we constrained transitions to be equal across states. We also eliminated dual transitions between hidden and observed states from our models, disallowing instantaneous simultaneous changes in these states. The seventeen HiSSE models involved configurations where certain transitions between states were set to zero. This richer suite of HiSSE models is necessary to accommodate the possible dynamics of the hidden state.

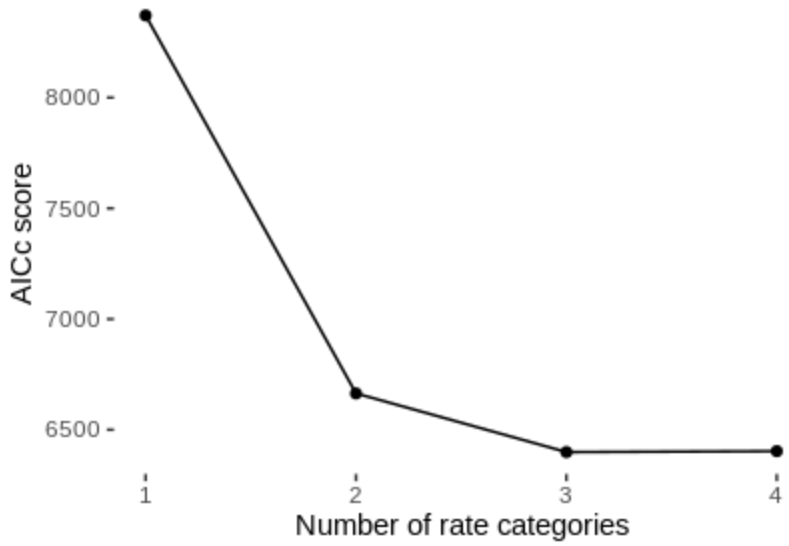
	Predator presence	Endemism	Small body size	Large body size	Male ornamentation	Polygamy	Rainforest	Arid	Elevation	Depth	Latitude
BiSSE: q's equal	7,663	9,273	8,909	9,401	7,746	8,040	9,044	8,603	8,879	9,379	8,687
BiSSE: ε's equal, q's equal	7,744	9,277	8,917	9,438	7,872	8,125	9,024	8,621	8,892	9,375	8,681
BiSSE: τ's equal, q's equal	8,806	9,778	8,909	9,428	8,543	8,837	9,414	8,602	8,996	9,679	8,697
BiSSE τ's equal, ε's equal	8,908	9,855	8,926	9,492	8,662	8,909	9,424	8,619	8,998	9,731	8,696
HiSSE: q's equal	7,105	8,030	7,267	7,803	6,901	7,123	7,699	6,965	7,174	7,839	7,040
HiSSE: ε's equal, q's equal	7,154	8,322	7,275	7,922	6,893	7,157	7,715	6,977	7,193	8,244	7,033
HiSSE: τ0A= τ1A= τ0B, ε0A=ε1A=ε0B, q's equal	7,599	8,468	8,871	9,367	6,983	7,272	8,924	8,597	8,166	8,109	7,050
HiSSE: τ0A= τ1A= τ0B, ε's equal, q's equal	7,777	8,225	8,921	9379	6,992	7,397	8,925	8,620	8,253	8,279	7,068
HiSSE: q0B1B=0,q1B0B=0, other q's equal	7,043	7,976	7,179	7,375	6,896	7,033	7,632	6,211	7,176	7,840	6,989
HiSSE: ε's equal, q0B1B=0, q1B0B=0, other q's equal	7,282	8,127	7,168	7,548	7,019	7,200	7,675	7,109	7,202	7,881	7,252
HiSSE: τ0A= τ1A= τ0B, ε0A=ε1A=ε0B, q0B1B=0, q1B0B=0, other q's equal	7,527	8,085	8,856	9,291	6,968	7,191	8,810	8,597	8,187	8,038	7,049
HiSSE: τ0A= τ1A= τ0B, ε's equal, q0B1B=0, q1B0B=0, other q's equal	7,677	8,103	8,854	9,296	6,986	7,189	8,802	8,596	8,212	8,037	7,042
HiSSE: τ0A= τ1A, ε0A=ε1A, q's equal	7,512	8,016	8,855	7,613	6,959	7,174	8,915	8,590	8,126	7,948	7,041
HiSSE: τ0A= τ1A, ε's equals, q's equal	7,630	8,199	8,869	9,360	6,954	7,199	8,911	8,616	8,238	8,274	7,094
HiSSE: τ0A= τ1A, ε0A=ε1A, q0B1B=0,q1B0B=0, other q's equal	7,495	7,953	8,862	9,310	7,006	7,150	8,798	8,590	8,086	7,848	7,158
HiSSE: τ0A= τ1A, ε's equals, q0B1B=0,q1B0B=0,	7,486	8,120	8,853	9,290	7,141	7,295	8,784	8,594	8,164	7,987	7,246

other q's equal											
HiSSE: τ0A= τ0B, ε0A=ε0B, q's equal	7,100	7,955	7,289	7,902	7,003	7,231	7,693	6,947	7,353	7,964	7,094
	Predator presence	Endemism	Small body size	Large body size	Male ornamentation	Polygamy	Rainforest	Arid	Elevation	Depth	Latitude
HiSSE: τ0A= τ0B, ε's, q's equal	7,408	8,493	7,290	7,932	6,970	7,252	7,743	6,964	7,352	7,910	7,061
HiSSE: τ0A= τ0B, ε0A=ε0B, q0B1B=0,q1B0B=0, other q's equal	7,059	8,037	7,375	8,182	6,862	7,082	7,649	6,993	7,228	7,871	6,792
HiSSE: τ0A= τ0B, ε's equal, q0B1B=0,q1B0B=0, other q's equal	7,141	8,279	7,395	8,144	6,906	7,103	7,675	7,022	7,282	8,261	7,420
HiSSE: τ's equal, q's equal	8,268	9,604	8,472	9,378	8,517	8,748	9,269	8,043	8,911	9,518	8,071
CID-2: q's equal	7,282	7,872	7,287	7,865	7,032	7,261	7,786	6,967	7,365	7,853	7,069
CID-4: q's equal	6,872	7,178	6,951	6,930	6,617	6,927	6,835	6,558	7,002	6,982	6,637
CID-4: ε's equal, q's equal	7,069	7,715	7,090	7,757	6,806	7,073	7,716	6,623	7,167	8,203	6,833

Supplementary Table S7: AIC scores of our seventeen BiSSE, HiSSE, CID-2, and CID-4 models on our eleven binary traits. Highlighted cells indicate the model with greater than 99.9%+ weight.

