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

Malinsky, Milan; Svardal, Hannes; Tyers, Alexandra; Miska, Eric.A.; Genner, Martin J.; Turner, George; Durbin, Richard

Nature Ecology and Evolution

DOI:
10.1038/s41559-018-0717-x

Published: 01/12/2018

Peer reviewed version

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

Authors: Milan Malinsky[1,2]†*, Hannes Svardal[1,3,4,5]†, Alexandra M. Tyers[6], Eric A. Miska[1,3,7], Martin J. Genner[8], George F. Turner[6], and Richard Durbin[1,3]*

Affiliations:

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

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

[3]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK.

[4]Department of Biology, University of Antwerp, 2020 Antwerp, Belgium.

[5]Naturalis Biodiversity Center, 2300 RA Leiden, The Netherlands.

[6]School of Natural Sciences, Bangor University, Bangor, LL57 2UW, UK.

[7]Gurdon Institute, University of Cambridge, Cambridge, CB2 1QN, UK.

[8]School of Biological Sciences, University of Bristol, Bristol BS8 1TQ, UK.

*Correspondence to: millanek@gmail.com (MM), rd@sanger.ac.uk (RD).

†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.

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 seen for example in Darwin's finches in the Galapagos[4,5], Anolis lizards of the Caribbean[6] and in East African cichlid fishes[7,8].

Cichlids are one of the most species-rich and diverse families 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[9]. 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[10,11], between some major lineages in Lake Tanganyika[12], and in describing the role of hybridization in the origins of the Lake Victoria radiation[13]. However, the task of reconstructing within-lake relationships remains challenging due both to retention of large amounts of ancestral genetic polymorphism (i.e. incomplete lineage sorting) and gene flow between taxa[12,14-18].

Initial genome assemblies of cichlids from East Africa suggest that an increased rate of gene duplication together with accelerated evolution of some regulatory elements and protein coding genes may have contributed to the radiations[11]. 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. The species that comprise the radiation can be divided into seven groups with differing ecology and morphology (see Supplementary Note): 1) the rock-dwelling 'mbuna'; 2) *Rhamphochromis* - typically midwater pelagic 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 lake bottom (here utaka corresponds to the genus *Copadichromis*); 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, as well as river systems to the east and south of the Lake Malawi catchment. This division into seven groups has been partially supported by previous molecular phylogenies based on mtDNA and amplified fragment length polymorphism (AFLP) data[18-20]. However, published phylogenies show numerous inconsistencies and, in particular, the question of whether the groups are genetically separate remained unanswered.

To characterize the genetic diversity, species relationships, and signatures of selection across the whole radiation, we obtained Illumina whole-genome sequence data from 134 individuals of 73 species distributed broadly across the seven groups (Fig. 1a; Supplementary Note). This includes 102 individuals at ~15× coverage and 32 additional individuals at ~6× (Supplementary Table 1).

Results

Low genetic diversity and species divergence

Sequence data were aligned to and variants called against a *Metriaclima zebra* reference genome[11]. Average divergence from the reference was 0.19% to 0.27% (Supplementary Fig. 1). 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. All the following analyses are based on biallelic SNPs.

To estimate nucleotide diversity (π) within the species, 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[21]. There does not appear to be a 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* (Supplementary Fig. 1).

Despite their extensive phenotypic differentiation, species within the Lake Malawi radiation are genetically closely related[22,23]. However, genome-wide genetic divergence has never been quantified. We calculated the average pairwise sequence differences (d_{XY}) between species and compared d_{XY} against heterozygosity, finding that the two distributions partially overlap (Fig. 1b). Thus, the sequence divergence within a single diploid individual is sometimes higher than the divergence between 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 therefore approximately one fifth of the divergence between human and chimpanzee[24]. In addition to the low ratio of divergence to diversity, most genetic variation is shared between species. On average both alleles are observed in other species for 82% of heterozygous sites within individuals, consistent with the expected and previously observed high levels of incomplete lineage sorting (ILS)[23]. Supplementary Fig. 2 shows d_{XY} and F_{ST} values for comparisons between the seven eco-morphological groups and Supplementary Fig. 3 shows patterns of linkage disequilibrium across the radiation, within groups, and within individual species.

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[25,26]. To obtain a direct estimate of the per-generation mutation rate, we reared offspring of three species from three different Lake Malawi groups (*A. 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 (Supplementary Fig. 4). 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 in human[27], although, given much shorter mean generation times, the per-year rate is still expected to be higher in cichlids than in humans. We note that Recknagel et al.[26] obtained a much higher mutation rate estimate (6.6×10^{-8} per bp per generation) in Midas cichlids, but from relatively low depth RADseq data that may have made accurate verification more difficult. We also note that our per generation rate estimate, although low, is still higher than the lowest μ estimate in vertebrates: 2×10^{-9} per bp per generation recently reported for Atlantic herring[28]. By combining our 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 $N_e = \pi/4\mu$).

Genome data support for eco-morphological groupings

Principal component analysis (PCA) of the whole-genome genotype data generally separates the major eco-morphological groups (Fig. 1c). The most notable exceptions to this are (1) the utaka, for which some species cluster more closely with deep benthics and others with shallow benthics, and (2) two species of the genus *Aulonacara*, *A. stuartgranti* and *A. steveni*, which are located between the shallow and deep benthic groups. Although these have enlarged lateral line sensory apparatus like many deep benthic species including other *Aulonocara*, they are typically found in shallower water[29]. Another interesting pattern in the PCA plot is that the utaka and benthic samples are often spread along principal component (PC) axes (Fig. 1c, Supplementary Fig. 5), a pattern typical for admixed populations (e.g. ref. 30). Along the two main PCs, the deeper water benthic species extend towards the deep water *Diplotaxodon*, an observation we will return to in the context of gene flow and shared mechanisms of depth adaptation.

To further verify the consistency of group assignments, we tested whether pairs of species from the same group always share more derived alleles with each other than with any species from other groups. Group assignments were again supported, except for the four species also highlighted in the PCA: the two shallow-living *Aulonocara* are closer to shallow benthics than to deep benthics in 71% and 82% of tests respectively when comparing these alternatives, and *Copadichromis trimaculatus* is closer to shallow benthics than to utaka in 58% of the comparisons. *Copadichromis cf. trewavasae* always clustered with shallow benthics; therefore, we treat it as a member of the shallow benthic group throughout the remainder of this manuscript. With the three intermediate samples removed and *C. cf. trewavasae* reassigned, all other species showed 100% consistency with their group assignment.

Allele sharing inconsistent with tree-like relationships

The above observations suggest that some species may be genetic intermediates between well-defined groups, consistent with previous studies which have suggested that hybridization and introgression subsequent to initial separation of species may have played a significant role in cichlid radiations, including in Lakes Tanganyika[12,14-16] and Malawi[18,20]. Where this happens, there is no single tree relating the species.

To assess the overall extent of violation of tree-like species relationships, we calculated Patterson's D statistic (ABBA-BABA test)[31,32] for all possible trios of Lake Malawi species, without assuming any *a priori* knowledge of their relationships. *N. brichardi* from Lake Tanganyika was always used as the outgroup. The test statistic D_{min} is the minimum absolute value of Patterson's D for each trio, across all possible tree topologies. Therefore, a significantly positive D_{min} score signifies that the sharing of derived alleles between the three species is inconsistent with a single species tree relating them, even in the presence of incomplete lineage sorting.

Overall, 62% of trios (75,616 out of 121,485) have significantly positive D_{min} score (Holm-Bonferroni FWER < 0.01). The D_{min} values are not independent: for example, a single gene flow event between ancestral lineages can affect multiple contemporary species and thus more trios than a more recent gene flow event would. However, tree violations are numerous and pervasive throughout the dataset, within all the major groups and also between groups (Fig. 2a), revealing reticulate evolution at multiple levels. Therefore, phylogenetic trees alone cannot fully describe the evolutionary relationships of Lake Malawi cichlids.

Phylogenetic framework

Despite no tree giving a complete and accurate picture of the relationships between species, standard phylogenetic approaches are useful to provide a framework for discussion. To obtain an initial picture 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 the full sequences within each window, obtaining trees with 2542 different topologies. We also calculated the maximum clade credibility (MCC) summary tree[33] and an ML phylogeny based on the full mtDNA genome (Fig. 1c and Supplementary Fig. 6).

We next applied a range of further phylogenomic methods which are known to be robust to incomplete lineage sorting. These included three multispecies coalescent methods[34,35]: the Bayesian SNAPP[36] (with a subset of 48,922 unlinked SNPs in 12 individuals representing the eco-morphological groups), the algebraic method SVDquartets[37,38], which allows for site-specific rate variation and is robust to gene-flow between sister taxa[39], and the summary method

ASTRAL[40,41], using the 2543 local ML trees that were described above as input. We also built a whole genome Neighbour-Joining (NJ) tree using the Dasarathy et al. algorithm, which has been shown to be a statistically consistent and accurate species tree estimator under ILS[42,43]. The above methods have also been applied to datasets where the individuals that are genetically intermediate between eco-morphological groups (*C. trimaculatus*, *A. stuartgranti*, and *A. steveni*) have been removed, thus likely reducing the extent of violation of the multispecies coalescent model.

Despite extensive variation among the 2543 individual ML trees (at least in part attributable to ILS), and, to a lesser extent, variation between the different genome-wide phylogenetic methods, there is some general consensus (Fig. 2c and Supplementary Figs. 6-10). Except for the three previously identified intermediate species, individuals from within each of the previously identified eco-morphological groups cluster together in all the whole genome phylogenies, forming well supported reciprocally monophyletic groups. The pelagic *Diplotaxodon* and *Rhamphochromis* together form a sister group to the rest of the radiation, except in the all-sample MCC and SVDquartets phylogenies. Perhaps surprisingly, all the methods place the generalist *A. calliptera* as the sister taxon to the specialized rocky-shore mbuna group in a position that is nested within the Lake Malawi radiation. On a finer scale, many similarities between the resulting phylogenies reflect features of previous taxonomic assignment, but some currently-recognized genera are always polyphyletic, including *Placidochromis*, *Lethrinops*, and *Mylochromis*.

The mtDNA phylogeny is an outlier, substantially different from all the whole-genome phylogenies and also from the majority of the local ML trees (Fig. 2b,c and Supplementary Figs. 6 and 11). Discordances between mtDNA and nuclear phylogenies in Lake Malawi have been reported previously and interpreted as a signature of past hybridization events[18,20]. However, as we discuss below, some of these previously suggested hybridization events are not reflected in the whole genome data. Indeed, large discrepancies between mitochondrial and nuclear phylogenies have been shown in many other systems, reflecting both that mtDNA as a single locus is not expected to reflect the consensus under ILS, and high incidence of mitochondrial selection[44-46]. This underlines the importance of evaluating species relationships in the Lake Malawi radiation from a genome-wide perspective.

Specific signals of introgression

We applied a variety of methods to identify the species and groups whose relationships violate the framework trees described in the previous section. First, we contrasted the pairwise genetic distances used to produce the NJ tree against the distances between samples along the tree branches, calculating the residuals (Supplementary Fig. 12). If the tree captured all the genetic relationships in our sample perfectly, the residuals would all be zero. However, as expected in the light of the D_{min} analysis above, we found numerous differences, affecting both groups of species and individual species, with some standout cases. Among the strongest signals on individual species, in addition to the previously discussed *C. trimaculatus*, we can see that 1) *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 2) our sample of *Otopharynx tetrastigma* (from Lake Ilamba) is much closer to *A. calliptera* (especially to the sample from Lake Kingiri, only 3.2km away) than expected from the tree.

