Title: Whole genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow

Authors: Milan Malinsky[1,2]†*, Hannes Svardal[1]†, Alexandra M. Tyers[4], Eric A. Miska[1,3], Martin J. Genner[5], George F. Turner[4], and Richard Durbin[1]*

Affiliations:

[1]Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK.

[2]Zoological Institute, University of Basel, 4051 Basel, Switzerland.

[3]School of Biological Sciences, Bangor University, Bangor, Gwynedd LL57 2UW, UK.

[4]Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge, CB2 1QN, UK.

[5]School of Biological Sciences, Life Sciences Building, 24 Tyndall Avenue, University of Bristol, Bristol BS8 1TQ, UK.

*Correspondence to: milan.malinsky@unibas.ch, rd@sanger.ac.uk.

†These authors contributed equally to this work.

Abstract: The hundreds of cichlid fish species in Lake Malawi constitute the most extensive recent vertebrate adaptive radiation. Here we characterize its genomic diversity by sequencing 134 individuals covering 73 species across all major lineages. Average sequence divergence between species pairs is only 0.1-0.25%. These divergence values overlap diversity within species, with 82% of heterozygosity shared between species. Phylogenetic analyses suggest that diversification initially proceeded by serial branching from a generalist *Astatotilapia*-like ancestor. However, no single species tree adequately represents all species relationships, with evidence for substantial gene flow at multiple times. Common signatures of selection on visual and oxygen transport genes shared by distantly related deep water species point to both adaptive introgression and independent selection. These findings enhance our understanding of genomic processes underlying rapid species diversification, and provide a platform for future genetic analysis of the Malawi radiation.

One Sentence Summary: The genomes of 73 cichlid fish species from Lake Malawi uncover evolutionary processes underlying a large adaptive evolutionary radiation.

Main Text: The formation of every lake or island represents a fresh opportunity for colonization, proliferation and diversification of living forms. In some cases, the ecological opportunities presented by underutilized habitats facilitate adaptive radiation - rapid and extensive diversification of the descendants of the colonizing lineages (*1*-*3*). Adaptive radiations are thus exquisite examples of the power of natural selection, as recognized by Darwin in the case of Galapagos finches (*4*, *5*), and seen for example in Anoles lizards of the Caribbean (*6*) and in East African cichlid fishes (*7*, *8*).

Cichlids are the most species rich and diverse family of vertebrates, and nowhere are their radiations more spectacular than in the Great Lakes of East Africa: Malawi, Tanganyika, and Victoria (*2*), each of which contains several hundred endemic species, with the largest number in Lake Malawi. Molecular genetic studies have made major contributions to reconstructing the evolutionary histories of these adaptive radiations, especially in terms of the relationships between the lakes (*9*), between some major lineages in Lake Tanganyika (*10*), and in describing the role of hybridization in the origins of the Lake Victoria radiation (*11*). However, the task of reconstructing within-lake relationships and of identifying sister species in lakes Malawi and Victoria remains challenging due both to retention of large amounts of ancestral genetic polymorphism (i.e. incomplete lineage sorting) and to evidence suggesting gene flow between taxa (*12*, *13*).

Initial genome assemblies of cichlids from East Africa suggest that an increased rate of gene duplications together with accelerated evolution of some regulatory elements and protein coding genes may have contributed to the radiations (*14*). However, understanding of the genomic mechanisms contributing to adaptive radiations is still in its infancy (*3*). Here we provide an overview of and insights into the genomic signatures of the haplochromine cichlid radiation of Lake Malawi.

Previous work on the phylogeny of Lake Malawi haplochromine cichlid radiation, mainly based on mitochondrial DNA (mtDNA), has divided the species into six groups with differing ecology and morphology (*15*): 1) the rock-dwelling ‘mbuna’; 2) *Rhamphochromis* - typically midwater zooplanktivores and piscivores; 3) *Diplotaxodon* - typically deep-water pelagic zooplanktivores and piscivores; 4) deep-water and twilight feeding benthic species; 5) ‘utaka’ feeding on zooplankton in the water column but breeding on or near the benthic substrate; 6) a diverse group of benthic species, mainly found in shallow non-rocky habitats. In addition, *Astatotilapia calliptera* is a closely related generalist that inhabits shallow weedy margins of Lake Malawi, and other lakes and rivers in the catchment and more widely. To characterize the genetic diversity, species relationships, and signatures of selection across the whole radiation, we obtained paired-end Illumina whole-genome sequence data from 134 individuals of 73 species distributed broadly across the seven groups (Fig. 1A) (*16*). This includes 102 individuals at ~15× coverage and 32 additional individuals at ~6× (Table S1).

Low genetic diversity and species divergence

Sequence data were aligned to the *Metriaclima zebra* reference assembly version 1.1 (*14*), followed by variant calling restricted to the 'accessible genome' (the portion of the genome where variants can be determined confidently with short read alignment), which comprised 653Mb or 91.5% of the assembly excluding gaps (*16*). Average divergence from the reference per sample was 0.19% to 0.27% (Fig. S1). Across all samples, after filtering and variant refinement, we obtained 30.6 million variants of which 27.1 million were single nucleotide polymorphisms (SNPs) and the rest were short insertions and deletions (indels) (*16*).

To estimate nucleotide diversity (π) within the sampled populations we measured the frequency of heterozygous sites in each individual. The estimates are distributed within a relatively narrow range between 0.7 and 1.8×10^{-3} per bp (Fig. 1B). The mean π estimate of 1.2×10^{-3} per bp is at the low end of values found in other animals (*17*), and there appears to be little relationship between π and the rate of speciation: individuals in the species-rich mbuna and shallow benthic groups show levels of π comparable to the relatively species-poor utaka, *Diplotaxodon*, and *Rhamphochromis* (Fig. S1).

Despite their extensive phenotypic differentiation, all species within the Lake Malawi haplochromine cichlid radiation are genetically closely related (*12*, *18*). However, genome-wide genetic divergence has never been quantified. To examine the extent of genetic differentiation between species across the radiation we calculated the average pairwise differences between all pairs of sequences from different species (d_{XY}). A comparison of d_{XY} against heterozygosity reveals that the two distributions partially overlap (Fig. 1B). Thus, the sequence divergence within a single diploid individual is sometimes higher than the divergence between sequences found in two distinct species. The average d_{XY} is 2.0×10^{-3} with a range between 1.0 and 2.4×10^{-3} per bp. The maximum d_{XY} is approximately one fifth of the divergence between human and chimpanzee (*19*). The low ratio of divergence to diversity means that most genetic variation is shared between species. On average both alleles are observed in other species for 82% of heterozygous sites within individuals.

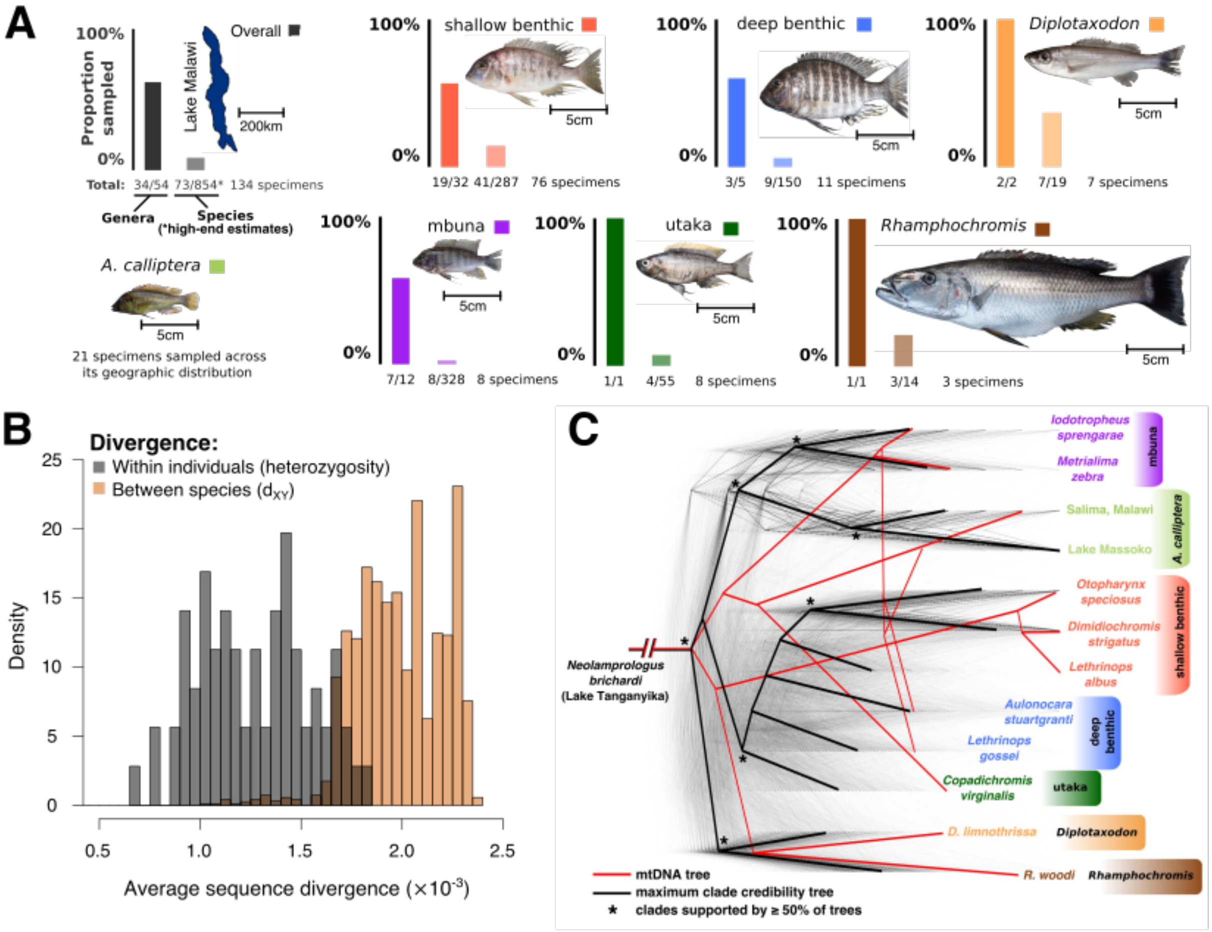

Fig. 1: The Lake Malawi haplochromine cichlid radiation. (A) The sampling coverage of this study: overall and for each of the seven main eco-morphological groups within the radiation. A representative specimen is shown for each group (Diplotaxodon: *D. limnothrissa*; shallow benthic: *Lethrinops albus*; deep benthic: *Lethrinops gossei*; mbuna: *Metriaclima zebra*; utaka: *Copadichromis virginalis*; Rhamphochromis: *R. woodi*). Numbers of species and genera are based on ref. (*20*). **(B)** The distributions of genomic sequence diversity within individuals (heterozygosity; $\hat{\pi}$) and of divergence between species (d_{XY}). **(C)** A set of 2543 Maximum Likelihood (ML) phylogenetic trees for non-overlapping regions along the genome. Branch lengths were scaled for visualization so that the total height of each tree is the same. The local trees were built with 71 species sequenced to 15x coverage and then subsampled to 12 individuals representing the eco-morphological groups. The maximum clade credibility

tree was build from the phylogenies with 12 individuals, and represents a 'consensus' of these. A maximum likelihood mitochondrial phylogeny is shown for comparison.