1.10 Missing state speciation and extinction:

We fit a series of MiSSE (missing state speciation and extinction) with the 'hisse' R package. We used the MiSSEGreedy function to identify the number of rate shifts across our phylogeny while stopping the iterative search after models increased by at least 10 AICc units for three consecutive iterations. Our best fit model contained three states (Supplementary Figure S1). These three states correspond to a background rate, a uniquely fast speciation rate for Lake Victoria, as well as a moderately fast rate shared by Lake Xiloa in Nicaragua, Lake Malawi, and the Lake Victoria Region Superflock excluding Lake Victoria.



Supplementary Figure S1. Plot of AIC scores, indicating best fit of a model with three rate categories, for a series of multiple-state speciation and extinction (MiSSE) models fit to the cichlid dataset.

2. Genomics

2.1 Variant calling

For our whole genome analyses, we utilized both existing short-read data (section 2.2) as well as newly sequenced short-read data (section 2.3). We used the 'McCortex' pipeline (22) to assemble and align de Bruijn graphs from the short reads of each sample, following the standard workflow for the breakpoint caller used for large structural variants. Briefly, each sample was assembled into its own de Brujin graph at a k-mer size of 31, then low-coverage bubbles were cleaned using default parameters. We then used the 'bubble calling' pipeline to identify possible variants associated with bubbles in each graph relative to a reference genome using the 'bubbles' command in McCortex via identification of k-mers in each sample unique to that variant. We then used the 'vcfcov' function of McCortex to calculate the coverage associated with the reference or alternate allele associated with each variant in our sample.

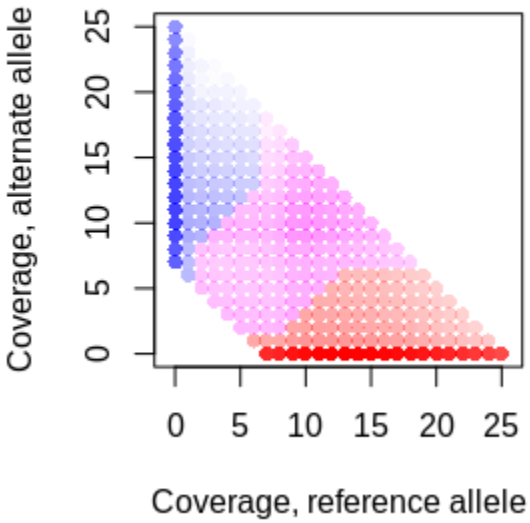
In order to go from coverage to SNP calling, we eschewed the final portion of the McCortex pipeline, as sites with no coverage of either the reference or alternate were being assigned as homozygous reference. Instead, we utilized a machine learning approach, hard competitive learning, to convert coverage information into variant calls for each sample. Sites with total coverage greater than 2.5 times the median threshold for sites with some coverage of both the reference and the alternate allele were called as missing, as were sites where total coverage of reference and alternate was less than half of the median total coverage of each site across the sample after removal of high coverage sites.

We then used the 'counts' function from the 'dplyr' package to calculate the number of every possible combination of reference and alternate counts of each site, creating a topological map of genotype density across each sample (Supplementary Figure S2). We then used this topological genotype map in conjunction with hard competitive learning, a type of self-organizing map, using the 'cclust' function in R package 'flexclust', using 3 clusters (homozygous reference, heterozygous, homozygous alternate), manhattan distance, weights corresponding to the count

of each genotype combination, 5000 total iterations, tolerance of 1e-4, gamma of 1, and an exponential method. Starting positions for the three clusters were set as the following.

Homozygous reference: (median value of reference counts for sites with an alternate allele count of zero, zero)
Homozygous alternate: (zero, median value of alternate counts for sites with a reference allele count of zero)
Heterozygous: (median value of the reference count for sites with nonzero reference and alternate counts, median value of the alternate count for sites with nonzero reference and alternate counts)

Because the clustering algorithm assigns every site as homozygous reference, heterozygous, and homozygous alternate, we treated borders, eg. count combinations adjacent to count combinations called differently, as missing data.



Supplementary Figure S2. Genotype topography of an example genome from our dataset, *Harpagochromis cavifrons*. Density plot showing counts of the reference and alternate allele for 19,711,572 total SNPs and indels. Each dot represents a particular combination of the count reference and the count of alternate alleles. Blue dots indicate counts of reference and alternate assigned as homozygous for the alternate allele, purple dots indicate counts of reference and alternate assigned as heterozygotes, and red dots indicate counts of reference and alternate alleles assigned as homozygous reference.

2.2 Genomic basis of speciation rate variation

We examine whole genome sequences from two species from all cichlid lake species flocks with more than two described species and some whole genome data available. We chose only two species per flock to avoid biasing indel discovery in favor of species-rich flocks, especially given the difficulty of properly correcting for the number of comparisons due to phylogenetic

structure within flocks. Each set of two species was chosen to maximize ecological divergence within that radiation; brief justifications of each choice are enclosed below.