Second, the 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[47] and calculated the 'coancestry matrix' of all species - a summary of nearest neighbour (therefore recent) haplotype relationships. The Lake Ilamba *O. tetrastigma* and Lake Kingiri *A. calliptera* also stand out in this analysis, showing a strong signature of recent gene flow between individual species from distinct eco-morphological groups (Supplementary Fig. 13). The other tree-violation signatures described above are also visible on the haplotype sharing level but are less pronounced, consistent with being older events involving the common ancestors of multiple present-day species. However, the chromopainter results indicate additional recent introgression events (e.g. the utaka *C. virginalis* with *Diplotaxodon*; more highlighted in Supplementary Fig. 13). Furthermore, the clustering based on recent co-ancestry is different from all phylogenetic trees: in particular a number of shallow benthics, including *P.* cf. *longimanus*, cluster next to the deep benthics.

Third, we used the f_4 admixture ratio[31,32,48] (*f* statistic; closely related to Patterson's D), computing *f*(A,B;C,O) for all groups of species that fit the relationships ((A, B), C) in the ASTRAL* tree (Supplementary Fig. 7), with the outgroup fixed as *N. brichardi*. When elevated

due to introgression, the *f* statistic is expected to be linear in relation to the proportion of introgressed material. The ASTRAL* tree has the lowest mean topological distance to all the other trees, and excludes the three species with intermediate group assignment, a choice made here because we were interested in identifying additional signals beyond the admixed status of *A. stuartgranti*, *A. steveni*, and *C. trimaculatus*. Out of the 164,320 computed *f* statistics, 97,889 were significant at FWER < 0.001.

As in the case of D_{min}, a single gene flow event can lead to multiple significant *f* statistics. Noting that the values for different combinations of ((A, B), C) groups are not independent as soon as they share branches on the tree, 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. 49, we developed "*f*-branch" or $f_b(C)$: a summary of *f* scores that, on a given tree, captures excess allele sharing between a species *C* and a branch *b* compared to the sister branch of *b* (Methods). Therefore, an $f_b(C)$ score is specific to the branch *b* (on the y-axis in Fig. 3), but a single introgression event can still lead to significant $f_b(C)$'s across multiple related C. There were 11,158 $f_b(C)$ scores of which 1,421 were significantly elevated at FWER<0.001 (Supplementary Fig. 14), and 238 scores were larger than 3% (the value inferred for human-Neanderthal introgression in ref. 31). The majority of nodes in the tree are affected: 92 of the 158 branches in the phylogeny show significant excess allele sharing with at least one other species *C* (Fig. 3).

Overall, the highest $f_b(C)$ (14.2%) is between the ancestor of the two sampled *Ctenopharynx* species from the shallow benthic group and the utaka *Copadichromis virginalis* (Fig. 3). Notably, *Ctenopharynx* species, particularly *C. intermedius* and *C. pictus* have very large numbers of long slender gill rakers, a feature shared with *Copadichromis* species, and believed to be related to a diet of small invertebrates[50]. Several other benthic lineages also share excess alleles with *C. virginalis* at a lower level. Next, the significantly elevated $f_b(C)$ scores between the shallow and the deep benthic lineages suggest that genetic exchanges between these two groups go beyond the clearly admixed shallow-living *Aulonacara* (not included in this analysis). The *f*-branch signals between *O. tetrastigma* and *A. calliptera* Kingiri are observed in both directions – *A. calliptera* Kingiri with shallow benthics (and most strongly *O. tetrastigma*) and *O. tetrastigma* with *A. calliptera* (most strongly *A. calliptera* Kingiri), suggesting bi-directional introgression.

At the level of the major eco-morphological groups, the strongest signal indicates that the ancestral lineage of benthics and utaka shares excess derived alleles with *Diplotaxodon* and, to a lesser degree, *Rhamphochromis*, as previously suggested by the PCA plot (Fig 1c). Furthermore, there is evidence for additional ancestry from the pelagic groups in utaka, which could be explained either by an additional, more recent, gene flow event or by differential fixation of introgressed material, possibly due to selection. Reciprocally, *Diplotaxodon* shares excess derived alleles (relative to *Rhamphochromis*) with utaka and deep benthics, as does *Rhamphochromis* with mbuna and *A. calliptera*. Furthermore, mbuna show excess allele sharing (relative to *A. calliptera*) with *Diplotaxodon* and *Rhamphochromis*. (Fig. 3) On the other hand, while ref. 18 suggested gene flow between the deep benthic and mbuna groups on the basis of a discrepancy between mtDNA and nuclear phylogenies, our genome-wide analysis did not find any signal of substantial genetic exchange between these groups.

The *f* statistic tests are robust to the occurrence of incomplete lineage sorting, in the sense that ILS alone cannot generate a significant test result[32]. We note, however, that pronounced population structure within ancestral species, coupled with rapid succession of speciation events, can also substantially violate the assumptions of a strictly bifurcating species tree and lead to significantly elevated *f* scores[32,51]. This needs to be taken into account when interpreting non-treelike relationships, for example among major groups early in the radiation. However, in cases of excess allele sharing between 'distant' lineages that are separated by multiple speciation events, ancestral population structure would have needed to segregate through these speciation events without affecting sister lineages, a scenario that is not credible in general. Therefore, we suggest that there is strong evidence for multiple cross-species gene flow events. Additionally, simulations suggest that, compared with treemix[52], $f_b(C)$ is robust to misspecification of the initial tree (Supplementary Note).

Overall, the NJ tree residuals, the haplotype sharing patterns, and the many elevated $f_b(C)$ scores paint a consistent picture. They confirm the extensive violations of the bifurcating species tree model initially revealed by the D_{min} analysis, and suggest many independent gene flow events at different times during the evolutionary history of the adaptive radiation.

Origins of the radiation

The generalist *Astatotilapia calliptera* has been referred to as the 'prototype' for the endemic Lake Malawi cichlids[29,53], and discussions concerning the origin of the radiation often centre on ascertaining its relationship to the Malawi species[20,54]. Previous phylogenetic analyses, using mtDNA and small numbers of nuclear markers, showed inconsistencies in this respect[18,20,54]. In contrast, our whole genome data indicated a clear and consistent position of the Lake Malawi catchment *A. calliptera* as a sister group to the mbuna, in agreement with the nuclear DNA phylogeny in ref. 18. While it is not certain whether the 320 remaining mbuna species form a monophyletic group with the eight species we used here, the eight species represent the majority of the genera of mbuna and therefore are likely to be representative of much of the genetic diversity within the group.

To explore the origins of the Lake Malawi radiation in greater detail, we obtained 24 additional *Astatotilapia* whole genome sequences from outside of Lake Malawi: five *A. calliptera* from Indian Ocean catchments (IOC), thus covering most of its geographical distribution, and 19 individuals from seven other *Astatotilapia* species (Supplementary Table 2). We generated new variant calls (Supplementary Methods) and first constructed a NJ tree, finding that all the *A. calliptera* (including IOC) cluster as a single group nested at the same place within the radiation, whereas the other *Astatotilapia* species branched off well before the lake radiation (Fig. 4a,b,c). All *A. calliptera* individuals cluster by geography (Fig. 4b,c), except for the specimen from crater Lake Kingiri, whose position in the tree is likely a result of the admixture signals with *O. tetrastigma*. Indeed, a NJ tree built only with *A. calliptera* samples (Supplementary Fig. 17) places the Kingiri individual according to geography with the specimens from the nearby crater Lake Massoko and Mbaka River.

Applying the same logic as above, we tested whether the position of the *A. calliptera* group in the NJ tree changes when the tree is built without mbuna (as would be expected if *A. calliptera* were affected by hybridization with mbuna). We found that the position of *A. calliptera* is unaffected (Supplementary Fig. 18), suggesting that the nested position is not due to later hybridization. The *f* statistics in Fig. 3 further support this, because the signals involving the whole mbuna or *A. calliptera* groups are modest and do not suggest erroneous placement of these groups in all phylogenetic analyses. Furthermore, the nested position of *A. calliptera* is also

supported by the vast majority of the genome. Searching for the basal branch in a set of 2638 local ML phylogenies, we found results that agree with the whole genome ASTRAL, SNAPP and NJ trees: the most common basal branches are the pelagic groups *Rhamphochromis* and *Diplotaxodon* (in 42.12% of the genomic windows). In comparison, *A. calliptera* (including IOC samples) were found to be basal only in 5.99% of the windows (Supplementary Fig. 19).

Joyce et al.[20] reported that the mtDNA haplogroup of Indian Ocean catchment (IOC) *A. calliptera* clustered with mbuna (as we confirm in Supplementary Fig. 15) and suggested that there had been repeated colonization of Lake Malawi by two independent *Astatotilapia* lineages with different mitochondrial haplogroups: the first founding the entire species flock, and the second, with the IOC mtDNA haplogroup, introgressing into the Malawi radiation and contributing strongly to the mbuna. This hypothesis predicts that, compared with the Malawi catchment *A. calliptera*, the IOC *A. calliptera* should be closer to mbuna. However, across the nuclear genome we found a strong signal in the opposite direction, with 30% excess allele sharing between Malawi catchment *A. calliptera* and mbuna (Fig. 4d). Therefore, the Joyce et al. hypothesis that the mbuna, the most species rich group within the radiation, may be a hybrid lineage formed from independent invasions is not supported by genome-wide data.

It has been repeatedly suggested that *A. calliptera* may be the direct descendant of the riverine-generalist lineage that seeded the Lake Malawi radiation[7,50,53,54]. Our interpretation of this argument is that the ancestor was likely a riverine generalist that was ecologically and phenotypically similar to *A. calliptera* and other *Astatotilapia*. This hypothesis is lent further support by geometric morphometric analysis. Using 17 homologous body shape landmarks 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 (Fig. 4a,e), and these together have a central position within the morphological space of the Lake Malawi radiation (Fig. 4e and Supplementary Fig. 16).

To reconcile the nested phylogenetic position of *A. calliptera* with its generalist 'prototype' phenotype, we propose a model where the Lake Malawi species flock consists of three separate radiations splitting off from the lineage leading to *A. calliptera*. The relationships between the major groups supported by the ASTRAL, SNAPP and NJ methods suggest that 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; the pelagic vs. utaka + benthic branching order is swapped in SVDquartets tree in Supplementary Fig. 9b). Applying our per-generation mutation rate to observed genomic divergences we obtained mean divergence time estimates between these lineages between 460 thousand years ago (ka) [95%CI: (350ka to 990ka)] and 390ka [95%CI: (300ka to 860ka)] (Fig. 4f), assuming three years per generation as in ref. 55. The point estimates all fall within the second most recent prolonged deep lake phase inferred from the Lake Malawi paleoecological record[56] while the upper ends of the confidence intervals cover the third deep lake phase at ~800ky. Considering that our split time estimates from sequence divergence are likely to be reduced by subsequent gene-flow, leading to underestimates, the data are consistent with a previous report based on fossil time calibration which put the origin of the Lake Malawi radiation at 700-800ka[12].

The fact that the common ancestor of all the *A. calliptera* appears younger than the Malawi radiation suggests that the Lake Malawi *A. calliptera* population has been a reservoir that has repopulated the river systems and more transient lakes following dry-wet transitions in East African hydroclimate[56,57]. 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, we suggest it is unlikely that the many strong phenotypic affinities of *A. calliptera* to the basal *Astatotilapia* (Fig. 4e; refs. [58,59]) would be reinvented from a lacustrine species.

Signatures and consequences of selection on coding sequences

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 20,664 genes and calculated the difference between these two values ($\delta_{N-S} = \bar{p}_N - \bar{p}_S$). 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 13% higher than in non-coding control regions ($p < 2.2 \times 10^{-16}$, one tailed Mann-Whitney test). One possible explanation of this observation would

be if intergenic regions were homogenized by gene-flow, whereas protein coding genes were more resistant to this.