Low per-generation mutation rate

It has been suggested that the species richness and morphological diversity of teleosts in general and of cichlids in particular might be explained by elevated mutation rates compared to other vertebrates (*21*, *22*). To obtain a direct estimate of the per-generation mutation rate, we reared offspring of three species from three different Lake Malawi groups (*Astatotilapia calliptera*, *Aulonocara stuartgranti*, and *Lethrinops lethrinus*). We sequenced both parents and one offspring of each to high coverage (40x), applied stringent quality filtering, and counted variants present in each offspring but absent in both its parents (Fig. S2) (*16*). There was no evidence for significant difference in mutation rates between species. The overall mutation rate (μ) was estimated at 3.5×10^{-9} (95%CI: 1.6×10^{-9} to 4.6×10^{-9}) per bp per generation, approximately three to four times lower than the rate obtained in similar studies using human trios (*23*). By combining this mutation rate with nucleotide diversity (π) values, we estimate the long term effective population sizes (N_e) to be in the range of approximately 50,000 to 130,000 breeding individuals (with $N_e = \pi/4\mu$). This result suggests that alleles at a locus will only rarely coalesce within the time between successive speciation events, because both the mean and standard deviation in the time to the common ancestor are expected to be in the order of hundreds of thousands of years ($2N_e$ generations).

Between-species relationships

To obtain a first estimate of between-species relationships we divided the genome into 2543 non-overlapping windows, each comprising 8000 SNPs (average size: 274kb), and constructed a Maximum Likelihood (ML) phylogeny separately for each window, obtaining trees with 2542 different topologies. We also calculated the maximum clade credibility (MCC) summary tree (*24*) and an ML phylogeny based on the full mtDNA genome (Figs. 1C, S4) (*16*). Despite extensive variation among the individual trees, it is apparent that there is some general consensus. Individuals from the previously identified eco-morphological groups tend to cluster together, with *Rhamphochromis*, *Diplotaxodon*, mbuna and *A. calliptera* forming well supported (in >80% of trees) apparently reciprocally monophyletic groups (Fig. S4B), while individuals from the utaka, and deep and shallow benthic were clustered within their respective groups, but with lower support. In the subset of 12 individuals shown in Fig. 1C, the pelagic clades *Diplotaxodon* and *Rhamphochromis* tend to cluster together and form an outgroup to the rest of the radiation. Perhaps surprisingly, the majority of the trees support the mbuna and the generalist *A. calliptera* being sister taxa phylogenetically nested within the radiation (Figs. 1C, S4B). The overall sub-structuring is also apparent from patterns of linkage disequilibrium (LD). Mean LD decays within a few hundred base-pairs across the set of all species, in a few kilobases (kb) for subsets of species from within eco-morphological subgroups (mbuna, *Diplotaxodon* etc.), and extends beyond 10kb within species (Fig. S3) (*16*).

It is also clear that the mtDNA phylogeny is substantially different both from the MCC summary and from the majority of the individual trees (Figs. 1C, S5). This underlines the importance of evaluating the Lake Malawi radiation from a genome-wide perspective, rather than relying on mtDNA.

Violations of the species tree concept

There are several pronounced differences in topology between the MCC phylogeny for a subset of 12 individuals (11 species, Fig. 1C) and the full 71 species MCC tree (Fig. S4), revealing that phylogenetic inference is sensitive to the number and kind of samples included. To further investigate genetic relationships between Lake Malawi cichlids, we used a variety of additional approaches, including Neighbor-Joining (NJ), UPGMA and Bayesian multispecies coalescent (*16*, *25*). Overall, the phylogenies reflect many features of previous taxonomic assignment, but some genera are clearly polyphyletic, including for example *Placidochromis*, *Lethrinops*, and *Mylochromis*. Furthermore, we also found considerable variation in topologies depending on which inference method was used and also within the Bayesian MCMC samples (Figs. 2A, S6). The fact that we are using over 25 million variable sites suggests these differences are not due to sampling noise, but reflect conflicting biological signals in the data. For example gene flow after the initial separation of species can distort the overall phylogeny and lead to intermediate placement of admixed taxa in the tree topology. Hence we investigated this possibility using a number of different approaches.

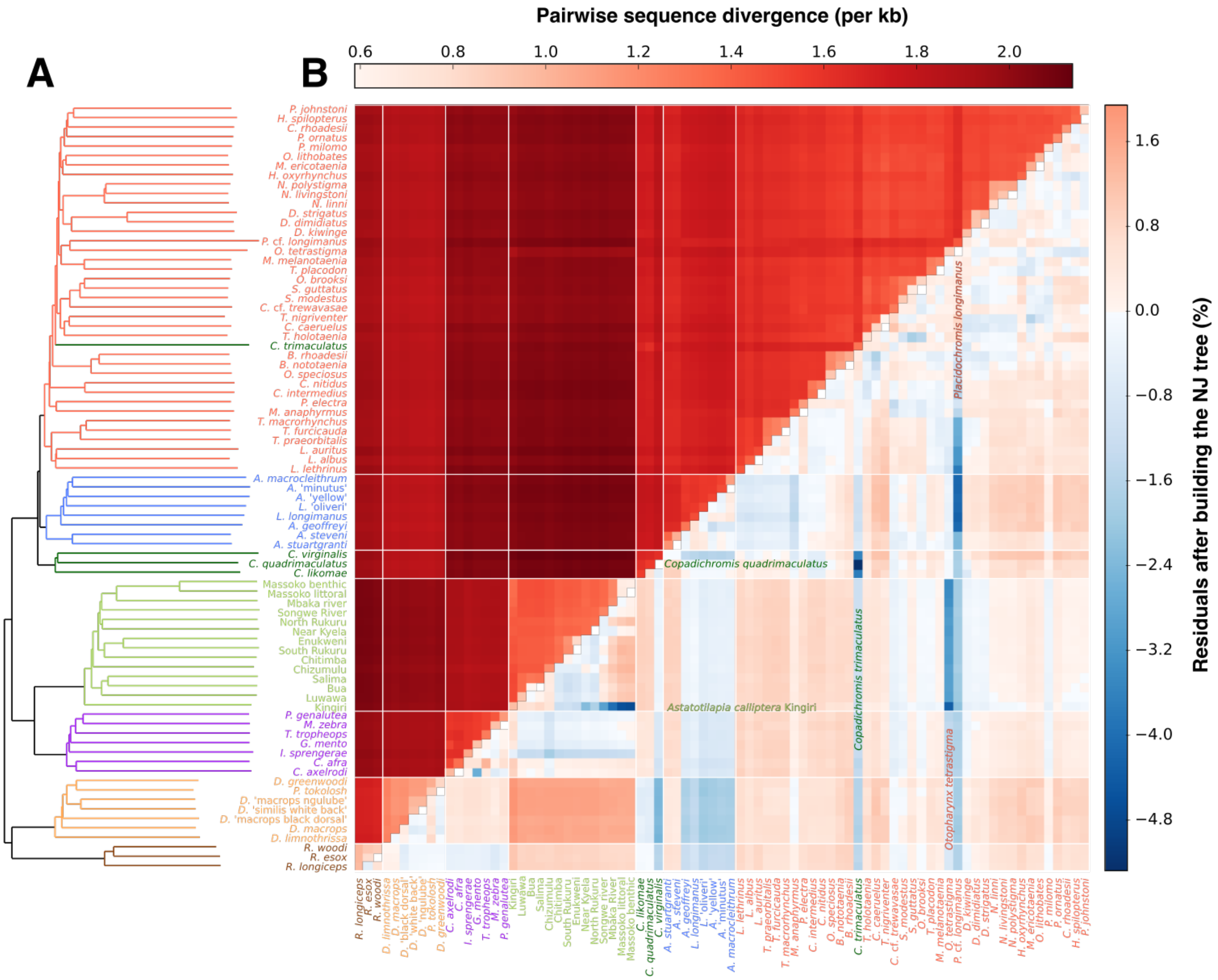

Fig. 2: Species relatedness. (A) Neighbor-joining tree of pairwise differences. Long terminal long terminal branches reflect the high ratio of within-species to between-species variation. **(B)** Pairwise genetic differences (above diagonal) and residuals of pairwise difference and tree distance (below diagonal). The residuals for each pair of individuals are calculated as: (sequence distance− tree distance)/sequence distance.