For Lake Ejagham, which has the smallest number of endemic species, we selected the two most ecologically and morphologically distinct species, the small elongate *Coptodon fusiformis* and the large, more piscivorous *Coptodon ejagham* (53). For Lake Barombi Mbo, we selected a large elongate species, *Stomatepia mariae*, the only member of the radiation known to feed on other fishes, and *Pungu maclareni*, the only member of the radiation known to feed by grazing and scraping sponges and rocks (54). Many of the other species in the radiation possess more generalized phytoplankton and insect diets typical of their riverine relatives in genus *Sarotherodon* (54). For Lake Tanganyika, we were limited to the *Neolamprologus brichardi* clade of plankton-feeders, the only group with whole genome Illumina sequencing data currently available. Within this clade, we selected a species with a larger mouth (*N. olivaceous*) and a species with a smaller mouth (*N. gracilis*); we suspect any pair of the five available genomes would have sufficed. While it would have been ideal to include more Tanganyikan species, we note that most ecological transitions within the lake are over 5 million years old, far older than any of our other comparisons.

Within the Malawi open-water clade, which contains only two major clades, the largely piscivorous *Rhamphochromis* and the piscivorous and planktivorous *Diplotaxodon*, we selected the most elongate large piscivore, *Rhamphochromis esox*, and a small planktivore, *Diplotaxodon limnothrissa* (55). Within the 'hap' or 'sand-dwelling' Malawi clade, we selected one of very few strict herbivores, *Hemitilapia oxyrhynchus*, which feeds by scraping algae from *Vallisneria* leaves, and a large piscivore species, *Tyrannochromis nigriventer* (55). While Malawi contains many piscivorous cichlids, those in genus *Tyrannochromis* are some of the largest and most morphologically specialized, and some of the few known to feed on adult cichlids of other species, rather than fry and juveniles (55). Within the Malawi mbuna clade, we selected the scale-biting *Genyochromis mento*, the only currently available genome for a high trophic level mbuna, and the algae scraping *Tropheops tropheops* (56). The mbuna clade contains a large number of specialized herbivores; *Tropheops* was selected for its extremely small jaws and therefore greater morphological divergence relative to *Genyochromis*. For Lakes Kivu and Victoria, we selected species previously identified as being on opposite ends of an ecomorphological continuum (57). Radiations where only pooled whole-genome sequence data was available (eg. Nicaraguan *Amphilophus*) were excluded, as were lakes with species pairs instead of species flocks. All genomes except Victorian cichlids were taken from previous studies (Supplementary Table S8); sequencing procedures for Victorians are described in the following section.

Adaptive radiation	**Species**	**Accession**
Barombi Mbo	*Stomatepia mariae*	SRR7591517
Barombi Mbo	*Pungu maclareni*	SRR7591525

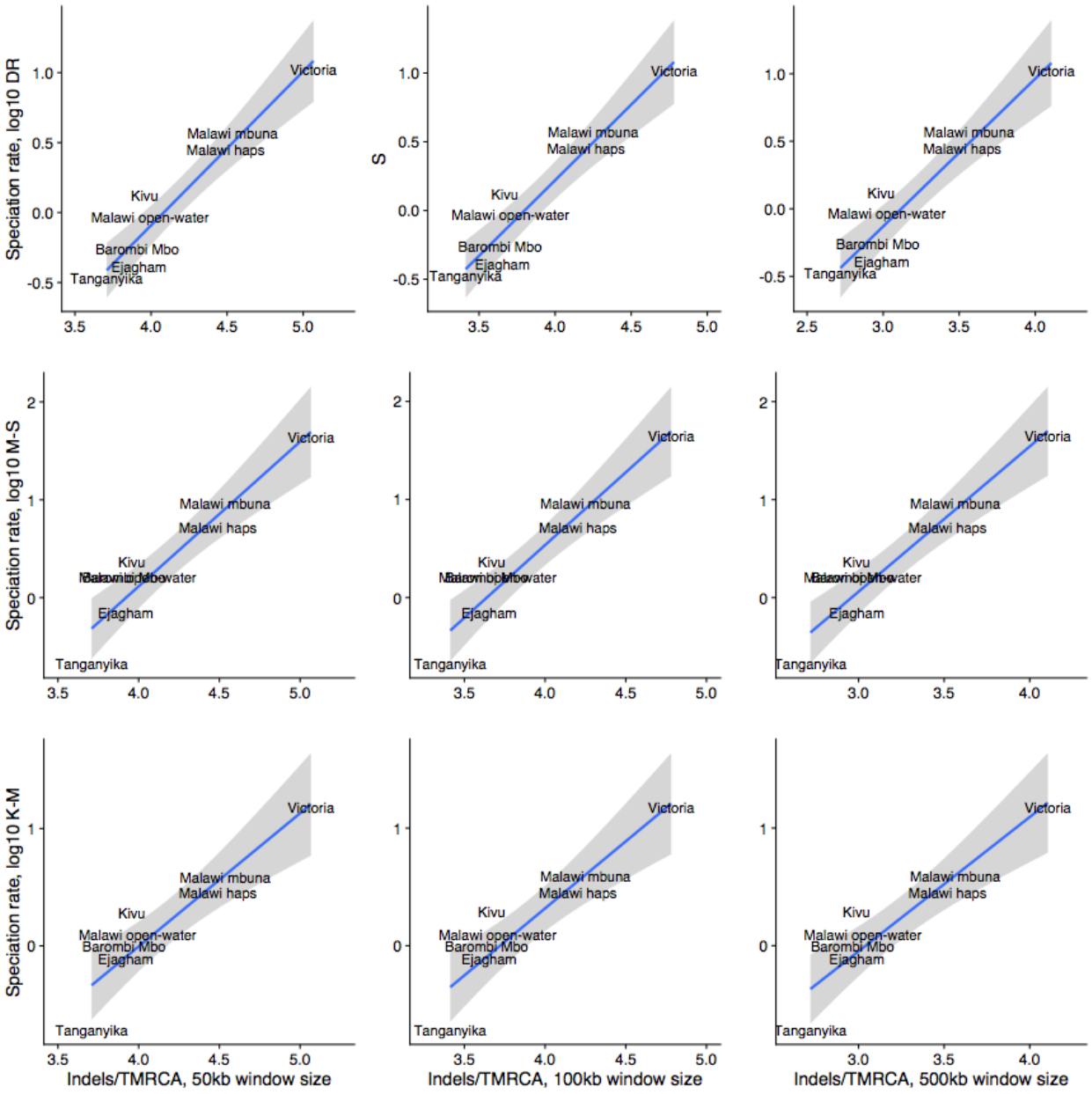
Ejagham	*Coptodon ejagham*	SRR7089205
Ejagham	*Coptodon fusiforme*	SRR7089212
Malawi ‘open water’	*Rhamphochromis esox*	ERR1081383
Malawi ‘open water’	*Diplotaxodon limnothrissa*	ERR715511
Malawi ‘hap’	*Tyrannochromis nigriventer*	ERR1081367
Malawi ‘hap’	*Hemitilapia oxyrhynchus*	ERR715518
Malawi ‘mbuna’	*Tropheops tropheops*	ERR715519
Malawi ‘mbuna’	*Genyochromis mento*	ERR715513
Kivu	*Prognathochromis vittatus*	SRX1406673
Kivu	*Paralabidochromis paucidens*	SRX1406675
Tanganyika	*Neolamprologus gracilis*	ERX1273318
Tanganyika	*Neolamprologus olivaceous*	ERX1273320
Victoria	*Harpagochromis vonlinnei*	This study
Victoria	*Mbipia mbipi*	This study