To control for statistical effects of variation in gene length and sequence composition we normalized the δ_{N-S} values per gene by taking into account the variance across all pairwise sequence comparisons for each gene, deriving the non-synonymous excess score Δ_{N-S} (see Methods). Values at the upper tail of the distribution of Δ_{N-S} are substantially overrepresented in the actual data when compared to a null model based on random sampling of codons (Fig. 5a). We focus below on the top 5% of the distribution ($\Delta_{N-S} > 40.2$, 1034 candidate genes). Genes with elevated Δ_{N-S} are expected to have been under positive selection at multiple non-synonymous sites, either recently repeatedly within multiple species or ancestrally. Therefore, the statistic reveals only a limited subset of positive selection events from the history of the radiation (e.g. a selection event on a single amino acid would not be detected). Furthermore, to minimise any effect of gene prediction errors, all the following analyses focus on the 15980 (77.3% of total) genes for which zebrafish homologs were found in ref. 11; selection scores of genes without homologs are briefly discussed in a Supplementary Note.

Cichlids have an unexpectedly large number of gene duplicates, which possibly contributed to their extensive adaptive radiations[3,11]. To investigate the extent of divergent selection on gene duplicates, we examined how the Δ_{N-S} 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 two or more copies in both genomes (N - N). 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) 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, rather than on genes with cichlid-specific duplications.

We used Gene Ontology (GO) annotation of zebrafish homologs to test whether candidate genes are enriched for particular functional categories (Methods). We found significant enrichment for 30 GO terms (range: $1.6\times10^{-8} < p < 0.01$, `weigh` algorithm[60]; Supplementary Table 3): 10 in the

Molecular Function, 4 in the Cellular Component and 16 in Biological Process category. Combining all the results in a network (connecting terms that share many genes) revealed clear clusters of enriched terms related to (i) haemoglobin function and oxygen transport; (ii) phototransduction and visual perception; and (iii) the immune system, especially inflammatory response and cytokine activity (Fig. 5c). That evolution of genes in these functional categories has contributed to cichlid radiations has been suggested previously (see below); it is nevertheless interesting that these categories stand out in a genome-wide analysis.

Shared mechanisms of depth adaptation

To gain insight into the distribution of adaptive alleles across the radiation, we built maximum likelihood trees from amino acid sequences of candidate genes, thus summarising potentially complex haplotype genealogy networks. Focusing on the significantly enriched GO categories, many haplotype trees have features that are unusual in the broader dataset: the haplotypes from the deep benthic group and the deep-water pelagic *Diplotaxodon* tend to group together (despite these two groups being distant in whole-genome phylogenies and monophyletic in only two out of 2638 local ML trees) and also tend to be disproportionally diverse when compared with the rest of the radiation. We quantified both excess similarity and diversity, and found that both measures are elevated for candidate genes in the 'visual perception' category (Fig. 6a; Mann-Whitney tests: p=0.007 for similarity, p=0.08 for shared diversity, and p=0.003 when the scores are added) and also for the 'haemoglobin complex' category (p values not significant due to the small number of genes).

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 *Diplotaxodon* groups have both adapted, either convergently or in parallel[61]. Shared signatures of selection in genes involved in vision and in oxygen transport therefore point to shared molecular mechanisms underlying this ecological parallelism. Further evidence of shared mechanisms of adaptation is that, for genes annotated with 'photoreceptor activity' and 'haemoglobin complex' GO terms, the $\Delta_{N\text{-}S}$ selection score is strongly correlated with the local levels of excess allele sharing between the two depth-adapted groups (Fig. 6b; $\rho_S = 0.63$ and 0.81, $p = 0.001$ and $p = 0.051$, respectively).

Vision genes with high similarity and diversity scores for the deep benthic and *Diplotaxodon* groups include three opsins: the green sensitive RH2Aβ and RH2B, and rhodopsin (Fig. 6a and Supplementary Fig. 20). 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 (Supplementary Fig. 20). RH2Aβ and RH2B are located less than 40kb apart on the same chromosome (Fig. 6c); a third paralog, RH2Aα, is located between them, but does not show signatures of shared depth adaptation (Supplementary Fig. 21), consistent with reports of functional divergence between RH2Aα and RH2Aβ[62,63]. 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[64]. Previously, we discussed the role of coding variants in rhodopsin in the early stages of speciation of *A. calliptera* in the crater Lake Massoko[55]. The haplotype tree presented here for the broader radiation shows that the Massoko alleles did not originate by mutation in that lake but were selected out of ancestral variation (Fig. 6a). The remaining opsin genes are less likely to be involved in shared depth adaptation (Supplementary Note).

There have been many studies of selection on opsin genes in fish[65-67], 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 benthics and *Diplotaxodon* in six other vision associated candidate genes (Fig. 6a). The functions of these genes, together with the fact that RH2Aβ and RH2B are expressed exclusively in double-cone photoreceptors, suggest a prominent role of cone cell vision in depth adaptation. The wavelength of maximum absorbance in cone 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[67].

Figure 6c illustrates interactions of the vision genes with shared selection patterns in the cichlid double-cone photoreceptor. The homeobox protein *six7* governs the expression of RH2 opsins and is essential for the development of green cones in zebrafish[68] (specific mutations are highlighted in Supplementary Fig. 20). The kinase GRK7 and the retinal cone arrestin-C have complementary roles in photoresponse recovery: arrestin produces the final shutoff of the cone pigment following phosphorylation by GRK7, thus determining the temporal resolution of motion vision[69]. 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 (Supplementary Fig. 20). 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[70]. Finally, peripherin-2 is essential to the development and renewal of the membrane system that holds the opsin pigments in both rod and cone cells[71].

Haemoglobin genes in teleost fish are found in two separate chromosomal locations: the minor 'LA' cluster and the major 'MN' cluster[72]. The region around the LA cluster has been highlighted by selection scans among four *Diplotaxodon* species by Hahn et al.[73], who also noted the similarity of the haemoglobin subunit beta (HBβ) haplotypes between *Diplotaxodon* and deep benthic species. We confirmed signatures of selection in the two annotated LA cluster haemoglobins. In addition, we found that four haemoglobin subunits (HBβ1, HBβ2, HBα2, HBα3) from the MN cluster are also among the genes with high selection scores (Supplementary Fig. 22). The shared patterns of depth selection may be particular to the β-globin genes (Supplementary Fig. 22), although this hypothesis remains tentative, because the repetitive nature of the MN cluster precludes us from confidently examining all haemoglobin genes.

A key question concerns the mechanism leading to the similarity of haplotypes in *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 trees and local patterns of excess allele sharing (f_{dM} statistics[55]), there is evidence for each of these processes acting on different genes. The haplotype trees for rhodopsin and HBβ have outgroup taxa (and also *A. calliptera*) appearing at multiple locations on their haplotype networks (Fig. 6a), suggesting that the haplotype diversity of these genes may reflect ancestral variation. In contrast, trees for the green cone genes show the Malawi radiation all being derived with respect to outgroups 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[74]. Finally, the peaks in f_{dM} around peripherin-2 and one of the arrestin-C genes are narrow, ending at the gene boundaries, and f_{dM} scores are elevated only for non-synonymous variants; synonymous variants do not show excess allele sharing (Supplementary Fig. 23). Due to the close proximity of non-synonymous and synonymous sites within the same gene, this

505 suggests that for these two genes there may have been independent selection on the same *de novo*
506 mutations.
507
508 ## Discussion

509 Variation in genome sequences forms the substrate for evolution. Here we described genome
510 variation at the full sequence level across the Lake Malawi haplochromine cichlid radiation. We
511 focused on ecomorphological diversity, representing more than half the genera from each major
512 group, rather than obtaining deep coverage of species within any particular group. Therefore, we
513 have more samples from the morphologically highly diverse benthic lineages than, for example,
514 the mbuna where there are relatively fewer genera and many species are largely recognised by
515 colour differences.

516 The observation that cichlids within an African Great Lake radiation are genetically very similar
517 is not new[75], but we now quantify the relationship of this to within-species variation, and the
518 consequences for variation in local phylogeny across the genome. The fact that between-species
519 divergence is generally only slightly higher than within species diversity, is likely the result of
520 the young age of the radiation, the relatively low mutation rate, and of gene flow between taxa.
521 Within-species diversity itself is relatively low for vertebrates, at around 0.1%, suggesting that
522 low genome-wide nucleotide diversity levels do not necessarily limit rapid adaptation and
523 speciation. This conclusion appears in contrast for example with a recent report that high
524 diversity levels may have been important for rapid adaptation in Atlantic killifish[76]. One
525 possibility is that in cichlids repeated selection has maintained diversity in adaptive alleles for a
526 range of traits that support ecological diversification, as we have concluded for rhodopsin and
527 HBβ and appears to be the case for some adaptive variants in sticklebacks[77].

528 We provide evidence that gene flow during the radiation, although not ubiquitous, has certainly
529 been extensive. Overall, the numerous violations of the bifurcating species tree model suggest
530 that full resolution of interspecies relationships in this system will require network approaches
531 (see e.g. ref. 35; section 6.2) and population genomic analyses within the framework of the
532 structured coalescent with gene flow. The majority of the signals affect groups of species,
533 suggesting events involving their common ancestors, or are between closely-related species
534 within the major ecological groups. The only strong and clear example of recent gene flow

between individual distantly-related species is not within Lake Malawi itself, but between *Otopharynx tetrastigma* from crater Lake Ilamba and local *A. calliptera*. Lake Ilamba is very turbid and the scenario is reminiscent of cichlid admixture in low visibility conditions in Lake Victoria[78]. It is possible that some of the earlier signals of gene flow between lineages we observed in Lake Malawi may have happened during low lake level periods when the water is known to have been more turbid[56].

Our 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[23], 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 (this includes Kocher's sand dwellers[23,29]), and finally of mbuna. Thus, we suggest that Lake Malawi contains three separate haplochromine cichlid radiations stemming from the generalist lineage, 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. Divergence via changes in expression patterns has previously been illustrated and discussed[11], and future studies addressing structural variation between cichlid genomes will 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 benthics and *Diplotaxodon* suggests different evolutionary mechanisms acting on 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 harbours specialist deep water species in least two different tribes[79] and has a similar light attenuation profile but a steeper oxygen gradient than Lake Malawi[61].

Over the last few decades, East African cichlids have emerged as a model for studying rapid vertebrate evolution[11,23]. Taking advantage of recently assembled reference genomes[11], our data and results provide unprecedented information about patterns of sequence sharing and adaptation across the Lake Malawi radiation, with insights into mechanisms of rapid phenotypic

diversification. The data sets are openly available (see Acknowledgements) and will underpin further studies on specific taxa and molecular systems. For example, we envisage that our results, clarifying the relationships between all the main lineages and many individual species, will facilitate speciation studies, which require investigation of taxon pairs at varying stages on the speciation continuum[80,81], and studies on the role of adaptive gene flow in speciation.

Methods

Samples. Ethanol preserved fin clips were collected by M.J. Genner and G.F. Turner between 2004 and 2014 from Tanzania and Malawi, in collaboration with the Tanzania Fisheries Research Institute (the MolEcoFish Project) and with the Fisheries Research Unit of the Government of Malawi (various collaborative projects). Samples were collected and exported with the permission of the Tanzania Commission for Science and Technology, the Tanzania Fisheries Research Institute, and the Fisheries Research Unit of the Government of Malawi

From sequencing to a variant callset. The analyses presented above are based on SNPs obtained from Illumina short (100bp-125bp) reads, aligned to the *Metriaclima zebra* reference assembly version 1.1[11] with `bwa-mem`[82], followed by GATK haplotype caller[83] and samtools/bcftools[84] variant calling restricted to 653Mb of 'accessible genome' where variants can be determined confidently with short reads, filtering, genotype refinement, imputation, and phasing in BEAGLE[85] and further haplotype phasing with `shapeit v2`[86], including the use of phase-informative reads[87]. For details please see Supplementary Methods.

Linkage disequilibrium (LD) calculations. Haplotype r^2 between pairs of SNPs was calculated along the phased scaffolds 0 to 201, using `vcftools v0.1.12b` with the options `--hap-r2 --ld-window-bp 50000`. To reduce the computational burden, we used a random subsample of 10% of SNPs. We binned the r^2 values according to the distance between SNPs into 1kb or 100bp windows and plotted the average values in each bin.