As a first step, we contrasted the pairwise genetic differences used to produce the raw NJ tree (Fig. 2B; above the diagonal) against the distances between samples along the tree branches (Fig. 2A), calculating the residuals (Fig. 2B; below the diagonal). If the tree were able to perfectly capture all the genetic relationships in our sample, we would expect the residuals to be zero. However, we found a large number of differences, with some standout cases of violations of the tree-like model. Among the strongest signals, we see that: 1) *Copadichromis trimaculatus*, which in the NJ tree clusters within the shallow benthic clade, is genetically much closer to other utaka, and in particular to *C. quadrimaculatus*, than its placement in the tree would suggest; 2) *Placidochromis* cf. *longimanus* is genetically closer to the deep benthic clade and to a subset of the shallow benthic (mainly *Lethrinops* species) than the tree suggests; and 3) our sample of *Otopharynx tetrastigma* is much closer to *Astatotilapia calliptera* from Lake Kingiri (and to a lesser degree to other *A. calliptera*) than would be expected from the tree. The *O. tetrastigma* specimen comes from Lake Ilamba, a satellite crater lake of Lake Malawi that also harbors a population of *A. calliptera* and is geographically close (3.2 km) to Lake Kingiri.

Sharing of long haplotypes between otherwise distantly related species is an indication of recent admixture or introgression. To investigate this type of gene-flow signature we used the chromopainter software package (*26*) and calculated a 'co-ancestry matrix' of all species (*16*) - a summary of nearest neighbor (therefore recent) haplotype relationships in the dataset. We found that the Lake Ilamba *O. tetrastigma* and Lake Kingiri *A. calliptera* stand out in this analysis by showing a strong signature of recent gene flow between distantly related species (Fig. S7). The other tree-violation signatures described above are also visible on the haplotype sharing level but are less pronounced, consistent with being older events perhaps involving the common ancestors of multiple present-day species. However, the chromopainter analysis reveals numerous other examples of excess co-ancestry between species that do not cluster immediately together (e.g. the utaka *C. virginalis* with *Diplotaxodon*; more highlighted in Fig. S7). Furthermore, the clustering based on recent co-ancestry is different from any tree generated using phylogenetic methods; in particular a number of shallow benthics including *P.* cf. *longimanus* cluster next to the deep benthics. Related to this, principal component analysis (PCA), while generally separating major clades, shows a 'smear' of utaka and benthic samples towards *Diplotaxodon* (Fig. 3A), a pattern that is typical for admixed populations (e.g. ref. (*27*)).

To test more specifically for non-tree-like relatedness patterns consistent with gene flow between the 'tree-violating' taxa identified above, we computed the f_4 admixture ratio (*28-30*) (*f* statistic in the following), which is closely related to Patterson's D (ABBA-BABA test) (*28*), and when elevated due to introgression is expected to be linear in the amount of introgressed material. For each of the three cases discussed in the analysis of Figure 2 we found strong signals of non-tree-like relatedness (Fig. 3B). In particular, we interpret the very high *f* statistics involving *C. trimaculatus* as suggesting that the gene-pool of this species is a product of hybridization between an utaka lineage and a shallow benthic lineage. It is notable that the position of *C.*

trimaculatus within the shallow benthic group is an unusual feature of the NJ tree in Fig. 2; it clusters with the other utaka in all other phylogenies.

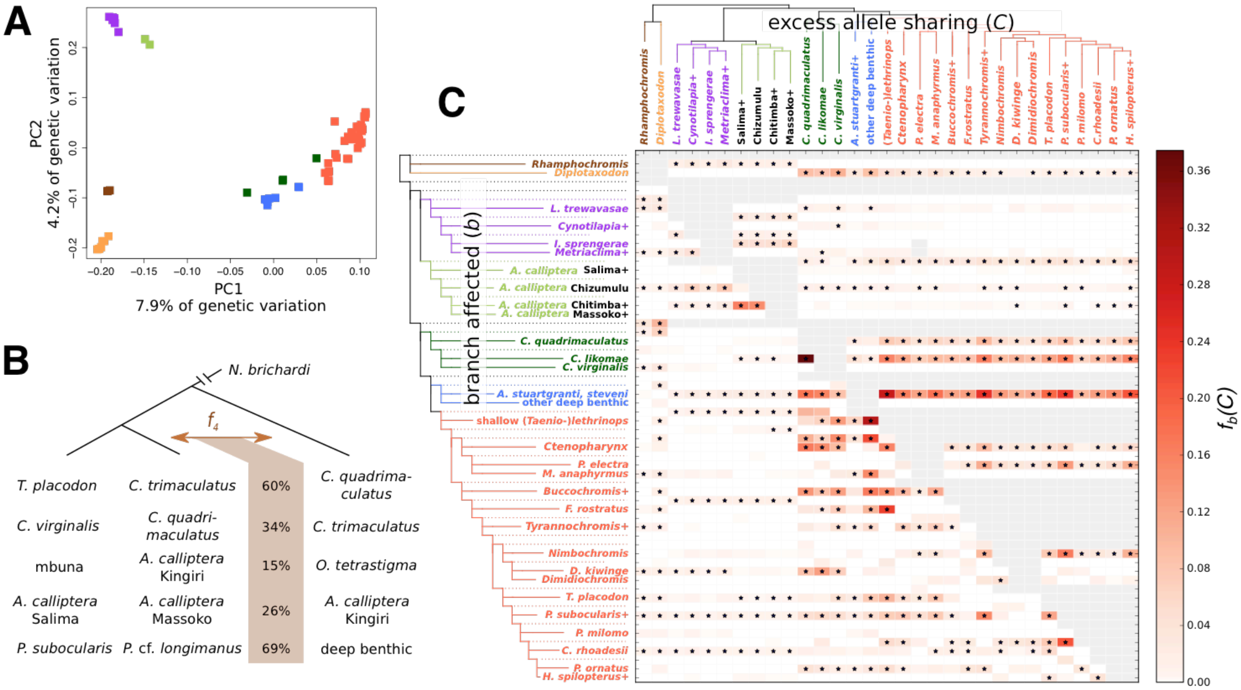

Fig. 3: Evidence for gene flow and non-tree like ancestral structure. Colors correspond to eco/morphological groups as in Fig. 1. **(A)** PCA analysis. **(B)** Selected strong f_4 admixture ratios (block jack-knifing p-values < 10-300). More comparisons can be found in Table S8. **(C)** Branch-specific statistic f_b amongst 31 subgroups. The ++ signs in labels signify that the subgroup includes multiple or additional taxa. For a full list of samples corresponding to each subgroup see Fig. S12. The * sign denotes block jack-knifing significance at |Z|>3.17 (Holm-Bonferroni FWER<0.001). Grey data points in the matrix correspond to tests that are not consistent with the NJ phylogeny.

To more exhaustively investigate cross-species gene flow and violation of a tree-like model of species relationships, we split the samples into 31 groups monophyletic with respect to the phylogeny shown in Fig. 2A and extended the *f* analysis to all trios of species groups that fit the relationships ((A, B), C) in that tree, excluding the strongly admixed taxa for whom *f* was already calculated above (*C. trimaculatus*, *A. calliptera* Kingiri, *O. tetrastigma, P.* cf. *longimanus*). Out of 4495 computed *f* statistics 3370 were significant at FWER < 0.001 (Fig. S10). However, a single gene flow event can lead to multiple significant *f* statistics: the values calculated for different combinations of ((A, B), C) groups are not independent as soon as they share internal or external branches. Therefore, we sought to obtain branch-specific estimates of excess allele sharing that would be less correlated. Building on the logic employed to understand correlated gene flow signals in ref. (*31*), we developed $f_b(C)$: a summary of *f* scores that captures excess allele sharing between a clade *C* and a branch *b* compared to the sister branch of *b* (*16*). The $f_b(C)$ score can be interpreted as a measure of how well a tree-like branching pattern captures the genetic relationships of samples descending from *b* against the rest of the phylogeny. This method reduces the number of calculated *f* statistics to 1,384. Of these, 351 scores are still significantly elevated (at FWER<0.001), while 39 of the 60 branches in the phylogeny show significant excess allele sharing with at least one other group *C* (Fig. 3C).

Pronounced population structure within ancestral species, coupled with rapid succession of speciation events, could also substantially violate the assumptions of a strictly bifurcating species tree and lead to significantly elevated *f* scores (*29*, *32*). However, to be able to explain many of the patterns reported here, ancestral population structure would need to segregate in remarkable ways through multiple speciation events, a scenario which appears unlikely. Therefore, we suggest that multiple cross-species gene flow events provide a more parsimonious explanation.

Not only are there many significant *f* scores, they also are unusually large: 183 out of the 351 significant scores (13% out of the total 1384; Fig. S11) are larger than 3%, corresponding to inferred human-neanderthal introgression (*28*). The strongest signal of $f_b(C)$ = 37% points to excess allele sharing of the branch leading to *C. likomae* with *C. quadrimaculatus* relative to *C. virginalis* suggests that the non-treelike relationships of utaka extend beyond the clearly admixed status of *C. trimaculatus* reported above. Underlining their complicated relationships, while *C. virginalis* and *C. quadrimaculatus* are sister species in Fig. 2A, removing *C. trimaculatus* leads to *C. likomae* becoming sister clade of *C. virginalis* with *C. quadrimaculatus* being the outgroup, as seen in Fig. 3C.

Several highly significant $f_b(C)$ scores point to multiple genetic exchanges between the deep and shallow benthic groups. *Aulonacara stuartgranti* and *A. steveni*, which overall cluster with deep benthics and have enlarged lateral line sensory apparatus like many of those, share excess derived alleles with all shallow benthic subgroups [max $f_b(C)$ = 28% relative to other deep benthics], reflecting that they are typically found in shallower water (*20*). Conversely, shallow benthic species of the genera *Lethrinops* and *Taeniolethrinops* show excess allele sharing with the remaining deep benthic taxa [$f_b(C)$ = 30% relative to other shallow benthics]. These signals suggest a combination of at least two gene flow events, one of which was between the common ancestors of shallow benthics and the *Aulonacara* species *A. stuartgranti* and *A. steveni*, explaining the inconsistent position of these *Aulonocara* in different phylogenies (compare Figs. 1C, 2A, S4, S6).