Supplementary Table S8. Accession numbers for lake radiation cichlid genomes, two species per radiation.

We then utilized the variant calling pipeline described above to call SNPs and indels using assembly graphs of each sample. Each set of two species per radiation was aligned separately to an appropriate reference genome - *O. niloticus* (OreNil2) for the Barombi Mbo and Ejagham radiations, *A. calliptera* (fAstCal1.2) for all haplochromine cichlids, and the *N. brichardi* (NeoBri1.0) genome for Tanganyika. Because the *N. brichardi* genome is not yet assembled to chromosome level, we first performed a genome alignment of the *N. brichardi* assembly to

Oreochromis niloticus using RaGOO (58). Because many of these studies were performed at different coverage levels, mean coverage was reduced to the lowest (5X) prior to calling SNPs and indels.

Speciation rate was calculated as the phylogenetic mean of the log10-transformed DR statistic for the clade defined by the MRCA of each lake pair. We also calculated several traditional clade-level measures of speciation rate using the bd.ms and bd.km functions in R package 'geiger'. We then performed a correlation test of speciation rate against our ratios of indels over TMRCA. (Fig 2a, Supplementary Fig S4).



Supplementary Figure S4. Speciation rates calculated using the log10-transformed DR statistic (top row), Magallon and Sanderson method (middle row), and Kendall-Moran method (bottom row), and number of LD-filtered indels using LD-filtered windows of 500kb, 100kb, and 500kb for two ecomorphologically

divergent species from cichlid lake radiations (n=8). Blue line indicates the regression line from a linear regression; grey shading indicates 95% confidence intervals for the mean.

2.3 Victorian cichlid species sampling:

All sampling in Tanzania was done with research permits from the Tanzania Commission for Science and Technology (COSTECH). All Uganda sampling was done with research permits from Uganda's Department of Agriculture, Animal Industry, and Fisheries obtained through Uganda's National Fisheries Resource Research Institute. We compiled information on the sampling location, date of collection, diet, habitat, and male coloration of each species in our dataset, using published data and our own unpublished information (Appendix 5). Using PCR-free library preparation (59) and Illumina HiSeq 3000 paired-end sequencing, we sequenced the genomes of Lake Victoria cichlids to a mean depth of coverage of 26.4x (14.2-54.6x). To avoid sequencing lane effects and to get an even read representation, barcoded DNA fragments of at least sixteen individuals were sequenced together on multiple Illumina lanes.

2.4 Indel sharing and ecological associations within Lake Victoria

We used custom R scripts to identify indels of 5bp and larger segregating within Lake Victoria at a minor allele frequency of at least 0.01, then identified which of these indels were also segregating across all of our outgroups, as well as within Kivu, Mweru, or the LVRS founding lineages (Fig 2b). We also performed association tests for indels currently assigned to chromosomes 1-22 using the chi-square test against each of our diet and habitat categories at a genome-wide significance level of $p<1x10^{-8}$ using the set of indels segregating within either Kivu, Mweru, and/or the LVRS founding lineages.