To estimate background LD, we calculated haplotype r^2 between variants mapping to different linkage groups (LG) in the *Oreochromis niloticus* genome assembly. First, we used the chain files generated by the whole genome alignment pipeline[88] (see Supplementary Methods) and the UCSC `liftOver` tool to translate the genomic coordinates of all SNPs to the *O. niloticus* coordinates. Then we calculated LD between variants mapping to LG1 and LG2.

***De novo* mutation rate estimation.** In each trio we looked for mutations in the child that were not present in either of its parents. Because the results of this analysis are very sensitive to false positives and false negative rates, we used higher coverage sequencing (~40x average) and applied more stringent genome masks than in the population genomic work. Increased coverage supports clean separation of sequencing errors and somatic mutations from true heterozygous calls in the offspring, and improved ability to distinguish single copy vs. multi-copy sequence on a per-individual basis.

First we determined the "Accessible Genome" (i.e. the regions of the genome where we can confidently call *de novo* mutations) for each trio by excluding:

1. Genomic regions where mapped read depth in any member of a trio is ≤25× or >50×

2. Bases where either of the parents has a mapped read that does not match the reference (the specific bases where any read has non-reference alleles in the parents were masked)
3. Sequences where indels were called in any sample (we also excluded +/- 3bp of sequence surrounding the indel)
4. Sites which were called as multiallelic among the nine samples in the overall trios dataset
5. Known segregating variable sites - i.e. sites with alternative alleles found in four and more copies in the overall Lake Malawi dataset
6. Sites in the reference where less than 90% of overlapping 50-mers (sub-sequences of length 50) could be matched back uniquely and without 1-difference. For this we used Heng Li's SNPable tool (http://lh3lh3.users.sourceforge.net/snpable.shtml), dividing the reference genome into overlapping k-mers (sequences of length k – we used k=50), and then aligning the extracted k-mers back to the genome (we used `bwa aln -R 1000000 -O 3 -E 3`).

After excluding sites in the categories above, we were left with an "Accessible Genome" of 516.6Mb in the *A. calliptera* trio, 459Mb in the *A. stuartgranti* trio and 404Mb in the *L. lethrinus* trio. Because any observed *de novo* mutation could have occurred either on the chromosome inherited from the mother or on the chromosome inherited from the father, the point estimate of the per generation per basepair mutation rate is: $\mu = nMutations/(2 \times AccessibleGenome)$.

Next we set out to search for *de novo* mutations: i.e. heterozygous sites in the offspring within the Accessible Genome. Under random sampling there is an equal probability of seeing a read with either of the two alleles at a heterozygous site. Therefore, N_a - the number of reads supporting the alternative allele is distributed as: ~Binomial(ReadDepth, 0.5). We filtered out variants with observed N_a below 2.5th or above 97.5th percentiles of this distribution, thus accepting a false negative rate of 5%. We also filtered out sites where the offspring call had Read Position or Base Quality rank sum test Z-score > 99.5th percentile of standard normal distribution or where the Strand Bias phred scaled p-value was ≥ 20 or where the phred scaled GQ (genotype quality) in either mother, father, or offspring was ≤ 30. For simplicity, assuming these filters are independent they are expected to introduce a false negative rate of 7.17%. The mutation rate estimate was adjusted to account for this.

After filtering, we found nine *de novo* mutations across the three offspring. For each mutation we double-checked the alignment in the IGV genome browser and found all of them were single base mutations supported by high number of reads (>8) in the offspring. The 95% confidence intervals for the number of observed mutations were calculated using the "exact" method relating chi-squared and Poisson distributions[89,90]. If N is the number of observed mutations, the lower (ciN_L) and upper (ciN_U) limits are:

$$ciN_L = \frac{P(\chi^2_{2N} \leq 0.025)}{2} \qquad ciN_U = \frac{P(\chi^2_{2(N+1)} \geq 0.975)}{2}$$

where 2N and 2(N+1) are degrees of freedom of the corresponding chi-squared distributions.

Principal Component Analysis. SNPs with minor allele frequency >= 0.05 were selected using the `bcftools (v1.2) view` option `--min-af 0.05:minor`. The program `vcftools v0.1.12b` was then used to export that data into PLINK format[91]. Next, the variants were LD-pruned to obtain a set of variants in approximate linkage equilibrium (unlinked sites) using the `--indep-pairwise 50 5 0.2` option in `PLINK v1.0.7`. Principal Component Analysis on the resulting set of variants was performed using the `smartpca` program from the `eigensoft v5.0.2` software package[92] with default parameters.

Genome-wide F_{ST} calculations. In addition to performing PCA, the `smartpca` program from the `eigensoft v5.0.2` software package also calculates genome-wide F_{ST} for all pairs of populations specified by the sixth column in the `.pedind` file. For the calculation, it uses the Hudson estimator, as defined by Bhatia, Patterson et al.[93] in equation 10, and the 'ratio of averages' is used to combine estimates of F_{ST} across multiple variants, as recommended in that manuscript. We used all SNPs (no minor allele frequency filtering).

Allele sharing test for group assignment. We tested if two individuals who come from the same group always share more derived alleles with each other than with any individuals from other groups. Technically, we implemented this using the D statistic (ABBA-BABA tests) framework[31,32], by calculating $D(A, G1, G2, O)$ for all permutations of individuals, where *G1* and *G2* come from the same eco-morphological group and *A* from a different group. The outgroup *O* was always *N. brichardi* from Lake Tanganyika. Note that this is an unusual use of the D statistics and our aim here was not to look for gene flow but to test if allele sharing is greater within eco-morphological groups (G1 with G2) compared to across groups (A with G2), in which case $D(A, G1, G2, O) > 0$. All results were statistically significant, which was assessed using block jackknife[31] on windows of 60k SNPs .

D_{min} statistic. Here we calculated the D statistic for each trio of species (A,B,C) and for all possible tree topologies (the outgroup again fixed as *N. brichardi*). Therefore, D_{min}=min(|D(A,B,C,O)|, |D(A,C,B,O)|, |D(C,B,A,O)|. If this is significantly elevated, then allele sharing within the trio of species is inconsistent with any simple tree topology. Note that this approach is conservative in the sense that the D_{min} score for each trio is considered in isolation and we ignore 'higher-order' inconsistencies where different D_{min} trio topologies are inconsistent with each other. Statistical significance was assessed using block jackknife[31] on windows of 60k SNPs and FWER was calculated following the Holm-Bonferroni method.

Sample selection for demographic analyses. To prevent potential confounding effects of uneven sequencing depth, we limited these analyses to one high coverage (15x) individual per species. Species without a high coverage sample (*P. subocularis, F. rostratus* and *L. trewavasae*) were not included.

Outgroup sequences/alleles. Outgroup (Supplementary Table 5) sequences in *M. zebra* genomic coordinates were obtained based on pairwise whole-genome alignments (Supplementary Methods). Insertions in the outgroup were ignored and deletions filled by `N` characters.

Local phylogenetic trees and maximum clade credibility. To generate a multiple alignment input in `fasta` format we used the `getWGSeq` subprogram of `evo` We set the window size in terms of the numbers of variants rather than physical length (8000 variants; `--split 8000` option) aiming for the local regions to have similar strengths of phylogenetic signal. Small windows at the ends of scaffolds were discarded. We limited the sequence output to the accessible genome using the `--accessibleGenomeBED` option. The *N. brichardi* outgroup sequence in *M. zebra* genomic coordinates was added via the `--incl-Pn` option.

Maximum likelihood phylogenies were inferred using `RAxML v7.7.8`[94] under the GTRGAMMA model. The best tree for each region was selected out of twenty alternative runs on distinct starting maximum parsimony trees (the `-N 20` option).

The maximum clade credibility (MCC) trees were calculated in `TreeAnnotator v.2.4.2`, a part of the BEAST2 platform[95]. Clade credibility is the frequency with which a clade appears in the tree set; the MCC tree is the tree (from among the trees in the set) that maximizes the product of the frequencies of all its clades[33]. The node heights for the MCC trees are derived as a summary from the heights of each clade in the whole tree set via the `Common Ancestor heights` option.

Mitochondrial (mtDNA) phylogenies. The mtDNA sequence corresponds to scaffolds 747 and 2036 in the *Metriaclima zebra* reference. Variants from these scaffolds were subjected to the same filtering as in the rest of the genome except for the depth filter because the mapped read depth was much higher (approximately 300-400x per sample). Because of the greater sequence diversity in the mtDNA genome, we found that more than 10% of variants were multiallelic. Therefore, we separated SNPs from indels at multiallelic sites using `bcftools norm` with the `--multiallelics -` option, then removed indels and the merged multiallelic SNPs back together with the `--multiallelics +` option. Sequences in the `fasta` format were generated using the `bcftools consensus` command, and missing genotypes in the VCF replaced by the `N` character with the `--mask` option. *N. brichardi* outgroup sequence in *M. zebra* genomic coordinates was added to the `fasta` files.

A maximum likelihood tree was inferred using `RAxML v7.7.8`[94] under the GTRGAMMA model. The best tree was selected out of twenty alternative runs on distinct starting maximum parsimony trees (using the `-N 20` option) and two hundred bootstrap replicates were obtained using `RAxML`'s rapid bootstrapping algorithm[96] satisfying the `-N autoFC` frequency-based bootstrap stopping criterion. Bipartition bootstrap support was drawn on the maximum likelihood tree using the `RAxML -f b` option.

Neighbour-joining trees and the residuals. For the Neighbour-Joining (NJ)[97] trees we calculated the average numbers of single nucleotide differences between haplotypes for each pair of species. This simple pairwise difference matrix was divided by the accessible genome size to obtain pairwise differences per bp, which are equivalent to $\hat{p}_{AB}$ of Dasarathy et al.[42]. Then we followed equation 8 from Dasarathy et al. and calculated their *corrected* measure of dissimilarity:

$$\hat{d}_{AB} = -\frac{3}{4} log\left(1 - \frac{4}{3}\hat{p}_{AB}\right)$$

The $\hat{d}_{AB}$ values were then used as input into the `nj()` tree-building function implemented in the `APE` package[98] in `R` language.

We measured the distances between all pairs of species in the reconstructed NJ tree (i.e. the lengths of branches) using the `get_distance()` method implemented in the `ETE3` toolkit for phylogenetic trees[99]. Our first measure of 'tree violation' is the difference between these distances and the distances between samples in the original matrix that was used to build the NJ tree.

Multispecies coalescent methods. We applied three different methods that attempt to reconstruct the species tree under the multispecies coalescent model. For a brief discussion of these approaches see Supplementary Methods.

For SNAPP[36] we used a random subset of ~0.5% of genome-wide SNPs (48,922 SNPs) for 12 individuals representing the eco-morphological groups and the Lake Victoria outgroup *P. nyererei* whose alleles were filled in based on the whole genome alignment. The *P. nyererei* alleles were assigned as 'ancestral' (`0` in the `nexus` input file). The 'forward' and 'backward' mutation rate parameters *u* and *v* were calculated directly from the data by SNAPP (the `Calc mutation rates` option). The default value 10 was used for the Coalescent rate parameter and the value of the parameter was sampled (estimated in the MCMC chain). We used uninformative priors as we don't assume strong a priori knowledge about the parameters. The prior for ancestral population sizes was chosen to be a relatively broad gamma distribution with parameters $\alpha = 4$ and $\beta = 20$. The tree height prior λ was set to the initial value of 100 but sampled in the MCMC chain with an uninformative uniform hyperprior on the interval [0,50000]. We ran three independent MCMC chains with the same starting parameters, each on 30 threads with a total runtime of over 10 CPU years. The first one million steps from each MCMC chain was discarded as burn-in. In total, more than 30 million MCMC steps were sampled in the three runs. For the MCMC traces for each run see Supplementary Fig. 24.