On a broader scale, there are strong signals of genetic exchange between major ecological groups. Most prominently, we infer excess allele sharing of all the utaka and benthics with *Diplotaxodon*, one of the two pelagic clades [$f_b(C)$ = 10% relative to mbuna and *A. calliptera*], as previously suggested by the PCA plot (Fig 3A). Furthermore, there is evidence for additional *Diplotaxodon* ancestry in utaka [$f_b(C)$ = 7%] and sub-clades of the benthics [most strongly for deep benthics at $f_b(C)$ = 3%] relative to their sister clades, which could be explained either by additional more recent gene flow events or by differential fixation of introgressed material, possibly due to selection. Reciprocally, *Diplotaxodon* shows excess allele sharing with all utaka and benthics (not significant for *Dimidiochromis kiwinge*) relative to *Rhamphochromis*.

Inference using the software treemix (*33*) also suggests strong evidence for gene flow within and between the major clades, mainly involving *Diplotaxodon*, deep and shallow benthics and utaka (Fig. S13). However, the overall topology and specific gene flow events in treemix results depend heavily on the value of a parameter specifying the number of allowed migration events. Using simulations we have shown that the accuracy of treemix results can be extremely sensitive to erroneous inferences of the initial phylogeny, whereas the interpretation of $f_b(C)$ scores as a measure of violation of a tree-like branching model is more robust in this respect (*16*).

Overall, the many elevated $f_b(C)$ scores in Fig. 3C reveal extensive violations of the tree model both across and within major groups. We therefore conclude that the evolutionary history of the Lake Malawi radiation is characterized by multiple gene flow events at different times during its evolutionary history and that no single phylogenetic tree can adequately capture the evolutionary relationships within the species flock.

History of the radiation

A. calliptera is the riverine-lacustrine generalist most closely related to the Malawi radiation, and has been referred to as the 'prototype' for Lake Malawi cichlids (*20*). Previous phylogenetic analyses, using mtDNA and small numbers of nuclear markers, showed inconsistencies with respect to *A. calliptera* (compare refs. (*13*) and (*34*)), with one study suggesting the phylogenetic discordances could be explained by gene-flow (*13*). On the other hand, our whole genome data indicate a clear and consistent position of all the *A. calliptera* individuals from the Lake Malawi catchment as a sister clade to the mbuna. This placement nested deep within the phylogeny (instead of being a basal sister clade) appears incompatible with previous suggestions that it could be the direct descendant of the riverine-generalist lineage that seeded the Lake Malawi radiation [e.g. refs. (*7*, *13*, *35*, *36*)]. It does not appear that the position within the radiation phylogeny is due to later hybridization: although the allele sharing patterns between *A. calliptera*, mbuna, and the rest of the radiation suggest some genetic exchange (Fig. 3C), the signals are relatively modest and insufficient to imply erroneous placement of the whole mbuna or *A. calliptera* clades in all phylogenetic reconstructions.

Therefore, we set out to acquire additional evidence regarding the origins of the Lake Malawi radiation. We obtained whole genome sequences from 19 individuals from seven *Astatotilapia* species not found in Lake Malawi (Table S2) and generated new variant calls (*16*). In addition, we sequenced five more *A. calliptera* specimens from Indian Ocean catchments, thus covering most of the geographical distribution of the species. We constructed NJ trees based on genetic distances and found that even with these additional data all the *A. calliptera* (including samples from outside the Lake Malawi catchment) remain as a single clade at the same place within the radiation as previously, whereas the other *Astatotilapia* species were clear outgroups (Fig. 4A,B). *A. calliptera* individuals cluster by geography (Fig. 4B,C), except for the sample from Kingiri, whose position in the clade is likely a result of the admixture signals shown in Fig. 3B. Indeed, the Kingiri individual clusters according to geography with Massoko and Mbaka if a NJ tree is built with *A. calliptera* samples only (Fig. S8).

We decided to further confirm that the position of *A. calliptera* is not an artifact of using trees based on genome-wide average distances. To do this, we specifically searched for the basal branch in a set of 2638 local ML phylogenies for non-overlapping genomic windows and found results that are consistent with the whole genome NJ tree: the most common basal branches are the pelagic groups *Rhamphochromis* and *Diplotaxodon* (in 42.51% of the genome). In comparison, *A. calliptera* (including all of the Indian Ocean catchment samples) were found to be basal only in 5.63% of the genome (Fig. 4D).

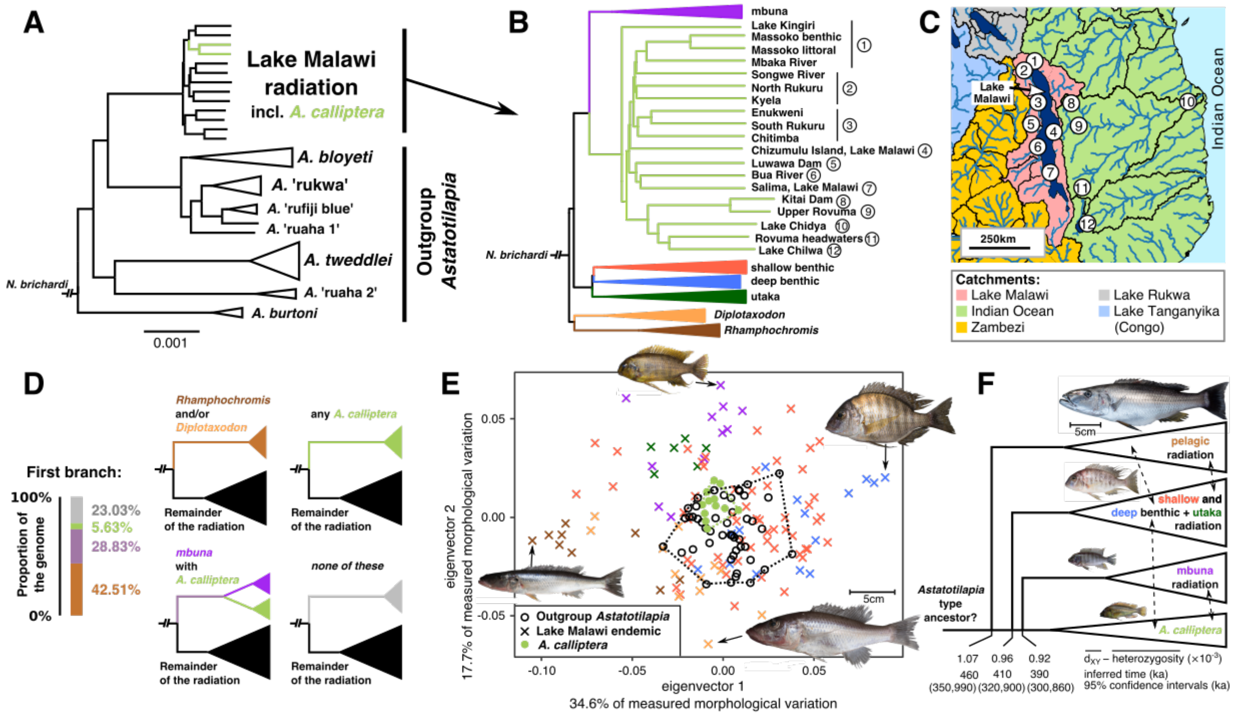

Fig. 4: Early history of the radiation. (A) An NJ phylogeny showing the Lake Malawi radiation in the context of East African *Astatotilapia* outgroups. **(B)** A Lake Malawi NJ phylogeny with expanded view of *A. calliptera* and all other groups collapsed. **(C)** Approximate *A. calliptera* sampling locations shown on a map of the broader Lake Malawi region. Black lines correspond to present day level 3 catchment boundaries from the US Geological Survey's HYDRO1k dataset. **(D)** Variation in the basal (or 'first') branch among ML phylogenies for 2638 non-overlapping genomic windows. This result is consistent with the tree in panel (B). **(E)** PCA of body shape variation of Lake Malawi endemics, *A. calliptera* and *Astatotilapia* outgroups, obtained from geometric morphometric analysis. **(F)** A phylogeny with the same topology as in panel (b) but displayed with a straight line between the ancestor and *A. calliptera*. For each branch off this lineage, we show mean sequence divergence (d_{XY}) minus mean heterozygosity, translation of this value into a mean time estimate, and 95% CI for the time estimate reflecting the statistical uncertainty in mutation rate. Dashed lines with arrows indicate likely instances of gene flow between major groups; their absolute timing (position along the x axis) is arbitrary.

Using geometric morphometrics with 17 homologous body shape landmarks (*16*) we established that, despite the relatively large genetic divergence, *A. calliptera* is nested within the morphospace of the other more distantly related but ecologically similar *Astatotilapia* species (Figs. 4A,E). This, and its central position within the morphological space of the Lake Malawi radiation (Figs. 4E, S9) are consistent with the ancestral species having a *Astatotilapia*-like riverine-lacustrine generalist phenotype. We suggest that the phylogenetic position of *A. calliptera* in Fig. 4B can be explained by a model in which the Lake Malawi species flock consists of three separate radiations; the pelagic radiation was seeded first, then the benthic + utaka, and finally the rock-dwelling mbuna, all in a relatively quick succession, followed by subsequent gene flow as described above (Fig. 4F).

Using our per-generation mutation rate we obtained mean separation time estimates for these lineages between 460 thousand years ago (ka) [95%CI: (350ka to 990ka)] and 390ky [95%CI: (300ka to 860ka)] (Fig. 4F), assuming three years per generation as in ref. (*37*). The point estimates all fall within the second most recent prolonged deep lake phase as inferred from the Lake Malawi paleoecological record (*38*) while the upper ends of the confidence intervals cover the third deep lake phase. We also note that split times estimated from sequence divergence are likely to be reduced by subsequent gene-flow, leading to underestimates. Therefore we conclude that the data are consistent with the previous reports based on fossil time calibration which put the origin of the Lake Malawi radiation at 700-800ka (*10*).