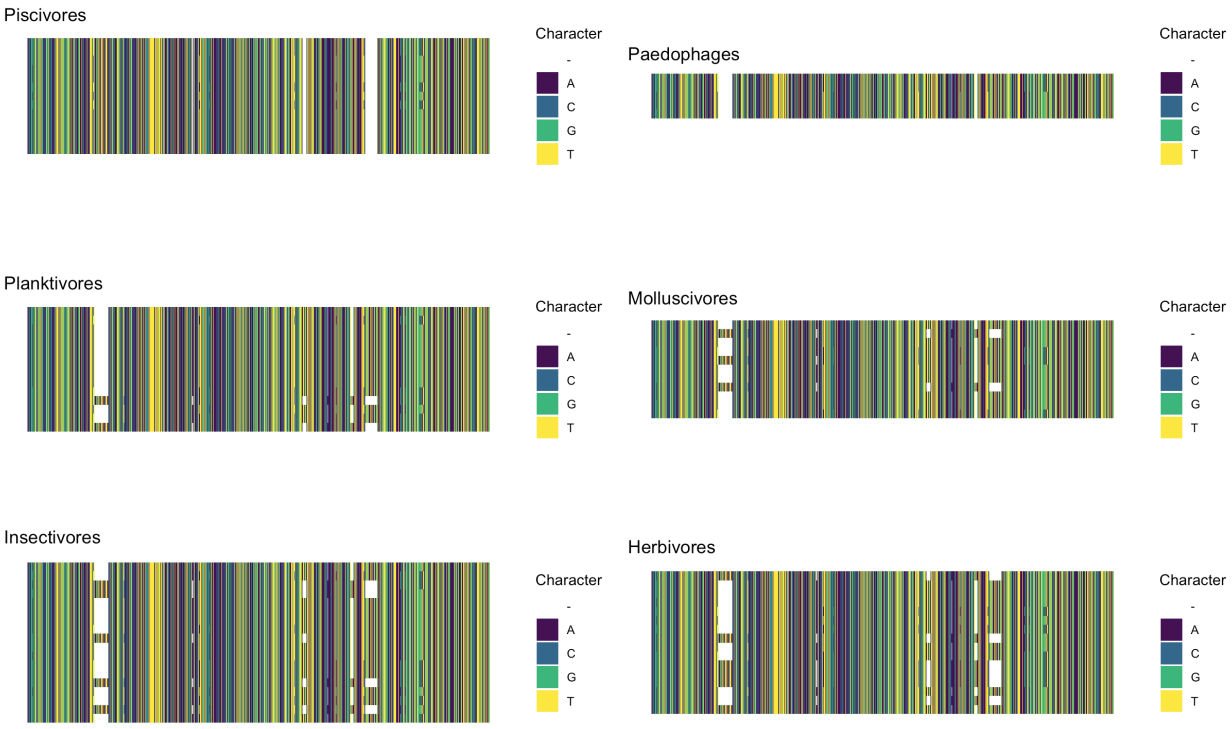
We also examined whether indels segregating within Kivu, Mweru, and/or the LVRS founding lineages were more likely to exhibit a stronger correlation with ecology than indels fixed across all outgroups. We calculated the maximum absolute value of the Pearson correlation coefficient between each indel and our ecological traits, then used STAN (16) to explore the relationship between these two indel types and ecological correlation, using default priors after standardizing our y variable to mean 0 and standard deviation 1, using an exponentially modified gaussian as our family object. We ran four chains with 1000 warm-up and 1000 sampling iterations. We detect a positive effect of ancestrally segregating indels and a negative effect of fixed indels (Supplementary Table S9). We do note, however, that we utilize here only a small number of potential outgroups, and several of these indels that appear as fixed across our outgroups might well involve segregating variants as outgroup sampling increases.

	Estimate	Est.Error	Lower 95% CI	Upper 95% CI	Rhat	ESS
Intercept	-0.01	0.00	-0.01	-0.00	1	3961
Segregating	0.05	0.00	0.04	0.06	1	3632
Fixed	-0.02	0.01	-0.04	-0.01	1	3113

Supplementary Table S9: Summary of the posterior distribution of population-level effects for our STAN regression of indel type against maximal ecological correlation (n=205,838 indels of MAF 0.01 or higher assigned to chromosomes).

2.5 An ancient piscivore-associated allele

One of the most notable regions occurred on chromosome 9, which was fixed in all of our piscivore genomes (Supplementary Figure S3), and also present in the Mweru piscivore *Serranochromis*. Mapping of the region against the PunNye2.0 assembly (60) revealed that these variants occurred within a small uncharacterized regulatory gene of the ubiquitin family adjacent to the stem cell gene THEM6, and that several of the variants were large indels >20bp. To test whether this haplotype had evolved in parallel, we used the software Bali-Phy (61) to generate a gene tree (Fig. 2b) of this 20kb genomic region using a model that includes both substitution rate variation and an indel model. We recover strong support for a clade of morphologically specialized piscivores, including a specialized fish-eating species (*Prognathochromis vittatus*) in the nearby Lake Kivu radiation, which is also part of the older LVRS. This suggests that at least one of the alleles strongly associated with ecological divergence in the young Victoria radiation was not the result of independent parallel evolution at the sequence level, but may in fact be as much as ten million years old. It is also possible that this allele originated in a third extinct lineage, especially given the presence of several paleolake radiations within East Africa over this time period.



Supplementary Figure S3. Indels and SNPs associated with the piscivore allele within an intron on chromosome 9. Alignment of one haplotype for each of our Victorian genomes (20kb), divided into groups

based on dietary guild. Rows correspond to species. Note the large indel near the end of the region fixed for all piscivores

2.6 Evaluating non-treelike evolution between Serranochromis *and Victorian cichlids:*

We ran our McCortex genotyping pipeline on a Lake Victoria predatory species (*Harpagochromis vonlinnei*), a non-predatory LV species (*Paralabidochromis flavus*), *Serranochromis* sp. 'checkerboard' from Mweru, and used version 2 of the *Oreochromis niloticus* genome as the outgroup. We produced a total of 17,643,194 genome-wide biallelic SNPs, and noted that the inclusion of *Oreochromis niloticus* as the outgroup contributes to the large number of SNPs seen here as opposed to our larger set of genomes. We first used the program Dsuite (62) using the Dtrios function with default parameters to calculate D-statistics between the trio defined by *Serranochromis* vs *Harpagochromis* and *Paralabidochromis*, finding no strong evidence of gene flow with a jackknife window of 10,000 ($p<0.08$). However, examination of the first 2 Mb of chromosome 9, which contains the likely introgressed region, reveals a weak signature of gene flow ($p<0.03$), and examining the first 2 Mb in windows of 10 SNPs with a step size of 5 SNPs reveals a very strong signal for admixture proportion (f_d). We also tested for the presence of recombination using the pairwise homoplasy test (Bruen et al. 2006), in Splitstree4 (63), using possible window sizes of 5kb ($p<0.001$), 10kb ($p>0.05$), 20kb ($p>0.05$), and 50kb ($p>0.05$). We found statistically significant evidence for recombination for all window sizes except 5kb.