Next we used SVDquartets[37,38] as implemented in `PAUP*` [v4.0a (build 159)][100]. We prepared the data into the NEXUS 'dna' format, using `evo` with the `getWGSeq --whole-genome --makeSVDinput -r` options. This command outputs for each individual the DNA base at each variable site, randomly sampling one of the two alleles at heterozygous sites, and ignoring sites that become monomorphic due to this random sampling of alleles. The final dataset contained 17,833,187 SNPs. Then we ran SVDquartets in `PAUP*` setting outgroup to *N. brichardi* and then executing `svdq evalq=all;` specifying that all quartets should be evaluated (not just a random subset). In the final step, `PAUP*` version of the QFM algorithm[101] is used to search for the overall tree that minimizes the number of quartets that are inconsistent with it.

Finally we used `ASTRAL`[40] (`v.5.6.1`) with default parameters and the full set of 2543 local trees generated by RAxML (see above) as input.

Tree comparisons. To summarise the degree of (dis)agreement between the topologies of trees produced by different phylogenetic methods (Fig. 2c), we calculated the normalised Robinson-Foulds distances between pairs of trees[102] using the `RF.dist` function from the `phangorn`[103] package in `R` with the option `normalize=TRUE`.

Chromopainter and fineSTRUCTURE. Singleton SNPs were excluded using the `bcftools v.1.1 -c 2:minor` option, before exporting the remaining variants in the PLINK format[91]. The `chromopainter v0.0.4` software[47] was then run for the 201 largest genomic scaffolds on shapeit phased SNPs. Briefly, we created a uniform recombination map using the `makeuniformrecfile.pl` script, then estimated the effective population size (N_e) for a subsample of 20 individuals using the chromopainter inbuilt expectation-maximization procedure[47], averaged over the 20 N_e values using the provided `neaverage.pl` script. The `chromopainter` program was then run for each scaffold independently, with the `-a 0 0` option to run all individuals against all others. Results for individual scaffolds were combined

using the `chromocombine` tool before running `fineSTRUCTURE v0.0.5` with 1,000,000 burn in iterations, and 200,000 sample iterations, recording a sample every 1,000 iterations (options `-x 1000000 -y 200000 -z 1000`). Finally, the sample relationship tree was built with `fineSTRUCTURE` using the `-m T` option and 20,000 iterations.

The *f*-branch statistic. The f4-admixture ratio (*f* statistic) statistic was developed to estimate the proportion of introgressed material in an admixed population [see SOM18 in ref. 31, and f_G in ref. 48]. However, when calculated for different subsets of samples within the same phylogeny, there are a very large number of highly correlated *f* values that are hard to interpret. To make the interpretation easier, we developed "*f*-branch" or $f_b(C)$: $f_b(C) = median_A\left[min_B[f(A,B,C,O)]\right]$, where B are samples descending from branch *b*, and A are samples descending from the sister branch of *b*. The outgroup *O* was always *N. brichardi.* The $f_b(C)$ score provides for each branch *b* of a given phylogeny and each sample *C* a summary of excess allele sharing of branch *b* with sample *C* (Fig 3, Supplementary Figure 26). Each $f_b(C)$ score was also assigned an associated z-score to assess statistical significance $Z_b(C) = median_A\left[min_B[Z(A,B,C,O)]\right]$. Additional information on the *f* and $f_b(C)$ statistics, including detailed reasoning behind the design of $f_b(C)$, are in Supplementary Methods.

Geometric morphometric analyses. A total of 168 photographs were used to compare the gross body morphology of *Astatotilapia calliptera* to that of endemic Lake Malawi species and other East African *Astatotilapia* lineages (Supplementary Table 7). Coordinates for 17 homologous landmarks [following ref. 104] were collected using `tpsDig2 v2.26`[105]. After landmark digitization, analysis of shape variation was carried out in R (v3.3.2) using the package `GeoMorph v3.0.2`[106]. First a General Procrustes Analysis was applied to remove non-shape variation and shape data were corrected for allometric size effects by performing a regression of Procrustes coordinates (10,000 iterations). The resulting allometry corrected residuals were used in PCA.

Maps. Present day catchment boundary maps are based on 'level 3' detail of Hydro1K dataset from the US Geological Survey. We downloaded the watershed boundary data from the United Nations Environment Programme website (http://ede.grid.unep.ch) and processed it using the QGIS geographic information system software (http://www.qgis.org/en/site/).

Protein-coding gene annotations. We used the `BROADMZ2` annotation generated by the cichlid genome project[11] and removed overlapping transcripts using Jim Kent's `genePredSingleCover` program. Genes whose annotated length in nucleotides was not divisible by three were discarded, as they typically had inaccuracies in annotation that would require manual curation (2495 out of 23698 genes). We also used the cichlid genome project[11] assignment of homologs between the *M. zebra* genome reference and zebrafish (*Danio rerio*).

Coding sequence positive selection scan. We used `evo` with the `getCodingSeq -H b --no-stats` options to obtain the coding sequences for each allele and each gene. The excess of non-synonymous variation (δ_{N-S}) and the non-synonymous variation excess score ($\Delta_{N\text{-}S}$) were calculated on a per-gene basis as follows. Let N_{TS} be the number of possible non-synonymous transitions and N_{TV} the number of possible non-synonymous transversion between two

sequences; analogously S_{TS} and S_{TV} represent possible synonymous differences. We do not specify the ancestral allele, and therefore consider it equally likely that allele i mutated into allele j or that allele j mutated to allele i. Then let N_d be the number of observed non-synonymous mutations and S_d the number of observed synonymous mutations. If there are more than one difference within a codon, all "mutation pathways" (i.e. the different orders in which mutations could have happened) have equal probabilities. When a particular allele contained a premature stop codon, the remainder of the sequence after the stop was excluded from the calculations.

Because the transition:transversion ratio in the Lake Malawi dataset was 1.73, and hence (because there are two possible transversions for each possible transition) the prior probability of each transition is 3.46 times that of each transversion, we account for the unequal probabilities of transitions and transversions in calculating the proportions of non-synonymous (p_N) and of synonymous differences (p_S) as follows:

$$p_N = \frac{N_d}{3.46 \times N_{TS} + N_{TV}} \qquad p_S = \frac{S_d}{3.46 \times S_{TS} + S_{TV}}$$

The excess of non-synonymous variation (δ_{N-S}) is the average of $p_N - p_S$ over pairwise sequence comparisons. Only between-species sequence comparisons are considered for the Lake Malawi dataset. We normalized the δ_{N-S} values in order to take into account the effect on the variance of this statistic introduced by differences in gene length and by sequence composition. To achieve this, we used the leave-one-out jackknife procedure across different pairwise comparisons for each gene, estimating the standard error. The non-synonymous variation excess score ($\Delta_{N\text{-}S}$) is then:

$$\Delta_{N-S} = \frac{\delta_{N-S}}{jackknife\_se(\delta_{N-S})}$$

Note that because the sequences are related by a genealogy, there is a correlation structure between the pairwise comparisons. Therefore, the jackknife approach substantially underestimates the true standard error of δ_{N-S} and is used here simply as a normalization factor.

The null model shown in Fig. 5a was derived by splitting all the coding sequence into its constituent codons, and then randomly sampling these codons with replacement to build new sequences that matched the actual coding genes in their numbers and the length distribution. Then we calculated the Δ_{N-S} scores, as we did for the actual genes and compared the two distributions. High positive values at the upper tail of the distribution are substantially overrepresented in the actual data when compared to a null model.

We also calculated the above statistics for random non-coding regions, matching the gene sequences in length. We used the `bedtools v2.26.0`[107] `shuffle` command to permute the locations of exons along the chromosomes. Of the total length of all the permuted sequences, 98.4% was within the 'accessible genome' and outside of coding sequences (we required at least 95% in any of the permuted locations). The specific command was:

```
bedtools shuffle -chrom -I exons.bed -excl InaccessibleGenome_andExons.bed -f
0.05 -g chrom.sizes
```

Gene Ontology enrichment. Zebrafish has the most extensive functional gene annotation of any fish species, providing a basis for Gene Ontology (GO)[108] term enrichment analysis. Gene Ontology (GO) enrichment for the genes that were candidates for being under positive selection (the top 5% of Δ_{N-S} values) was calculated in `R` using the `topGO v2.26.0` package[109] from

the `Bioconductor` project[110]. The GO hierarchical structure was obtained from the `GO.db v3.4.0` annotation and linking zebrafish gene identifiers to GO terms was accomplished using the `org.Dr.eg.db v3.4.0` annotation package. Genome-wide, between 9024 and 9353 genes had a GO annotation that could be used by `topGO`, the exact number depending on the GO category being assessed. The `nodeSize` parameter was set to 5 to remove GO terms which have fewer than 5 annotated genes, as suggested in the `topGO` manual.

There is often an overlap between gene sets annotated with different GO terms, in part because the terms are related to each other in a hierarchical structure[108]. This is partly accounted for by our use in `topGO` of the `weight` algorithm that accounts for the GO graph structure by down-weighing genes in the GO terms that are neighbours of the locally most significant terms in the GO graph[60]. All the p-values we report are from the `weight` algorithm, which the authors suggest should be reported without multiple testing correction[109].

Some interdependency between significant GO terms remains after using the `weight` algorithm. Therefore, we used the Enrichment Map[111] app for Cytoscape (http://www.cytoscape.org) to organize all the significantly enriched terms into networks where terms are connected if they have a high overlap, i.e. if they share many genes.

***Diplotaxodon* and deep benthic convergence.** To obtain a quantitative measure of the similarity between and the extent of excess diversity in the *Diplotaxodon* and deep benthic amino acid sequences, we calculated simple statistics based on the proportions of non-synonymous differences (p_N scores). Intuitively, the similarity score is high if *Diplotaxodon* and deep benthic jointly have higher p_N then all the others, but are not very different from each other relative to how much diversity there is within *Diplotaxodon* and deep benthic.

Specifically, the similarity score s is calculated as follows:

$$s_{raw} = \bar{p}_N^O - (\bar{p}_N^B - \bar{p}_N^W)$$

and

$$s = \frac{s_{raw}}{jackknife\_se(p_N)} - mean\left(\frac{s_{raw}}{jackknife\_se(p_N)}\right)$$

where $\bar{p}_N^O$ is the mean p_N between *Diplotaxodon* jointly with deep benthic and all the other Lake Malawi species, $\bar{p}_N^B$ is the mean p_N between *Diplotaxodon* and deep benthic, and $\bar{p}_N^W$ is the mean p_N within *Diplotaxodon* and deep benthic. The jackknife normalization is analogous to the one used for $\Delta_{\text{N-S}}$ and the mean ($\bar{s}_{raw}$) is subtracted to center the statistic at zero.

The excess diversity score is high when the mean p_N scores within *Diplotaxodon* and within deep benthic are high relative to the mean p_N in the rest of the radiation. Specifically, the excess score ex is defined as:

$$ex = \frac{[(\bar{p}_N^D + \bar{p}_N^{DB})/2] - \bar{p}_N^R}{jackknife\_se(p_N)}$$

where $\bar{p}_N^D$ is the mean p_N within *Diplotaxodon*, $\bar{p}_N^{DB}$ is the mean p_N within deep benthic, and $\bar{p}_N^R$ is the mean p_N within the rest of the radiation.

Haplotype trees. To view the relationship between haplotypes for genes of interest, we translated nucleotide sequences to amino acid sequences and loaded these into `Haplotype Viewer` (http://www.cibiv.at/~greg/haploviewer). This software requires that a tree is loaded

together with the sequences. Therefore, we inferred gene trees using `RAxML v7.7.8`[94] with the `PROTGAMMADAYHOFFF` model of substitution.