The fact that even the most diverged of the *A. calliptera* have a relatively recent common ancestor, with divergence at ~75% of the most distinct species in the Malawi radiation and corresponding to 340ka [95%CI: (260ka, 740ka)], suggests that the Lake Malawi population has been a reservoir that has repopulated the river systems and more transient lakes following dry-wet transitions in East African hydroclimate (*38*, *39*). Our results do not fully resolve whether the lineage leading from the common ancestor to *A. calliptera* retained its riverine generalist phenotype throughout or whether a lacustrine species evolved at some point (e.g. the common ancestor of *A. calliptera* and mbuna) and later de-specialized again to recolonize the rivers. However, while it is a possibility, it does not appear likely that the many strong phenotypic affinities of *A. calliptera* to outgroup *Astatotilapia* [see refs. (*40*, *41*); Fig. 4E], would be reinvented from a lacustrine species.

Signatures and consequences of selection

To gain insight into the functional basis of diversification and adaptation in Lake Malawi cichlids, we next turned our attention to protein coding genes. We compared the between-species levels of non-synonymous variation $\bar{p}_N$ to synonymous variation $\bar{p}_S$ in over 20,000 genes and calculated the difference between these two values ($\delta_{N-S} = \bar{p}_N - \bar{p}_S$) (*16*). Overall, coding sequence exhibits signatures of purifying selection: the average between-species $\bar{p}_N$ was 54% lower than in a random matching set of non-coding regions. Interestingly, the average between-species synonymous variation $\bar{p}_S$ in genes was slightly but significantly higher than in non-coding controls (13% lower mean; $p < 2.2{\times}10^{-16}$, one tailed Mann-Whitney test). A possible explanation of this observation is that protein coding regions are more resistant to gene flow.

Average per-gene non-synonymous excess variation (δ_{N-S}) calculated between Lake Malawi species correlates only relatively weakly with δ_{N-S} for genes in the five cichlid genome assemblies presented in Brawand *et al.* (*14*) which represent one from each of the major lineages and radiations of East African cichlids, and are approximately an order of magnitude more divergent than Malawi cichlids (Spearman $\rho_S = 0.32$; Fig. S14A). Thus the majority of genic selection within the Malawi radiation appears distinct from selection acting on these longer timescales between radiations. However, Malawi between-species δ_{N-S} correlates substantially more ($\rho_S = 0.49$; Fig. S14B) with δ_{N-S} calculated between *A. calliptera* populations, suggesting that the direction of within-Malawi selection could have been influenced by the diversity of alleles present in the common ancestor of the radiation. An alternative explanation could be that different *A. calliptera* populations are acquiring these alleles through gene flow with the derived Lake Malawi species.

To control for statistical effects stemming from variation in gene length and sequence composition we normalized the δ_{N-S} values per gene by taking into account the variance in $p_N - p_S$ across all pairwise sequence comparisons for each gene, deriving the non-synonymous excess score (Δ_{N-S}) (*16*). The genes with highly positive Δ_{N-S} are likely to be under positive selection. We focus below on the top 5% of the Δ_{N-S} distribution ($\Delta_{N-S} > 40.2$, 1034 candidate genes; Fig. 5A).

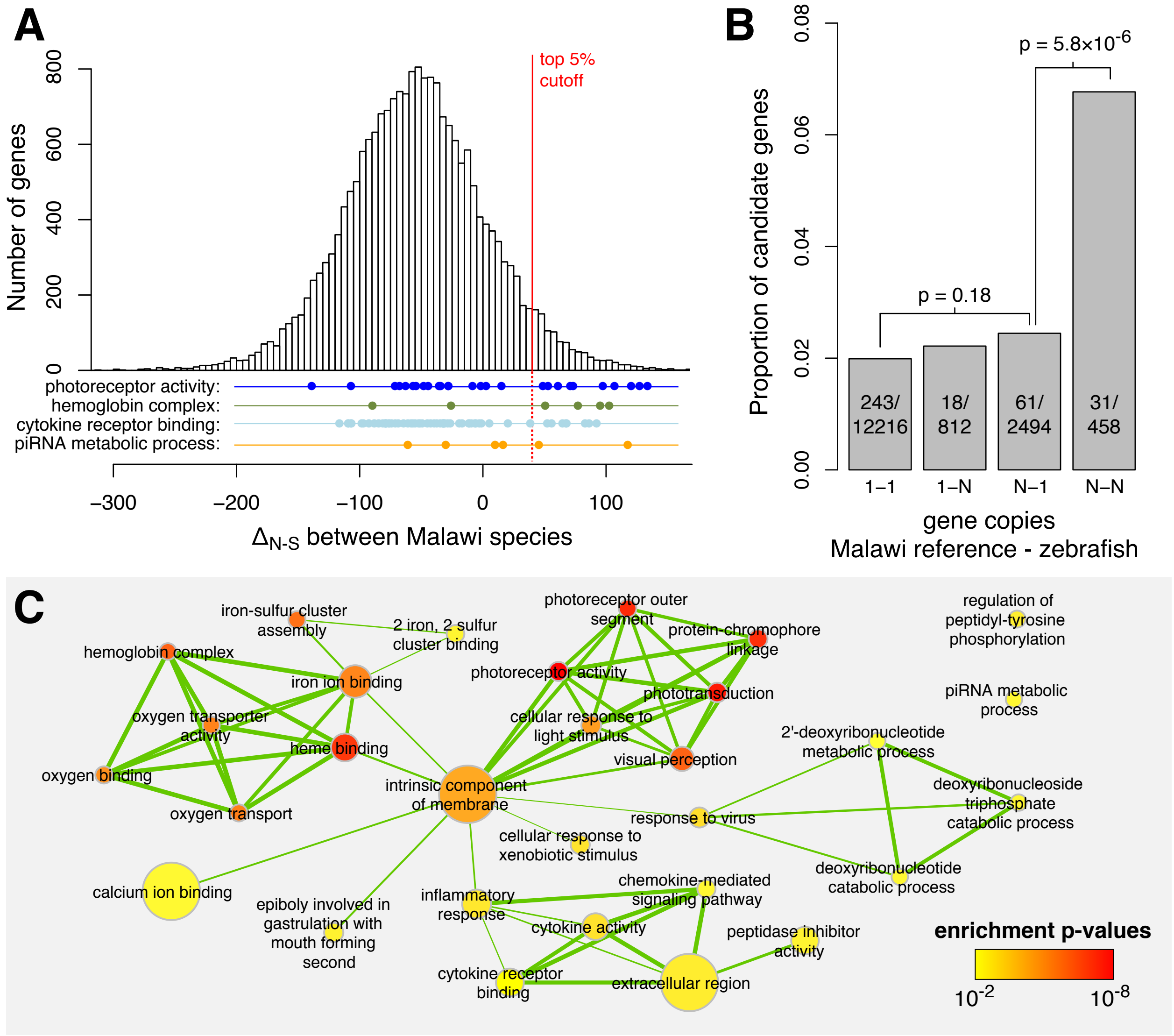

Figure 5: Gene selection scores, copy numbers, and ontology enrichment. (A) The distribution of the non-synonymous variation excess scores (**Δ_{N-S}**) highlighting the 5%FDR cutoff, and the distributions of genes in selected Gene Ontology (GO) categories. **(B)** The relationship between the probability of **Δ_{N-S}** being positive and in the top 5% and the relative copy numbers of genes in the Lake Malawi reference (*M. zebra*) and zebrafish. The p-values are based on χ^2 tests of independence. **(C)** An enrichment map for significantly enriched GO terms (cutoff at $p \leq 0.01$). The level of overlap between GO enriched terms is indicated by the thickness of the edge between them. The color of each node indicates the p-value for the term and the size of the node is proportional to the number of genes annotated with that GO category.

This candidate gene set is highly enriched for genes for which no homologs were found in any of Medaka, Stickleback, Tetraodon or Zebrafish (other teleosts) when examined in ref. (*14*) (606 out of 4,190 without vs. 428 out of 16,472 with homology assignment; χ^2 test $p < 2.2\times10^{-16}$).

Genes without homologs tend to be short (median coding length is 432bp) and some of the signal may be explained by a component of gene prediction errors. However a comparison of short genes (≤450bp) without homologs to a set of random noncoding sequences (Fig. S15) showed significant differences ($p < 2.2\times10^{-16}$, Mann-Whitney test), with both a substantial component of genes with low $\bar{p}_N$, reflecting genes under purifying selection, and also an excess of genes with high $\bar{p}_N$ (Fig. S16).

Cichlids have an unexpectedly large number of gene duplicates and it has been suggested that this phenomenon has contributed to their extensive adaptive radiations (*3*, *14*). To investigate the extent of divergent selection on gene duplicates, we examined how the non-synonymous excess scores are related to gene copy numbers in the reference genomes. Focusing on homologous genes annotated both in the Malawi reference (*M. zebra*) and in the zebrafish genome, we found that the highest proportion of candidate genes was among genes with many to many (N - N) relationships between the two genomes. The relative enrichment in this category is both substantial and highly significant (Fig. 5B). On the other hand, the increase in proportion of candidate genes in the N - 1 category (multiple copies in the *M. zebra* genome but only one copy in zebrafish) relative to 1 - 1 genes, is of a much lesser magnitude and is not significant (χ^2 test $p < 0.18$), suggesting that selection is occurring more often within ancient multi-copy gene families.

Next we used Gene Ontology (GO) annotation of zebrafish homologs to test whether candidate genes are enriched for particular functional categories. We found significant enrichment for 30 GO terms [range: $1.6\times10^{-8} < p < 0.01$, `weigh` algorithm (*16*, *42*); Table S3], 10 in the Molecular Function (MF), 4 in the Cellular Component (CC), and 16 in Biological Process (BP) category. Combining the results from all three GO categories in a network connecting terms with high overlap (i.e. sharing many genes) revealed clear clusters of enriched terms related to (i) hemoglobin function and oxygen transport; (ii) phototransduction and visual perception; and (iii) the immune system, especially inflammatory response and cytokine activity (Fig. 5C) .