To distinguish whether non-treelike evolution between *Serranochromis* and Lake Victoria Superflock piscivores was due to introgression or incomplete lineage sorting, we used QuIBL (64). We divided up a 1Mb region on chromosome 9 flanking the piscivore-associated region into 5kb windows, and then converted the SNPs in these windows in combination with the *Pundamilia nyererei* reference genome into an alignment for each window, excluding windows with fewer than 25 SNPs, for a total of 1876 trees. We then ran RaxML 8.12.2 (41) with a GTR+G model on each window, rooting each tree on *Oreochromis niloticus*, then ran QuIBL with a likelihood threshold of 0.01, 50 steps, and a gradient scalar of 0.5. For the triplet made from our Lake Victoria pair and Serranochromis, we find strong support for a two-distribution model (BIC -20233) vs a one-distribution model (BIC -18396), supporting introgression from *Serranochromis* into the Lake Victoria Superflock over incomplete lineage sorting. This suggests that the topologies in this region that link *Serranochromis* with Victorian piscivores are not because topologies are unstable enough at this evolutionary scale that simple incomplete lineage sorting could be responsible.

2.7 Identity by descent segments:

Genomic segments that are shared between individuals because of inheritance from the same parent (identical by descent, IBD) become broken up by recombination through successive generations. Unless such IBD segments are protected against mutation and recombination by selection or recombination suppression, the number and length of segments shared between individuals from two different species are informative about the time since the most recent common ancestor of these species, or since their most recent hybridization (25).

We estimated IBD segments with IBDseq (26), using a LOD score of 5, no trimming, and an error rate of 0.01. We did not utilize LD filtering, preferring to identify regions with an excess of IBD segment sharing using the reshuffling procedure described in the following section. We

converted physical position to map position to the nearest 0.01 cM, using the recombination map from the most recent *Pundamilia nyererei* assembly (60). We retained only segments large enough (0.1cM+) to confidently confirm a region as being in identity by descent (26). Because estimates of IBD sharing at chromosome ends can be overestimated due to reduced recombination rate information, we trimmed IBD segments falling within the bottom 5% and top 95% of map position for each chromosome before running IBDseq. We then split all segments overlapping gaps in the assembly larger than 0.1 cM, then removed any segments below 0.1 cM. Eight genomes (*Paralabidochromis* 'short head chilotes', *Harpagochromis* 'orange rock hunter', *Yssichromis* 'plumbus', *Pyxichromis* 'stripe', *Labrochromis* 'grey', *Haplochromis* 'brown snout', *Lipochromis* 'velvet black cryptodon', *Haplochromis* cf. 'supramacrops') in our sample exhibted extremely high levels of IBD sharing relative to a described or better-known undescribed species; these were removed from the IBD database prior to the reshuffling described below, as they likely represented conspecific individuals.

2.8 IBD sharing simulations:

Stabilizing selection within a population can slow or prevent recombination breaking IBD segments, resulting in 'oversharing' of IBD segments, and these segments may alter inferred genetic relationships. IBD patterns can also be affected by genomic features such as suppressed recombination, the presence of inversions, and by centromeres. To separate IBD segments likely not associated with selection or genome structure, we compared our distributions of IBD sharing to patterns observed from datasets where IBD segments were randomly assigned to new non-overlapping genomic positions using 'bedtools shuffle'. The total IBD sharing between individuals remained the same, but locations of IBD segments were permuted, excluding the first and last 5% of each genome, as well as gaps of 0.1 cM and larger. IBD segments were shuffled based on cM distance rather than physical distance. For each simulation, we then created networks of IBD sharing intersecting each 0.1 cM interval along the genome, then calculated the largest clique size present at that location. In graph theory, cliques represent nodes where each member of the clique is connected to each other member.

Clique size never exceeded a size of four in our reshuffled segments, suggesting that in our empirical data, locations in the genome where five or more genomes all shared IBD segments with each other could involve processes or structure not associated with shared ancestry. We then examined each 0.1 cM interval, recording which IBD segments were involved in cliques of size 3 and larger. We then identified IBD segments where <5% of the length was assigned to cliques larger than observed in our shuffling procedure, which we hereafter refer to as 'non-overshared'.

We modeled the effect of our diet, habitat, male nuptial coloration, and sympatry factors (Appendix 5) on sharing of IBD segments among species in the Lake Victoria radiation in a Bayesian modelling framework using STAN (16). Much as phylogenetic analyses must take evolutionary nonindependence into account, we take into account the non-independence in our network via a multi-member random effect. Briefly, this allows each of the two variables in our dataset associated with the pair of genomes sharing IBD to belong to the same random effect variable, thus ensuring that all comparisons involving two genomes are not independent of all comparisons involving either of the two genomes.

2.9 Temporal sequence of divergence:

We examined three different IBD segment length classes corresponding to three different relative time windows in the history of the radiation, with shorter segments representing more ancient and longer segments more recent ancestry (Fig. 3b). We converted these into three networks, with each link in the network representing the presence of at least one IBD segment within the given length class shared between the genomes of two species. We tested whether diet, macrohabitat or male nuptial coloration best explained the genetic structure of the radiation at different time points during its unfolding. For this, we fitted a Bayesian regression model using STAN (16) to the relationship between the number of IBD segments shared between any two species and their difference in diet, habitat and nuptial color, while simultaneously considering the non-independent nature of pairwise IBD sharing between species.