Local excess allele sharing between *Diplotaxodon* and deep benthic. We used an extension of the f_d statistic[48]; this extension is referred to as f_{dM}[55]. f_{dM} is a conservative version of the f statistic that is particularly suited for analysis of small genomic windows[48,55]. For the gene scores shown in Fig. 6b, we calculated f_{dM} (mbuna, deep benthic, *Diplotaxodon*, *N. brichardi*) for each gene in window from the transcription start site (TSS) to 10 kb into the gene. For the 'along the genome' plots, as shown in Fig. 6d and Supplementary Fig. 23, we used a product of two f_{dM} statistics [f_{dM}(shallow benthic, deep benthic, *Diplotaxodon, N. brichardi*) x f_{dM}(*Rhamphochromis*, *Diplotaxodon,* deep benthic, *N. brichardi*)], an approach which we found to increase the local resolution. This score was calculated in sliding windows of 100 SNPs across a region of +-100kb around the genes. Finally, we also calculated f_{dM} (mbuna, deep benthic, *Diplotaxodon*, *N. brichardi*) separately for synonymous and non-synonymous mutations in each gene.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The majority of the custom code used in this project is available on Github as a part of the evo package (https://github.com/millanek/evo). All other custom codes are available from the authors upon request.

Data availability. All raw sequencing reads have been deposited to the NCBI Short Read Archive: (BioProjects PRJEB1254 and PRJEB15289). Sample accessions are listed in Supplementary Table 4. In addition, we are making whole-genome variant calls in the Variant Call Format (VCF), phylogenetic trees and protein coding sequence alignments, and tables with f4 statistics available through the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.7rj8k6c).

References and Notes:

1. Losos, J. B. & Ricklefs, R. E. Adaptation and diversification on islands. *Nature* **457,** 830–836 (2009).
2. Wagner, C. E., Harmon, L. J. & Seehausen, O. Ecological opportunity and sexual selection together predict adaptive radiation. *Nature* **487,** 366–369 (2012).
3. Berner, D. & Salzburger, W. The genomics of organismal diversification illuminated by adaptive radiations. *Trends Genet.* **31,** 491–499 (2015).
4. Darwin, C. *On the Origin of Species*. (OUP Oxford, 2008).
5. Lamichhaney, S. *et al.* Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518,** 371–375 (2015).
6. Losos, J., Jackman, T., Larson, A., Queiroz, K. & Rodriguez-Schettino, L. Contingency and determinism in replicated adaptive radiations of island lizards. *Science* **279,** 2115–2118 (1998).
7. Fryer, G. & Iles, T. D. *The cichlid fishes of the great lakes of Africa: their biology and evolution*. (Oliver and Boyd, 1972).
8. Salzburger, W., Van Bocxlaer, B. & Cohen, A. S. Ecology and Evolution of the African

Great Lakes and Their Faunas. *Annu. Rev. Ecol. Evol. Syst.* **45,** 519–545 (2014).
9. Genner, M. J. *et al.* How does the taxonomic status of allopatric populations influence species richness within African cichlid fish assemblages? *Journal of Biogeography* **31,** 93–102 (2004).
10. Meyer, A. Phylogenetic relationships and evolutionary processes in East African cichlid fishes. *Trends Ecol Evol* **8,** 279–284 (1993).
11. Brawand, D. *et al.* The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513,** 375–381 (2014).
12. Meyer, B. S., Matschiner, M. & Salzburger, W. Disentangling incomplete lineage sorting and introgression to refine species-tree estimates for Lake Tanganyika cichlid fishes. *Syst. Biol.* (2016). doi:10.1093/sysbio/syw069
13. Meier, J. I. *et al.* Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat Commun* **8,** 14363 (2017).
14. Koblmüller, S., Egger, B., Sturmbauer, C. & Sefc, K. M. Rapid radiation, ancient incomplete lineage sorting and ancient hybridization in the endemic Lake Tanganyika cichlid tribe Tropheini. *Mol. Phylogenet. Evol.* **55,** 318–334 (2010).
15. Weiss, J. D., Cotterill, F. P. D. & Schliewen, U. K. Lake Tanganyika—A 'Melting Pot' of Ancient and Young Cichlid Lineages (Teleostei: Cichlidae)? *PLoS ONE* **10,** e0125043 (2015).
16. Gante, H. F. *et al.* Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. *Mol Ecol* (2016). doi:10.1111/mec.13767
17. Wagner, C. E. *et al.* Genome□wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol Ecol* **22,** 787–798 (2012).
18. Genner, M. J. & Turner, G. F. Ancient hybridization and phenotypic novelty within Lake Malawi's cichlid fish radiation. *Mol. Biol. Evol.* **29,** 195–206 (2012).
19. Moran, P., Kornfield, I. & Reinthal, P. N. Molecular Systematics and Radiation of the Haplochromine Cichlids (Teleostei: Perciformes) of Lake Malawi. *Copeia* **1994,** 274 (1994).
20. Joyce, D. A. *et al.* Repeated colonization and hybridization in Lake Malawi cichlids. **21,** R108–9 (2011).
21. Leffler, E. M. *et al.* Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* **10,** e1001388 (2012).
22. Albertson, R. C., Markert, J. A., Danley, P. D. & Kocher, T. D. 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).
23. Kocher, T. D. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.* **5,** 288–298 (2004).
24. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
25. Ravi, V. & Venkatesh, B. Rapidly evolving fish genomes and teleost diversity. *Curr. Opin. Genet. Dev.* **18,** 544–550 (2008).
26. Recknagel, H., Elmer, K. R. & Meyer, A. 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).

27. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15,** 47–70 (2014).
28. Feng, C. *et al.* Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. *Elife* **6,** e23907 (2017).
29. Konings, A. *Malawi Cichlids in Their Natural Habitat*. (Cichlid Press, 2007).
30. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
31. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
32. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28,** 2239–2252 (2011).
33. Heled, J. & Bouckaert, R. R. Looking for trees in the forest: summary tree from posterior samples. *BMC Evol. Biol.* **13,** 221 (2013).
34. Edwards, S. V. Is a new and general theory of molecular systematics emerging? *Evolution* **63,** 1–19 (2009).
35. Edwards, S. V. *et al.* Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* **94,** 447–462 (2016).
36. Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A. & RoyChoudhury, A. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* **29,** 1917–1932 (2012).
37. Chifman, J. & Kubatko, L. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics* **30,** 3317–3324 (2014).
38. Chifman, J. & Kubatko, L. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *J. Theor. Biol.* **374,** 35–47 (2015).
39. Long, C. & Kubatko, L. The Effect of Gene Flow on Coalescent-Based Species-Tree Inference. *arXiv.org* 1–19 (2017).
40. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30,** i541–8 (2014).
41. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19,** 153 (2018).
42. Dasarathy, G., Nowak, R. & Roch, S. Data Requirement for Phylogenetic Inference from Multiple Loci: A New Distance Method. *IEEE/ACM Trans Comput Biol Bioinform* **12,** 422–432 (2015).
43. Rusinko, J. & McPartlon, M. Species tree estimation using Neighbor Joining. *J. Theor. Biol.* **414,** 5–7 (2017).
44. Ballard, J. W. O. & Whitlock, M. C. The incomplete natural history of mitochondria. *Mol Ecol* **13,** 729–744 (2004).
45. Toews, D. P. L. & Brelsford, A. The biogeography of mitochondrial and nuclear discordance in animals. *Mol Ecol* **21,** 3907–3930 (2012).
46. Consuegra, S., John, E., Verspoor, E. & de Leaniz, C. G. Patterns of natural selection acting on the mitochondrial genome of a locally adapted fish species. *Genet. Sel. Evol.* **47,** 58 (2015).
47. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLoS Genet.* **8,** e1002453 (2012).

48. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32,** 244–257 (2015).
49. Martin, S. H. *et al.* Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* **23,** 1817–1828 (2013).
50. Eccles, D. H. & Trewavas, E. *Malawian Cichlid Fishes*. (Lake Fish Movies, 1989).
51. Eriksson, A. & Manica, A. 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).
52. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8,** e1002967 (2012).
53. Peterson, E. N., Cline, M. E., Moore, E. C., Roberts, N. B. & Roberts, R. B. Genetic sex determination in Astatotilapia calliptera, a prototype species for the Lake Malawi cichlid radiation. *Naturwissenschaften* **104,** 41 (2017).
54. Genner, M. J., Ngatunga, B. P., Mzighani, S., Smith, A. & Turner, G. F. Geographical ancestry of Lake Malawi's cichlid fish diversity. *Biol. Lett.* **11,** 20150232 (2015).
55. Malinsky, M. *et al.* Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* **350,** 1493–1498 (2015).
56. Ivory, S. J. *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).
57. Lyons, R. P. *et al.* Continuous 1.3-million-year record of East African hydroclimate, and implications for patterns of evolution and biodiversity. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 15568–15573 (2015).
58. Greenwood, P. H. *Towards a phyletic classification of the 'genus' Haplochromis (Pisces, Cichlidae) and related taxa. Part 1.* **35,** 265–322 (1979).
59. Lippitsch, E. 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).
60. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22,** 1600–1607 (2006).
61. Van Bocxlaer, B., SCHULTHEIß, R., Plisnier, P.-D. & Albrecht, C. Does the decline of gastropods in deep water herald ecosystem change in Lakes Malawi and Tanganyika? *Freshwater Biology* **57,** 1733–1744 (2012).
62. Spady, T. C. *et al.* Evolution of the cichlid visual palette through ontogenetic subfunctionalization of the opsin gene arrays. *Mol. Biol. Evol.* **23,** 1538–1547 (2006).
63. Weadick, C. J. & Chang, B. S. W. Complex patterns of divergence among green-sensitive (RH2a) African cichlid opsins revealed by Clade model analyses. *BMC Evol. Biol.* **12,** 206 (2012).
64. Sugawara, T. *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).
65. Bowmaker, J. K. & Hunt, D. M. Evolution of vertebrate visual pigments. *Current Biology* **16,** R484–R489 (2006).
66. Davies, W. I. L., Collin, S. P. & Hunt, D. M. Molecular ecology and adaptation of visual photopigments in craniates. *Mol Ecol* **21,** 3121–3158 (2012).

67. Carleton, K. L., Dalton, B. E., Escobar-Camacho, D. & Nandamuri, S. P. Proximate and ultimate causes of variable visual sensitivities: Insights from cichlid fish radiations. *Genesis* **54,** 299–325 (2016).
68. Ogawa, Y., Shiraki, T., Kojima, D. & Fukada, Y. Homeobox transcription factor Six7 governs expression of green opsin genes in zebrafish. *Proceedings of the Royal Society B: Biological Sciences* **282,** 20150659 (2015).
69. Renninger, S. L., Gesemann, M. & Neuhauss, S. C. F. Cone arrestin confers cone vision of high temporal resolution in zebrafish larvae. *Eur. J. Neurosci.* **33,** 658–667 (2011).
70. Brockerhoff, S. E. *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).
71. Boesze-Battaglia, K. & Goldberg, A. F. X. Photoreceptor renewal: a role for peripherin/rds. *Int. Rev. Cytol.* **217,** 183–225 (2002).
72. Opazo, J. C., Butts, G. T., Nery, M. F., Storz, J. F. & Hoffmann, F. G. Whole-genome duplication and the functional diversification of teleost fish hemoglobins. *Mol. Biol. Evol.* **30,** 140–153 (2013).
73. Hahn, C., Genner, M. J., Turner, G. F. & Joyce, D. A. The genomic basis of cichlid fish adaptation within the deepwater 'twilight zone' of Lake Malawi. *Evolution Letters* **1,** 184–198 (2017).
74. Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487,** 94–98 (2012).
75. Meyer, A., Kocher, T. D., Basasibwaki, P. & Wilson, A. C. Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. *Nature* **347,** 550–553 (1990).
76. Reid, N. M. *et al.* The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science* **354,** 1305–1308 (2016).
77. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484,** 55–61 (2012).
78. Seehausen, O. Cichlid Fish Diversity Threatened by Eutrophication That Curbs Sexual Selection. *Science* **277,** 1808–1811 (1997).
79. Konings, A. *Tanganyika cichlids in their natural habitat*. (Cichlid Press, 2015).
80. Coyne, J. A. & Orr, H. A. Speciation. (2004).
81. Feder, J. L., Egan, S. P. & Nosil, P. The genomics of speciation-with-gene-flow. *Trends in Genetics* **28,** 342–350 (2012).
82. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org* **q-bio.GN,** (2013).
83. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
84. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).
85. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).
86. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9,** 179–181 (2012).

87. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93,** 687–696 (2013).
88. Miller, W. *et al.* 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.* **17,** 1797–1808 (2007).
89. Ulm, K. A simple method to calculate the confidence interval of a standardized mortality ratio (SMR). *Am. J. Epidemiol.* **131,** 373–375 (1990).
90. Dobson, A. J., Kuulasmaa, K., Eberle, E. & Scherer, J. Confidence intervals for weighted sums of Poisson parameters. *Stat Med* **10,** 457–462 (1991).
91. Purcell, S. *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).
92. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190–e190 (2006).
93. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23,** 1514–1521 (2013).
94. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
95. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10,** e1003537 (2014).
96. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57,** 758–771 (2008).
97. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).
98. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20,** 289–290 (2004).
99. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33,** 1635–1638 (2016).
100. Swofford, D. L. *PAUP*: Phylogenetic analysis using parsimony (and other methods), 2002*. (Sinauer Associates).
101. Reaz, R., Bayzid, M. S. & Rahman, M. S. Accurate Phylogenetic Tree Reconstruction from Quartets: A Heuristic Approach. *PLoS ONE* **9,** e104008
102. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Mathematical Biosciences* (1981).
103. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27,** 592–593 (2011).
104. Theis, A., Ronco, F., Indermaur, A., Salzburger, W. & Egger, B. Adaptive divergence between lake and stream populations of an East African cichlid fish. *Mol Ecol* **23,** 5304–5322 (2014).
105. Rohlf, F. J. tpsDig2.
106. Adams, D. C. & Castillo, E. O. geomorph: an R package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and …* (2013).
107. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
108. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).
109. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. *R package version* (2010).

110. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12,** 115–121 (2015).
111. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5,** e13984 (2010).

Acknowledgments
This work was supported by the Wellcome Trust (097677/Z/11/Z to MM, WT206194 and WT207492 for 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). We want to thank the Sanger Institute sequencing core for DNA sequencing, Mingliu Du for DNA extractions, David Swofford and Michael Matschiner for advice on phylogenomic analyses, Walter Salzburger and Ian Wilson for comments on the manuscript. We also thank the Tanzania Fisheries Research Institute, and the Fisheries Research Unit of the Government of Malawi for their assistance and support.

Author contributions:
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 analysed the genomic data. All authors participated in interpretation of the results. MM, HS and RD drafted the manuscript, and all others commented.

Competing interests
RD declares that he owns stock in Illumina from previous consulting. The authors declare no other competing interests.

Figure captions:

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. 29. **b,** The distributions of genomic sequence diversity within individuals (heterozygosity; π) and of divergence between species (d_{XY}). **c,** Principal component analysis (PCA) of whole genome variation data.

Fig. 2: Excess allele sharing and patterns of species relatedness. a, Derived allele sharing reveals non-tree-like relationships among trios of species. The bars show the proportion of significantly elevated D_{min} scores (see main text). Shading corresponds to FWER q values of (from light to dark) 10^{-2}, 10^{-4}, 10^{-8}, 10^{-14}. The scatterplots show the D_{min} scores that were significant at FWER<0.01. Results are shown separately for comparisons where all three species in the trio are from the same group, and for cases where the species come from two or three different groups. *Rhamphochromis* and utaka within-group comparisons are not shown due to the low number of data points. **b,** 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 and then subsampled for display to 12 individuals representing the eco-morphological groups. The maximum clade credibility tree shown here was built from the subsampled local trees. A ML mitochondrial phylogeny is shown for comparison. **c,** A summary of all phylogenies from this study and the normalised Robinson-Foulds distances between them, reflecting the topological distance between pairs of trees on the scale from zero to 100%. The least controversial 12 sample tree is SNAPP.t1, with an average distance to other trees of 17.7%, while ASTRAL* is the least controversial among the 'main trees' (mean distance of 25.3%). To compare trees with differing sets of taxa, the trees were downsampled so that only matching taxa were present. The position of the outgroup/root was considered in all comparisons.

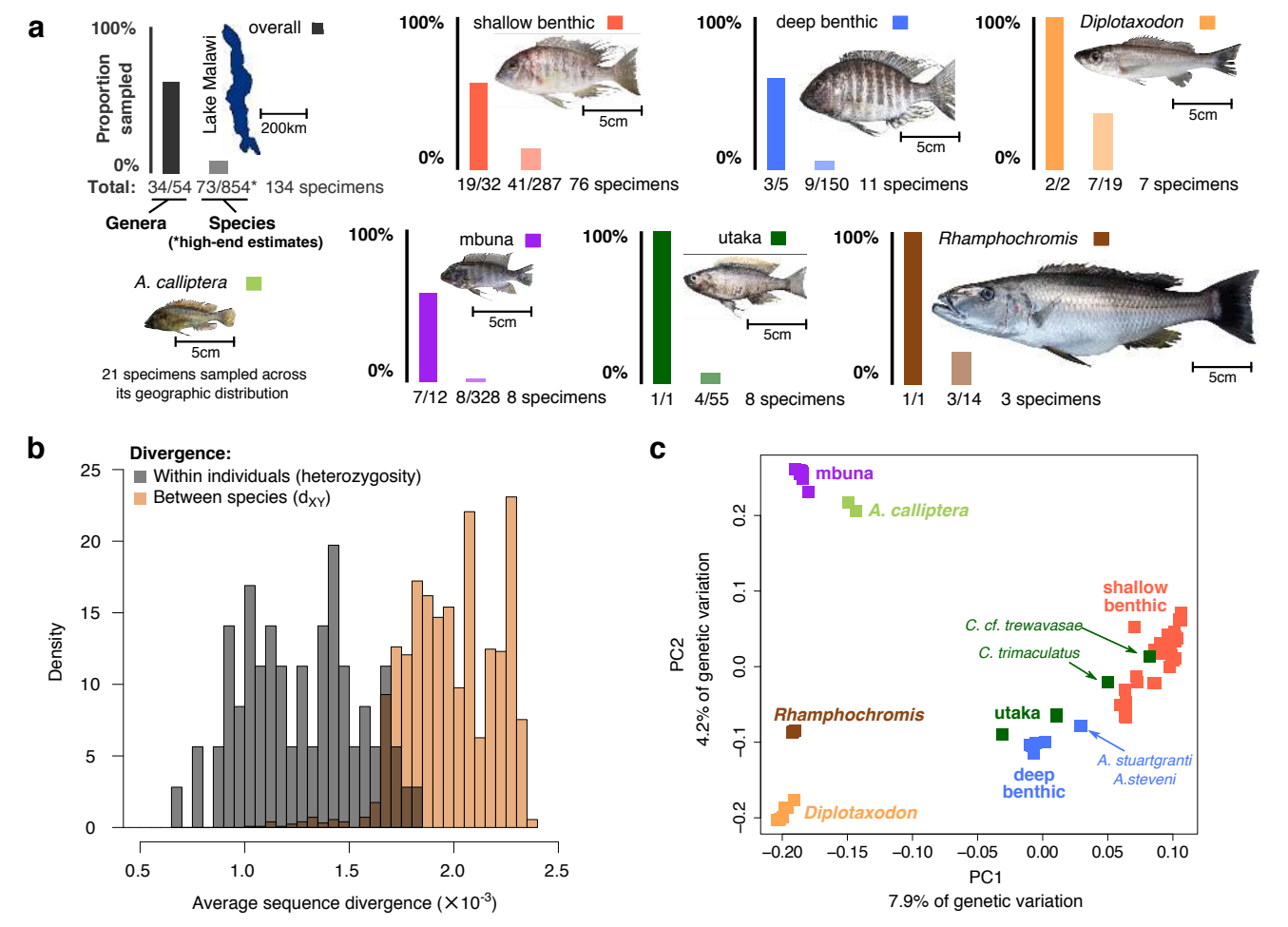
Fig. 3: Identifying tree violating branches and possible gene flow events. The branch-specific statistic $f_b(C)$ identifies excess sharing of derived alleles between the branch of the tree on the y-axis and the species C on the x-axis (see Supplementary Note). The ASTRAL* tree was used as a basis for the branch statistic and grey data points in the matrix correspond to tests that are not consistent with the phylogeny. Colours correspond to eco-morphological groups as in Fig. 1. The * sign denotes block jack-knifing significance at $|Z|>3.17$ (Holm-Bonferroni FWER<0.001).

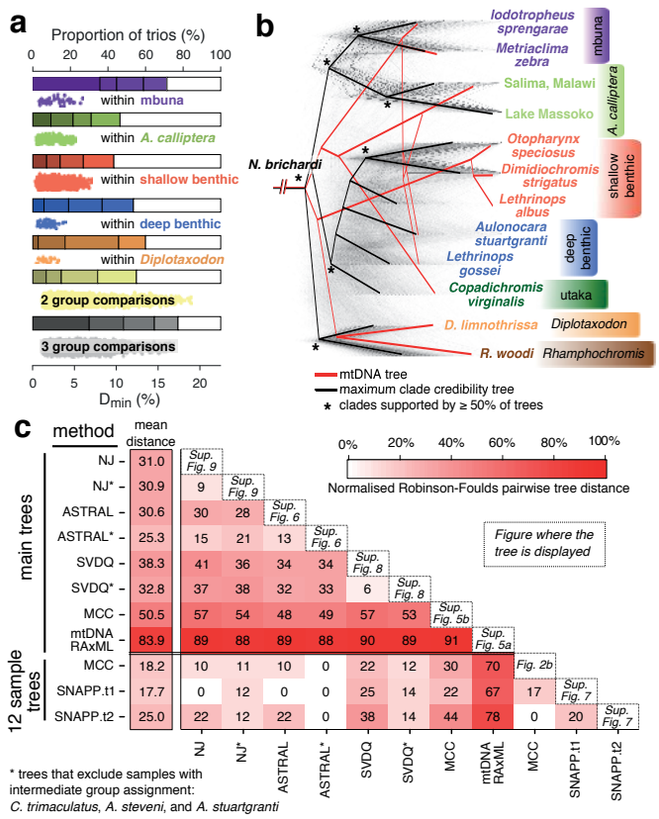
Fig. 4: Origins of the radiation and the role of *A. calliptera*. a, An NJ phylogeny showing the Lake Malawi radiation in the context of other East African *Astatotilapia* taxa. **b,** A Lake Malawi NJ phylogeny with expanded view of *A. calliptera*, with all other groups collapsed. **c,** Approximate *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,** Strong f_4 admixture ratio signal showing that Malawi catchment