Shared mechanisms of depth adaptation

To gain insight into the distribution of adaptive alleles across the radiation, we examined the haplotype genealogies for amino acid sequences of candidate genes, focusing on the genes in significantly enriched GO categories. It became apparent that many of the genealogies in the 'visual perception' category have common features that are unusual in the broader dataset: the haplotypes from the deep benthic group and the deep-water pelagic *Diplotaxodon* tend to be disproportionally diverse when compared with the rest of the radiation, and tend to group together despite these two groups being relatively distant in the whole-genome phylogenetic reconstructions.

Sharply decreasing levels of dissolved oxygen and low light intensities with narrow short wavelength spectra are the hallmarks of the habitats at below ~50 meters to which the deep benthic and pelagic *Diplotaxodon* groups have both adapted, either convergently or in parallel (*43*). Signatures of selection on similar haplotypes in the same genes involved in vision and in oxygen transport would therefore point to shared molecular mechanisms underlying this ecological parallelism.

To obtain a quantitative measure of shared molecular mechanisms, we calculated for each gene a similarity score for deep benthic and *Diplotaxodon* amino acid sequences and also compared the amounts of non-coding variation in these depth-adapted groups against the rest of the radiation (*16*). Both measures are elevated for candidate genes in the 'visual perception' category (Fig. 6A; p=0.007 for similarity, p=0.08 for shared diversity, and p=0.003 when similarity and diversity scores are added; all p-values based on Mann-Whitney test). The measures are also elevated for the 'hemoglobin complex' category, although due to the small number of genes the differences are not statistically significant in this case. Furthermore, the level of excess allele sharing between *Diplotaxodon* and deep benthic [measured by the local *f* statistic f_{dM} (*30*, *37*)] is strongly correlated with the $\Delta_{\text{N-S}}$ selection score for genes annotated with photoreceptor activity and hemoglobin complex GO terms ($\rho_S = 0.63$ and 0.81, $p = 0.001$ and $p = 0.051$, respectively, Fig. 6B).

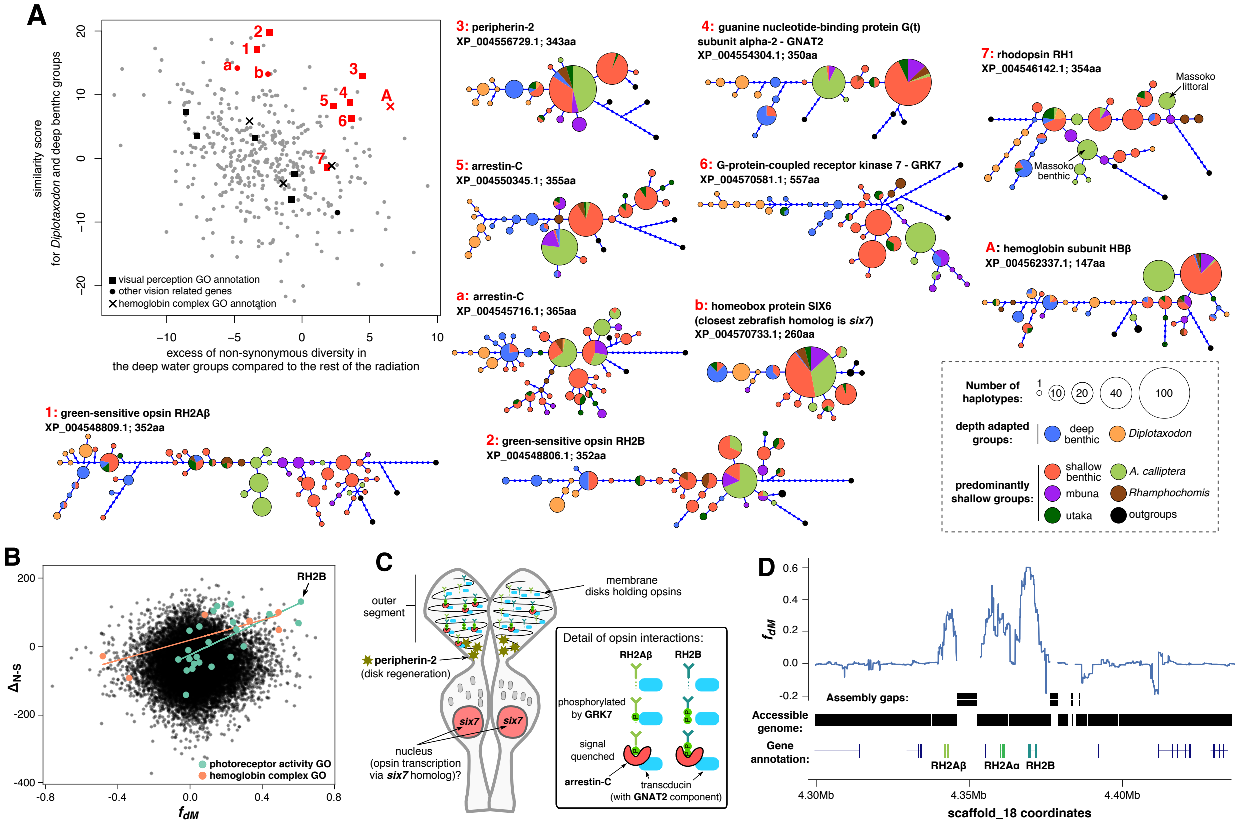

Fig. 6: Shared selection between the deep water adapted groups *Diplotaxodon* and deep benthic. (A) The scatterplot shows the distribution of genes with high Δ_{N-S} scores (candidates for positive selection) along axes reflecting shared selection signatures. Only genes with zebrafish homologs are shown. Amino acid haplotype genealogies are shown for genes as indicated by the red symbols and numbers. Outgroups include *Oreochromis niloticus*, *Neolamprologus brichardi*, *Astatotilapia burtoni*, and *Pundamilia nyererei*. **(B)** Selection scores plotted against f_{dM} (mbuna, deep benthic, *Diplotaxodon, N. brichardi*), a measure of excess allele sharing between deep benthic and *Diplotaxodon*. **(C)** A schematic drawing of a double cone photoreceptor expressing the green sensitive opsins and illustrating the functions of other genes with signatures of shared selection. **(D)** f_{dM} calculated in sliding windows of 100 SNPs around the green opsin cluster, revealing that excess allele sharing between deep benthic and *Diplotaxodon* extends far beyond the coding sequences.

Vision genes with high similarity and diversity scores for the deep benthic and *Diplotaxodon* groups include three opsin genes: the green sensitive opsins RH2Aβ, RH2B, and rhodopsin (Figs. 6A, S17A). The specific residues that distinguish the deep adapted groups from the rest of the radiation differ between the two RH2 copies, with only one shared mutation out of a possible fourteen (Fig. S17B). RH2Aβ and RH2B are located within 40kb from each other on the same chromosome (Fig. 6C); a third paralog, RH2Aα, is located between them, but it has very little coding diversity specific to deep benthic and *Diplotaxodon* (Fig. S18). This finding is consistent with previous reports suggesting functional divergence between RH2Aα and RH2Aβ following the duplication of RH2A early in the cichlid lineage (*44*, *45*). A similar, albeit weaker signature of shared depth-related selection is apparent in rhodopsin, which is known to play a role in deep-water adaptation in cichlids (*46*). Previously, we discussed the role of coding variants in rhodopsin in the early stages of speciation of *A. calliptera* in the crater Lake Massoko (*37*). The haplotype genealogy presented here for the broader radiation strongly suggests that the Massoko alleles did not originate by mutation in that lake but were selected out of ancestral variation (Fig. 6A).

The long wavelength, red sensitive opsin (LWS) has been shown to play a role in speciation along a depth gradient in Lake Victoria (*47*). While it is not particularly diverse in *Diplotaxodon* and deep benthic, it is interesting to note that *Diplotaxodon* have haplotypes that are clearly distinct from those in the rest of the radiation, while the majority of deep benthic haplotypes are their nearest neighbors (Fig. S18). The short wavelength opsin SWS1 is among genes with high $\Delta_{N\text{-}S}$ scores but does not exhibit shared selection between *Diplotaxodon* and deep benthic - it is most variable within the shallow benthic group. Finally, the short wavelength opsins SWS2A and SWS2B have negative $\Delta_{N\text{-}S}$ scores in our Lake Malawi dataset and thus are not among the candidate genes.

There have been many previous studies of selection on opsin genes in fish [e.g. reviewed in (*48*-*50*)], including selection associated with depth preference, but having whole genome coverage allows us to investigate other components of primary visual perception in an unbiased fashion. We found shared patterns of selection between deep benthic and *Diplotaxodon* in the genealogies of six other vision associated candidate genes: a homolog of the homeobox protein *six7*, the G-protein-coupled receptor kinase GRK7, two copies of the retinal cone arrestin-C, the α subunit of cone transducin GNAT2, and peripherin-2 (Fig. 6A). The functions of these genes suggest a prominent role of cone cell vision in depth adaptation. The homeobox protein *six7* governs the expression of RH2 opsins and is essential for the development of green cones in zebrafish (*51*). One of the variants in this gene that distinguishes deep benthic and *Diplotaxodon* is just a residue away from the DNA binding site of the HOX domain, while another is located in the SIX1_SD domain responsible for binding with the transcriptional activation co-factor of *six7* (*52*) (Fig. S17C). The kinase GRK7 and the retinal cone arrestin-C genes have complementary roles in photoresponse recovery, where arrestin produces the final shutoff of the cone pigment following phosphorylation by GRK7, thus determining the temporal resolution of motion vision (*53*). Note that bases near to the C terminus in RH2Aβ mutated away from serine (S290Y and S292G), thus reducing the number of residues that can be modified by GRK7 (Fig. S17B). The transducin subunit GNAT2 is located exclusively in the cone receptors and is a key component of the pathway which converts light stimulus into electrical response in these cells (*54*). The final gene, peripherin-2, is essential to the development and renewal of the membrane system in the outer cell segments that hold the opsin pigments in both rod and cone cells (*55*). Cichlid green sensitive opsins are expressed exclusively in double cone photoreceptors and the wavelength of maximum absorbance in cells expressing a mixture of RH2Aβ with RH2B (λ_{max} = 498nm) corresponds to the part of light spectrum that transmits the best into deep water in Lake Malawi (*50*). Figure 6C illustrates the possible interactions of all the above genes in a double cone photoreceptor of the cichlid retina.