We divided our non-overshared IBD segments into three segment size categories representing different time stages in the radiation: 0.1-0.19, 0.2-39, and 0.4-0.79 cM. We then scored whether each pairwise combination of genomes possessed one or more IBD segments in that size category, then used this binary variable as our response variable for each model. We also ran a series of models where we scored whether each pairwise combination of genomes possessed five or more IBD segments of the size category (Supplementary Table S10). We modeled the presence or absence of IBD segments with a Bernoulli family object. We ran four chains with 10,000 iterations using a set of weakly informative priors: mean 0 and standard deviation 1 for intercept, slope, and a gamma(2,1) prior for the multi-member random effects term. Neither diet, habitat, sympatry, or male nuptial coloration exhibited 95% credible intervals outside zero for our smallest segment size, representing the earliest stages of divergence in the Lake Victoria radiation. However, the subsequent three models with larger segment sizes exhibited a positive 95% credible interval for both diet and habitat. Male nuptial coloration exhibited a negative effect, but only exhibited a 66% credible interval outside zero.

	Estimate	Est. Error	Lower 95% CI	Upper 95% CI	Rhat	ESS
0.1-0.19 cM						
Intercept	-6.67	0.52	-7.75	-5.74	1	21741
Diet	-0.07	0.71	-1.54	1.126	1	23880
Habitat	-0.03	0.65	-1.34	1.2	1	22063
NuptialColor	-0.04	0.65	-1.34	1.21	1	24217
Multi-member random effect	0.63	0.35	0.1	1.45	1	16702
0.2-0.39 cM						
Intercept	2.39	0.17	2.07	2.74	1	9784
Diet	0.72	0.17	0.39	1.06	1	40415
Habitat	0.38	0.14	0.11	0.67	1	28804

NuptialColor	-0.13	0.13	-0.38	0.12	1	33432
Multi-member random effect	1.37	0.16	1.09	1.71	1	7049
0.4-0.79 cM						
Intercept	-1.61	0.1	-1.80	-1.42	1	16997
Diet	0.59	0.09	0.41	0.77	1	41722
Habitat	0.42	0.09	0.25	0.59	1	32243
NuptialColor	-0.1	0.09	-0.27	0.07	1	36903
Multi-member random effect	0.68	0.09	0.53	0.86	1	9440

Supplementary Table S10: Summary of the posterior distribution of population-level effects for our STAN regression of presence or absence of IBD segments of a given size class against diet, habitat, and male nuptial coloration (n=4186 pairwise comparisons).

2.10 D-statistics across the Lake Victoria radiation

As a complement to our IBD models, we calculated D-statistics for all combinations of our 100 Victorian cichlid genomes, using a taxon from Lake Kivu (*P. paucidens*) as the outgroup. To reduce the influence of regions with very high SNP density, we performed LD filtering using R package 'SNPRelate', using a window size of 50kb and a correlation threshold of 0.5, reducing our dataset to 1,756,447 biallelic SNPs. We analyzed a total of 161,700 quartets, using 20 jackknife blocks for each quartet to calculate p-values for the D-statistic. We then utilized the "BBAA" output from Dsuite, which arranges quartets into their most treelike configuration (Malinsky 2019). Quartets with Bonferroni-corrected significant p-values (n=22,764, p<3.09e-7) were marked as belonging to a "high introgression" category. We then passed the rooted quartets not in the high introgression category (n=138,936) to ASTRAL (65) to generate a multispecies coalescent tree (Extended Data Fig 4a). ASTRAL recovered a treelike history of the radiation, but it is critical to note that within our 'high introgression' category, there is evidence of gene flow involving all of our Victoiran cichlid genomes, indicating that non-treelike evolution is ubiquitous within the Victorian radiation (Extended Data Fig 4b). We then tested whether diet, macrohabitat or male nuptial coloration best explained assignment to the 'high introgression' category by fitting a Bayesian regression model using STAN that examines whether a high D-statistic is associated with the two ingroup taxa sharing the same diet, habitat, or male nuptial coloration, as well as whether the two potentially introgressing taxa share those same traits (Extended Data Fig 4). We also incorporated a multi-member random effect to account for the non-independence of comparisons involving the same species, as we did for our IBD analysis.

We find that a more treelike pattern is more likely to occur when the two ingroup taxa share the same diet or habitat, analogous to phylogenetic signal. This suggests that both diet and habitat are associated with the macro-scale genomic structure of the radiation. However, a treelike pattern is more likely when the two ingroup taxa have different male coloration. We find that a quartet is more likely to belong to the 'high introgression' category if the two taxa potentially exchanging genes have the same diet or habitat, but male nuptial coloration had no effect.

Data and Code Availability

Genomic data is available at NCBI BioProject ID: PRJNA626405. Data files and scripts are available via Dryad DOI: https://doi.org/10.5061/dryad.fn2z34tr0

Supplemental References:

31. Smith, S. A., Beaulieu, J. M., & Donoghue, M. J. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evol. Biol.*, *9*(1), 37. (2009)

32. Borstein, S.R & O'Meara, B.C., AnnotationBustR: an R package to extract subsequences from GenBank annotations. *PeerJ 6*, p.e5179. (2018)

33. Costa F.O. & Carvalho G.R. The Barcode of Life Initiative: synopsis and prospective societal impacts of DNA barcoding of fish. *Gen. Soc. Pol.* Dec;3(2):29. (2007)

34. Friedman, M., Keck, B.P., Dornburg, A., Eytan, R.I., Martin, C.H., Hulsey, C.D., Wainwright, P.C. & Near, T.J.. Molecular and fossil evidence place the origin of cichlid fishes long after Gondwanan rifting. *Proc. Roy. Soc. B* 280(1770), 20131733. (2013)

35. Martin, C.H., Cutler, J.S., Friel, J.P., Dening Touokong, C., Coop, G., & Wainwright, P. C. Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution*, *69*(6), 1406-1422. (2015)

36. McGee, M. D., Faircloth, B. C., Borstein, S. R., Zheng, J., Darrin Hulsey, C., Wainwright, P. C., & Alfaro, M. E. Replicated divergence in cichlid radiations mirrors a major vertebrate innovation. *Proc. Roy. Soc. B*, *283*(1822), 20151413. (2016).

37. Rican, O., Pialek, L., Dragova, K., & Novak, J. Diversity and evolution of the Middle American cichlid fishes (Teleostei: Cichlidae) with revised classification. *Vert. Zool. 66*(1), 3-102. (2016).

38. Burress, E. D., Alda, F., Duarte, A., Loureiro, M., Armbruster, J. W., & Chakrabarty, P. Phylogenomics of pike cichlids (Cichlidae: Crenicichla): the rapid ecological speciation of an incipient species flock. *J. Evol. Biol. 31*(1), 14-30. (2018)

39. Irisarri, I., Singh, P., Koblmüller, S., Torres-Dowdall, J., Henning, F., Franchini, P., ... & Sturmbauer, C. Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. *Nat. Comm.*, *9*(1), 1-12. (2018)

40. Paradis E, Claude J, & Strimmer K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2), 289-90. (2004)

41. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30(9), 1312-3.(2014)

42. Smith, S.A., & O'Meara, B.C. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics 28*(20), 2689-2690. (2012).

43. Matschiner, M., Musilová, Z., Barth, J.M., Starostová, Z., Salzburger, W., Steel, M., & Bouckaert, R. Bayesian phylogenetic estimation of clade ages supports trans-Atlantic dispersal of cichlid fishes. *Sys. Biol.*. 66(1):3-22. (2017)

44. GBIF: The Global Biodiversity Information Facility (year) What is GBIF?. Available from https://www.gbif.org/what-is-gbif (2018)

45. Boettiger C, Lang DT, Wainwright PC. rfishbase: exploring, manipulating and visualizing FishBase data from R. *J. Fish Biol.*. 81(6):2030-9. (2012)

46. Abell R., Thieme M.L., Revenga C., Bryer M., Kottelat M., Bogutskaya N., Coad B., Mandrak N., Balderas S.C., Bussing W., & Stiassny M.L.. Freshwater ecoregions of the world: a new map of biogeographic units for freshwater biodiversity conservation. *BioScience* 58(5), 403-14. (2008)

47. Andreadis K.M., Schumann G.J., & Pavelsky T. A simple global river bankfull width and depth database. *Wat. Res. Res.* 49(10), 7164-8. (2013)

48. Jarvis, A., Reuter, H.I., Nelson, A., & Guevara, E. Hole-filled SRTM for the globe Version 4. (2008)

49. Fick S.E. & Hijmans R.J.. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Clim.*. 37(12), 4302-15. (2017)
50. Bürkner PC. brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Soft* 80(1), 1-28. (2017)
51. Brooks S.P., & Gelman A.. General methods for monitoring convergence of iterative simulations. *J. Comp. Graph. Stat.* 7(4), 434-55. (1998)
52. Muschick, M., Russell, J.M., Jemmi, E., Walker, J., Stewart, K.M., Murray, A.M., Dubois, N., Stager, J.C., Johnson, T.C., & Seehausen, O. Arrival order and release from competition does not explain why haplochromine cichlids radiated in Lake Victoria. *Proc. Roy. Soc. B* 285(1878), p.20180462. (2018)
53. Dunz, A.R. & Schliewen, U.K., Description of a Tilapia (Coptodon) species flock of Lake Ejagham (Cameroon), including a redescription of Tilapia deckerti (Thys van den Audenaerde, 1967). *Spixiana 33*(2), 251-280. (2010)
54. Trewavas, E., Green, J., & Corbet, S. A. Ecological studies on crater lakes in West Cameroon Fishes of Barombi Mbo. *J. Zool.* 167(1), 41–95. (2009)
55. Turner, G.F. Offshore Cichlids of Lake Malawi. Cichlid Press, El Paso, Texas. (1996)
56. Genner, M.J., & Turner, G.F.. The mbuna cichlids of Lake Malawi: a model for rapid speciation and adaptive radiation. *Fish and Fisheries*, 6(1), 1-34. (2005)
57. McGee, M.D., Neches, R.Y. & Seehausen, O. Evaluating genomic divergence and parallelism in replicate ecomorphs from young and old cichlid adaptive radiations. *Mol. Ecol. 25*(1), 260-268. (2016)
58. Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., ... & Schatz, M. C. (2019). RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Gen. Biol. 20*(1), 1-17.
59. Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., & Turner, D.J.. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+ C)-biased genomes. *Nat. Meth.* 6(4), 291. (2009)
60. Feulner, P.G., Schwarzer, J., Haesler, M.P., Meier, J.I., & Seehausen, O. A dense linkage map of Lake Victoria cichlids improved the Pundamilia genome assembly and revealed a major QTL for sex-determination. *G3*, 200207. (2018)
61. Suchard, M.A. & Redelings, B.D. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* 22(16), 2047-8. (2006)
62. Malinsky, M. Dsuite:fast D-statistics and related admixture evidence from VCF files. *BioRxiv*, 634477. (2019)
63. Huson, D.H. & Bryant, D. Estimating phylogenetic trees and networks using SplitsTree 4. *Manuscript in preparation, software available from www.splitstree.org*.
64. Edelman, N.B., Frandsen, P.B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., ... & Challis, R. Genomic architecture and introgression shape a butterfly radiation. *Science* 366(6465), 594-599. (2019)
65. Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinf. 19*(6), 153. (2018).
66. Trewavas, E. *Tilapiine fishes of the genera Sarotherodon, Oreochromis and Danakilia*. British Museum (Natural History). (1983)
67. Lamboj, A. The cichlid fishes of western Africa. (2004)
68. Konings, A. (Ed.). *Enjoying cichlids*. Cichlid Press. (1993)