Hemoglobin genes in teleost fish are located in two separate chromosomal locations: the minor 'LA' cluster and the major 'MN' cluster (*56*). The region around the LA cluster has been highlighted by selection scans among four *Diplotaxodon* species by Hahn et al. (*57*), who also noted the similarity of the hemoglobin subunit beta (HBβ) haplotypes between *Diplotaxodon* and deep benthic species. We confirmed signatures of selection in the two annotated LA cluster hemoglobins. In addition, we found that four hemoglobin subunits (HBβ1, HBβ2, HBα2, HBα3) from the MN cluster are also among the genes with high selection scores (Fig. S19). It appears that shared patterns of depth selection may be particular to the β-globin genes (Fig. S19B), although this hypothesis must remain tentative due to the highly repetitive nature of the MN cluster limiting our ability to confidently examine variation in all the hemoglobin genes in the region.

A key question concerns the mechanism leading to related haplotypes between *Diplotaxodon* and deep benthics. Possibilities include parallel selection on variation segregating in both groups due to common ancestry, selection on the gene flow that we described in a previous section, or independent selection on new mutations. From considering the haplotype genealogies and f_{dM} statistics summarizing local patterns of excess allele sharing, there is evidence for each of these processes acting, for different genes. The haplotype genealogies for rhodopsin and HBβ have outgroups at multiple locations on their haplotype networks and discontinuous positions of *A. calliptera* (Fig. 6A), suggesting that their haplotype diversity may reflect ancient differences in the founders. In contrast, networks for the green cone genes show patterns more consistent with the Malawi radiation being all derived with respect to outgroups (or with us not having sampled a source of ancestral variation) and we found substantially elevated f_{dM} scores extending for around 40kb around the RH2 cluster (Fig. 6D), consistent with adaptive introgression in a pattern reminiscent of mimicry loci in *Heliconius* butterflies (*58*). In contrast, the peaks in f_{dM} scores around peripherin-2 and one of the arrestin-C genes are relatively narrow, with boundaries that correspond almost exactly to the gene boundaries. Furthermore, these two genes have elevated f_{dM} scores only for non-synonymous variants Fig. S20, while synonymous variants do not show any excess allele sharing between *Diplotaxodon* and deep benthic. Due to the close proximity of non-synonymous and synonymous sites within the same gene, this suggests that for these two genes there may have been independent selection on the same *de novo* mutations.

Discussion

Genome sequences form the substrate for evolution. Here we have described genome variation at the full sequence level across the Lake Malawi haplochromine cichlid radiation. We focused on diversity, representing more than half the genera from each top-level lineage rather than obtaining deep coverage of any particular group. Therefore, we have more samples from the morphologically highly diverse benthic lineages than for example the species-rich mbuna.

The observation that cichlids within an African Great Lake radiation are genetically very close is not new (*59*), but we now quantify the relationship of this to within-species variation, and the consequences for variation in local phylogeny across the genome. The observation of within-species diversity being relatively low for vertebrates, at around 0.1%, suggests that low genome-wide nucleotide diversity levels are not necessarily limiting factors for rapid adaptation and speciation. This contrasts with the suggestion that high diversity levels may have been important for rapid adaptation in Atlantic killifish (*60*). One possibility is that in cichlids repeated selection has maintained diversity in adaptive alleles for a range of traits that support ecological diversification, as appears to be the case for sticklebacks (*61*).

We provide evidence that gene flow during the radiation has been extensive, but it does not appear to be ubiquitous. We see only one strong and clear example of recent gene flow between more distantly related species, not within Lake Malawi itself but between *Otopharynx tetrastigma* from crater Lake Ilamba and local *A. calliptera*. Lake Ilamba is very turbid and this apparent admixture is reminiscent of cichlid admixture in low visibility conditions in Lake Victoria (*62*). It is possible that some of the earlier gene flow signals we observed in Lake Malawi may have happened during low lake level periods when the water is known to have been more turbid (*38*).

Our suggested model of the early stages of radiation in Lake Malawi (Fig. 4F) is broadly consistent with the model of initial separation by major habitat divergence (*12*), although we propose a refinement in which there were three relatively closely-spaced separations from a generalist *Astatotilapia* type lineage, initially of pelagic genera *Rhamphochromis* and *Diplotaxodon*, then of shallow and deep water benthics and utaka [a clade which includes Kocher's sand dwellers (*12*, *20*)], and finally of mbuna. Thus, Lake Malawi appears to contain three haplochromine cichlid radiations that have separate origins, but are interconnected by subsequent gene flow.

The finding that cichlid-specific gene duplicates do not tend to diverge particularly strongly in coding sequences (Fig. 5B), suggests that other mechanisms of diversification following gene duplications may be more important in cichlid radiations. Divergence via changes in expression patterns has been illustrated and discussed in ref. (*14*). Future studies that address larger scale structural variation between cichlid genomes will be able to assess the contribution of differential retention of duplicated genes.

The evidence concerning shared adaptation of the visual and oxygen transport systems to deep water environments between deep benthic and *Diplotaxodon* suggests different evolutionary mechanisms for different genes, even within the same cellular system. It will be interesting to see whether the same genes or even specific mutations underlie depth adaptation in Lake Tanganyika, which harbors specialist deep water species in least two different tribes (*63*) and has a similar light attenuation profile but a steeper oxygen gradient than Lake Malawi (*43*).

Overall, our data and results provide unprecedented information about patterns of sequence sharing and adaptation across one of the most dramatic adaptive radiations, providing insights into mechanisms of rapid phenotypic diversification. The data sets we have generated are openly available (see Acknowledgements) and will underpin further studies on specific taxa and molecular systems. Given the extent of shared variation, we suggest that future studies that take into account variation within as well as between species will be important in future to reveal finer scale details of adaptive selection.

References and Notes:

1. J. B. Losos, R. E. Ricklefs, Adaptation and diversification on islands. *Nature*. **457**, 830–836 (2009).

2. C. E. Wagner, L. J. Harmon, O. Seehausen, Ecological opportunity and sexual selection together predict adaptive radiation. *Nature*. **487**, 366–369 (2012).

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

4. C. Darwin, *On the Origin of Species* (OUP Oxford, 2008).

5. S. Lamichhaney *et al.*, Evolution of Darwin/'s finches and their beaks revealed by genome sequencing. *Nature*. **518**, 371–375 (2015).

6. J. Losos, T. Jackman, A. Larson, K. Queiroz, L. Rodriguez-Schettino, Contingency and determinism in replicated adaptive radiations of island lizards. *Science*. **279**, 2115–2118 (1998).

7. G. Fryer, T. D. Iles, *The cichlid fishes of the great lakes of Africa: their biology and distribution* (Oliver and Boyd, 1972).

8. W. Salzburger, B. Van Bocxlaer, A. S. Cohen, Ecology and Evolution of the African Great Lakes and Their Faunas. *Annu. Rev. Ecol. Evol. Syst.* **45**, 519–545 (2014).

9. A. Meyer, Phylogenetic relationships and evolutionary processes in East African cichlid fishes. *Trends Ecol. Evol. (Amst.)*. **8**, 279–284 (1993).

10. B. S. Meyer, M. Matschiner, W. Salzburger, Disentangling incomplete lineage sorting and introgression to refine species-tree estimates for Lake Tanganyika cichlid fishes. *Syst. Biol.* (2016), doi:10.1093/sysbio/syw069.

11. J. I. Meier *et al.*, Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat Commun*. **8**, 14363 (2017).

12. T. D. Kocher, Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.* **5**, 288–298 (2004).

13. D. A. Joyce *et al.*, Repeated colonization and hybridization in Lake Malawi cichlids. *Curr. Biol.* **21**, R108–9 (2011).

14. D. Brawand *et al.*, The genomic substrate for adaptive radiation in African cichlid fish. *Nature*. **513**, 375–381 (2014).

15. P. Moran, I. Kornfield, P. N. Reinthal, Molecular Systematics and Radiation of the Haplochromine Cichlids (Teleostei: Perciformes) of Lake Malawi. *Copeia*. **1994**, 274 (1994).

16. See supplementary materials.

17. E. M. Leffler *et al.*, Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* **10**, e1001388 (2012).

18. R. C. Albertson, J. A. Markert, P. D. Danley, T. D. Kocher, Phylogeny of a rapidly evolving clade: the cichlid fishes of Lake Malawi, East Africa. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 5107–5110 (1999).

19. Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. **437**, 69–87 (2005).

20. A. Konings, *Malaŵi Cichlids in Their Natural Habitat* (Cichlid Press, ed. 4, 2007).

21. V. Ravi, B. Venkatesh, Rapidly evolving fish genomes and teleost diversity. *Curr. Opin. Genet. Dev.* **18**, 544–550 (2008).

22. H. Recknagel, K. R. Elmer, A. Meyer, A hybrid genetic linkage map of two ecologically and morphologically divergent Midas cichlid fishes (Amphilophus spp.) obtained by massively parallel DNA sequencing (ddRADSeq). *G3 (Bethesda)*. **3**, 65–74 (2013).

23. L. Ségurel, M. J. Wyman, M. Przeworski, Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet*. **15**, 47–70 (2014).

24. J. Heled, R. R. Bouckaert, Looking for trees in the forest: summary tree from posterior samples. *BMC Evol. Biol.* **13**, 221 (2013).

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

26. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet.* (2012).

27. 1000 Genomes Project Consortium *et al.*, An integrated map of genetic variation from

1,092 human genomes. *Nature*. **491**, 56–65 (2012).

28. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *Science*. **328**, 710–722 (2010).

29. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).

30. S. H. Martin, J. W. Davey, C. D. Jiggins, Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257 (2015).

31. S. H. Martin *et al.*, Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* **23**, 1817–1828 (2013).

32. A. Eriksson, A. Manica, Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13956–13960 (2012).

33. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).

34. M. J. Genner, G. F. Turner, Ancient hybridization and phenotypic novelty within Lake Malawi's cichlid fish radiation. *Mol. Biol. Evol.* **29**, 195–206 (2012).

35. D. H. Eccles, E. Trewavas, *Malawian Cichlid Fishes* (Lake Fish Movies, 1989).

36. E. N. Peterson, M. E. Cline, E. C. Moore, N. B. Roberts, R. B. Roberts, Genetic sex determination in Astatotilapia calliptera, a prototype species for the Lake Malawi cichlid radiation. *Naturwissenschaften*. **104**, 41 (2017).

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

38. S. J. Ivory *et al.*, Environmental change explains cichlid adaptive radiation at Lake Malawi over the past 1.2 million years. *Proceedings of the National Academy of Sciences*. **113**, 11895–11900 (2016).

39. R. P. Lyons, C. A. Scholz, A. S. Cohen, Continuous 1.3-million-year record of East African hydroclimate, and implications for patterns of evolution and biodiversity. *Proceedings of the …*. **112**, 15568–15573 (2015).

40. P. H. Greenwood, *Towards a phyletic classification of the"genus" Haplochromis (Pisces, Cichlidae) and related taxa* (1979).

41. E. Lippitsch, A phyletic study on lacustrine haplochromine fishes (Perciformes, Cichlidae) of East Africa, based on scale and squamation characters. *Journal of Fish Biology*. **42**, 903–946 (1993).

42. A. Alexa, J. Rahnenführer, T. Lengauer, Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*. **22**, 1600–1607 (2006).

43. B. Van Bocxlaer, R. SCHULTHEIß, Does the decline of gastropods in deep water herald ecosystem change in Lakes Malawi and Tanganyika? *Freshwater …* (2012).

44. T. C. Spady *et al.*, Evolution of the cichlid visual palette through ontogenetic subfunctionalization of the opsin gene arrays. *Mol. Biol. Evol.* **23**, 1538–1547 (2006).

45. C. J. Weadick, B. S. W. Chang, Complex patterns of divergence among green-sensitive (RH2a) African cichlid opsins revealed by Clade model analyses. *BMC Evol. Biol.* **12**, 206 (2012).

46. T. Sugawara *et al.*, Parallelism of amino acid changes at the RH1 affecting spectral sensitivity among deep-water cichlids from Lakes Tanganyika and Malawi. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 5448–5453 (2005).

47. O. Seehausen *et al.*, Speciation through sensory drive in cichlid fish. *Nature*. **455**, 620–626 (2008).

48. J. K. Bowmaker, D. M. Hunt, Evolution of vertebrate visual pigments. *Current Biology*. **16**, R484–R489 (2006).

49. W. I. L. DAVIES, S. P. COLLIN, D. M. Hunt, Molecular ecology and adaptation of visual photopigments in craniates. *Mol Ecol*. **21**, 3121–3158 (2012).

50. K. L. Carleton, B. E. Dalton, D. Escobar-Camacho, S. P. Nandamuri, Proximate and ultimate causes of variable visual sensitivities: Insights from cichlid fish radiations. *Genesis*. **54**, 299–325 (2016).

51. Y. Ogawa, T. Shiraki, D. Kojima, Y. Fukada, Homeobox transcription factor Six7 governs expression of green opsin genes in zebrafish. *Proceedings of the Royal Society B: Biological Sciences*. **282**, 20150659 (2015).

52. A. Marchler-Bauer *et al.*, CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).

53. S. L. Renninger, M. Gesemann, S. C. F. Neuhauss, Cone arrestin confers cone vision of high temporal resolution in zebrafish larvae. *Eur. J. Neurosci.* **33**, 658–667 (2011).

54. S. E. Brockerhoff *et al.*, Light stimulates a transducin-independent increase of cytoplasmic Ca2+ and suppression of current in cones from the zebrafish mutant nof. *J. Neurosci.* **23**, 470–480 (2003).

55. K. Boesze-Battaglia, A. F. X. Goldberg, Photoreceptor renewal: a role for peripherin/rds. *Int. Rev. Cytol.* **217**, 183–225 (2002).

56. J. C. Opazo, G. T. Butts, M. F. Nery, J. F. Storz, F. G. Hoffmann, Whole-genome duplication and the functional diversification of teleost fish hemoglobins. *Mol. Biol. Evol.* **30**, 140–153 (2013).

57. C. Hahn, M. J. Genner, G. T. Turner, D. A. Joyce, The genomic basis of adaptation to the deep water"twilight zone"in Lake Malawi cichlid fishes. *bioRxiv* (2017).

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

59. A. Meyer, T. D. Kocher, P. Basasibwaki, A. C. Wilson, Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. *Nature*. **347**, 550–553 (1990).

60. N. M. Reid *et al.*, The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science*. **354**, 1305–1308 (2016).

61. F. C. Jones *et al.*, The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*. **484**, 55–61 (2012).

62. O. Seehausen, Cichlid Fish Diversity Threatened by Eutrophication That Curbs Sexual Selection. *Science*. **277**, 1808–1811 (1997).

63. A. Konings, *Tanganyika cichlids in their natural habitat* (Cichlid Press, ed. 3, 2015).

64. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. **q-bio.GN** (2013).

65. M. A. M. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

66. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

67. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. **27**, 2987–2993 (2011).

68. S. R. Browning, B. L. Browning, 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).

69. O. Delaneau, J. Marchini, J.-F. Zagury, A linear complexity phasing method for thousands of genomes. *Nat. Methods*. **9**, 179–181 (2012).

70. O. Delaneau, B. Howie, A. J. Cox, J.-F. Zagury, J. Marchini, Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93**, 687–696 (2013).

71. R. S. Harris, thesis, The Pennsylvania State University (2007).

72. W. Miller *et al.*, 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.* **17**, 1797–1808 (2007).

73. K. Ulm, A simple method to calculate the confidence interval of a standardized mortality ratio (SMR). *Am. J. Epidemiol.* **131**, 373–375 (1990).

74. A. J. Dobson, K. Kuulasmaa, E. Eberle, J. Scherer, Confidence intervals for weighted sums of Poisson parameters. *Stat Med*. **10**, 457–462 (1991).

75. A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. **22**, 2688–2690 (2006).

76. R. Bouckaert *et al.*, BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).

77. A. Stamatakis, P. Hoover, J. Rougemont, A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**, 758–771 (2008).

78. N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

79. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. **20**, 289–290 (2004).

80. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

81. J. Heled, A. J. Drummond, Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27**, 570–580 (2010).

82. S. Purcell *et al.*, PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. **81**, 559–575 (2007).

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

84. F. J. Rohlf, tpsDig2, (available at http://http//life.bio.sunysb.edu/morph/).

85. D. C. Adams, E. O. Castillo, geomorph: an R package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and …* (2013).

86. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

87. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

88. A. Alexa, J. Rahnenfuhrer, topGO: enrichment analysis for gene ontology. *R package version* (2010).

89. W. Huber *et al.*, Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*. **12**, 115–121 (2015).

90. D. Merico, R. Isserlin, O. Stueker, A. Emili, G. D. Bader, Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE*. **5**, e13984 (2010).

91. A. J. Ribbink, B. A. Marsh, A. C. Marsh, A. C. Ribbink, B. J. Sharp, A preliminary survey of the cichlid fishes of rocky habitats in Lake Malawi. *S. Afr. J. Zool.* **18**, 149–310 (1983).

92. J. Snoeks, A. Konings, Eds., *The cichlid diversity of Lake Malawi/Nyasa/Niassa* (Cichlid Press, ed. 4, 2004).

93. M. J. Genner *et al.*, Reproductive isolation among deep-water cichlid fishes of Lake Malawi differing in monochromatic male breeding dress. *Mol Ecol*. **16**, 651–662 (2007).

94. D. A. Joyce *et al.*, An extant cichlid fish radiation emerged in an extinct Pleistocene lake. *Nature*. **435**, 90–95 (2005).

95. J. Schwarzer *et al.*, Repeated trans-watershed hybridization among haplochromine cichlids (Cichlidae) was triggered by Neogene landscape evolution. *Proceedings of the Royal Society B: Biological Sciences*. **279**, 4389–4398 (2012).

96. J. Kelleher, A. M. Etheridge, G. McVean, Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).

Acknowledgments: This work was supported by the Wellcome Trust (097677/Z/11/Z to MM, WT098051 to RD and HS), the Royal Society-Leverhulme Trust Africa Awards (AA100023 and AA130107 to MG and GFT) and the European Molecular Biology Organization (ALTF 456-2016 to MM). Raw sequencing reads are in the Sequence Read Archive: (BioProjects PRJEB1254 and PRJEB15289: sample accessions listed in Table S4). Whole-genome variant calls in the Variant Call Format (VCF), phylogenetic trees and protein coding sequence alignments are available from the Dryad Digital Repository (http://dx.doi.org/accession). RD declares that he owns stock in Illumina from previous consulting. We want to thank the Sanger Institute sequencing core for DNA sequencing; Walter Salzburger and Ian Wilson for comments on the manuscript, and the Tanzania Fisheries Research Institute and the Malawi Government Fisheries Research Unit for support. EM, GFT, MJG, MM and RD devised the study. GFT and MJG collected the samples. AMT bred parent-offspring trios and performed geometric morphometric analyses. MM performed the DNA extractions. HS and MM analyzed the genomic data. All authors participated in interpretation of the results. MM, HS and RD drafted the manuscript, and all others commented.

Supplementary Materials:

Materials and Methods

Supplementary Text

Figures S1-S30

Tables S1-S8

References (*64-96*)