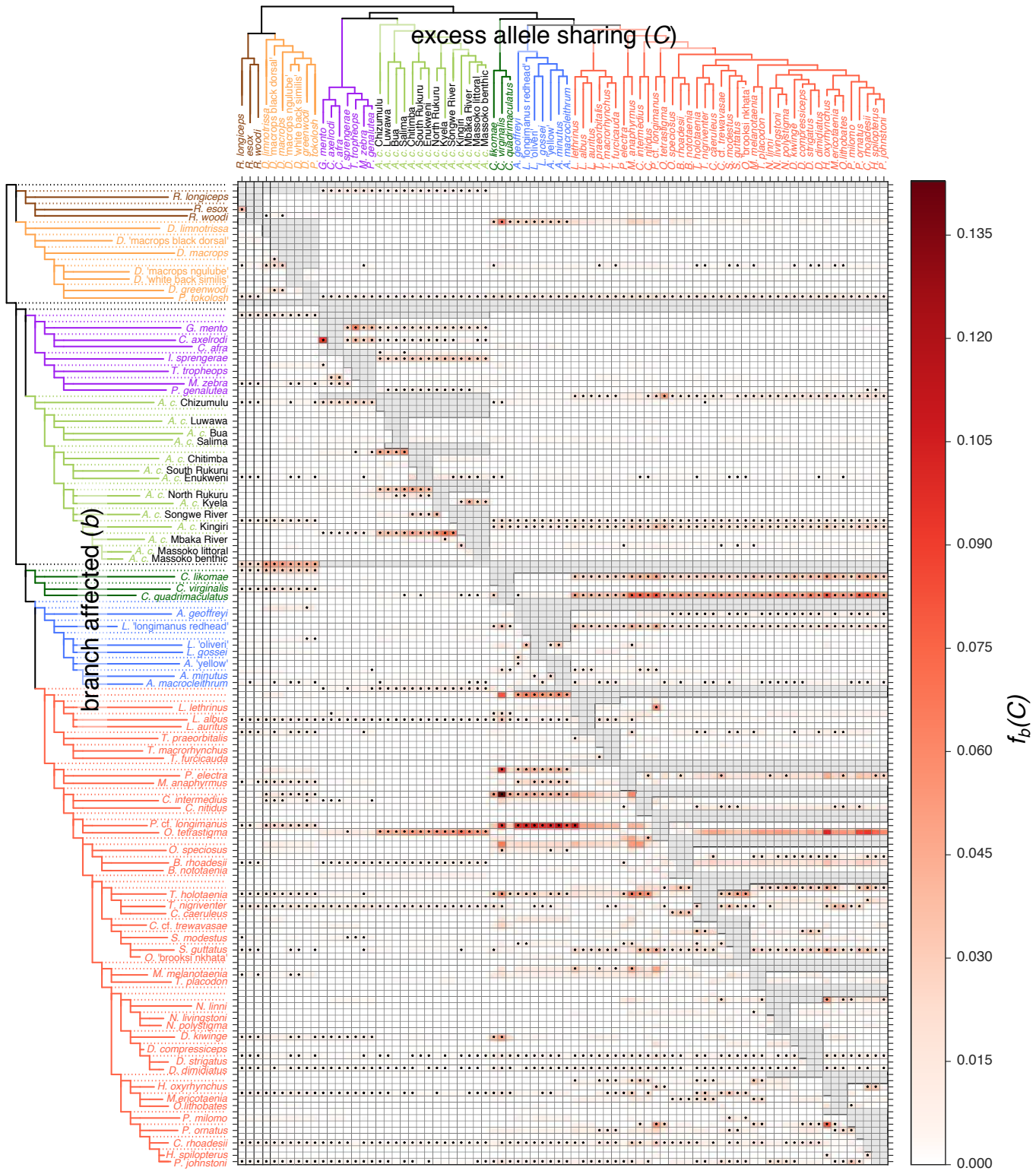
A. calliptera are closer to mbuna than their Indian Ocean catchment counterparts. **e,** PCA of body shape variation of Lake Malawi endemics, *A. calliptera* and other *Astatotilapia* taxa, obtained from geometric morphometric analysis. **f,** A phylogeny with the same topology as in panel (b) but displayed 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, and translation of this value into a mean divergence time estimate with 95% CI reflecting the statistical uncertainty in mutation rate. Dashed lines with arrows indicate likely instances of gene flow between major groups; their true timings are uncertain.

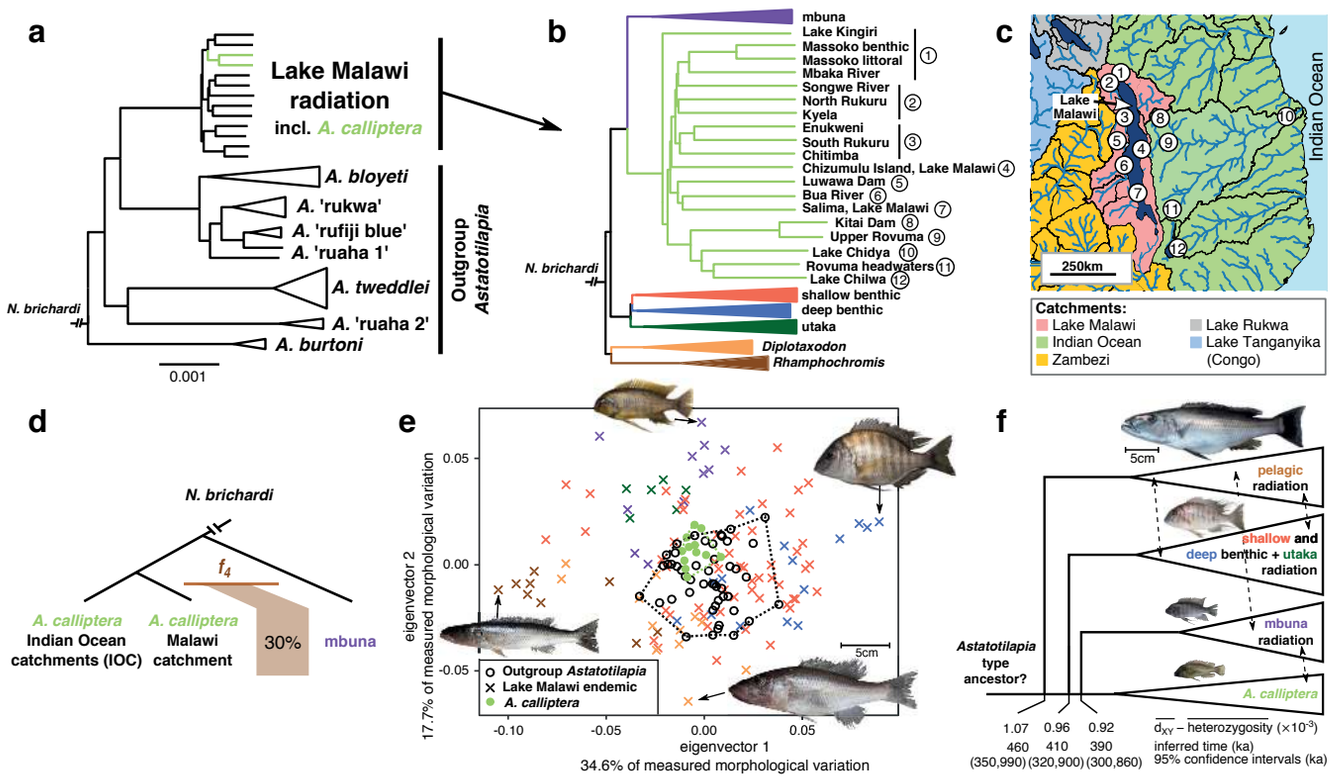
Fig. 5: Gene selection scores, copy numbers, and ontology enrichment. **a,** The distribution of the non-synonymous variation excess scores ($\boldsymbol{\Delta_{N-S}}$) highlighting the top 5% cutoff, compared against a null model. The null was derived by calculating the statistic on randomly sampled combinations of codons. We also show the distributions of genes in selected Gene Ontology (GO) categories which are overrepresented in the top 5%. **b,** The relationship between the probability of $\boldsymbol{\Delta_{N-S}}$ being 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 $\boldsymbol{\chi^2}$ tests of independence. Genes existing in two or more copies in both zebrafish and Malawi cichlids are disproportionately represented among candidate selected genes. **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 colour 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.

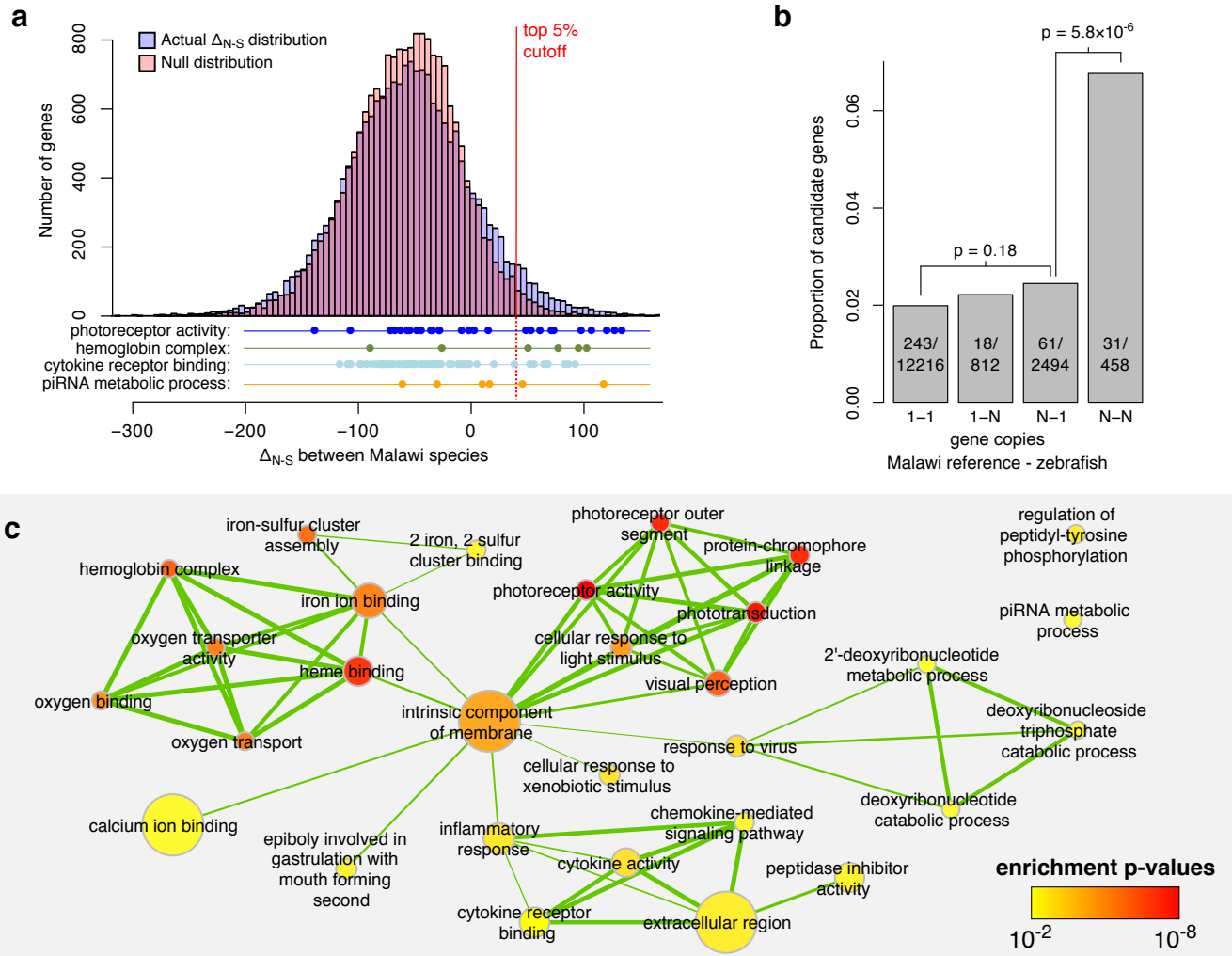
Fig. 6: Shared selection between the deep water adapted groups *Diplotaxodon* and deep benthic. **a,** The scatterplot shows the distribution of genes with high Δ_{N-S} scores (candidates for positive selection) along axes reflecting shared selection signatures. Only genes with zebrafish homologs are shown. Amino acid haplotype trees, shown for genes as indicated by the red symbols and numbers, indicate that *Diplotaxodon* and deep benthic species are often divergent from other taxa, but similar to each other. 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 local excess allele sharing between deep benthic and *Diplotaxodon*. Overall there is no correlation between Δ_{N-S} and f_{dM}. However, the strong correlation between Δ_{N-S} and f_{dM} in the highlighted GO categories suggests that positively selected alleles in those categories tend to be subject to introgression or convergent selection between *Diplotaxodon* and the deep benthic group. **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.

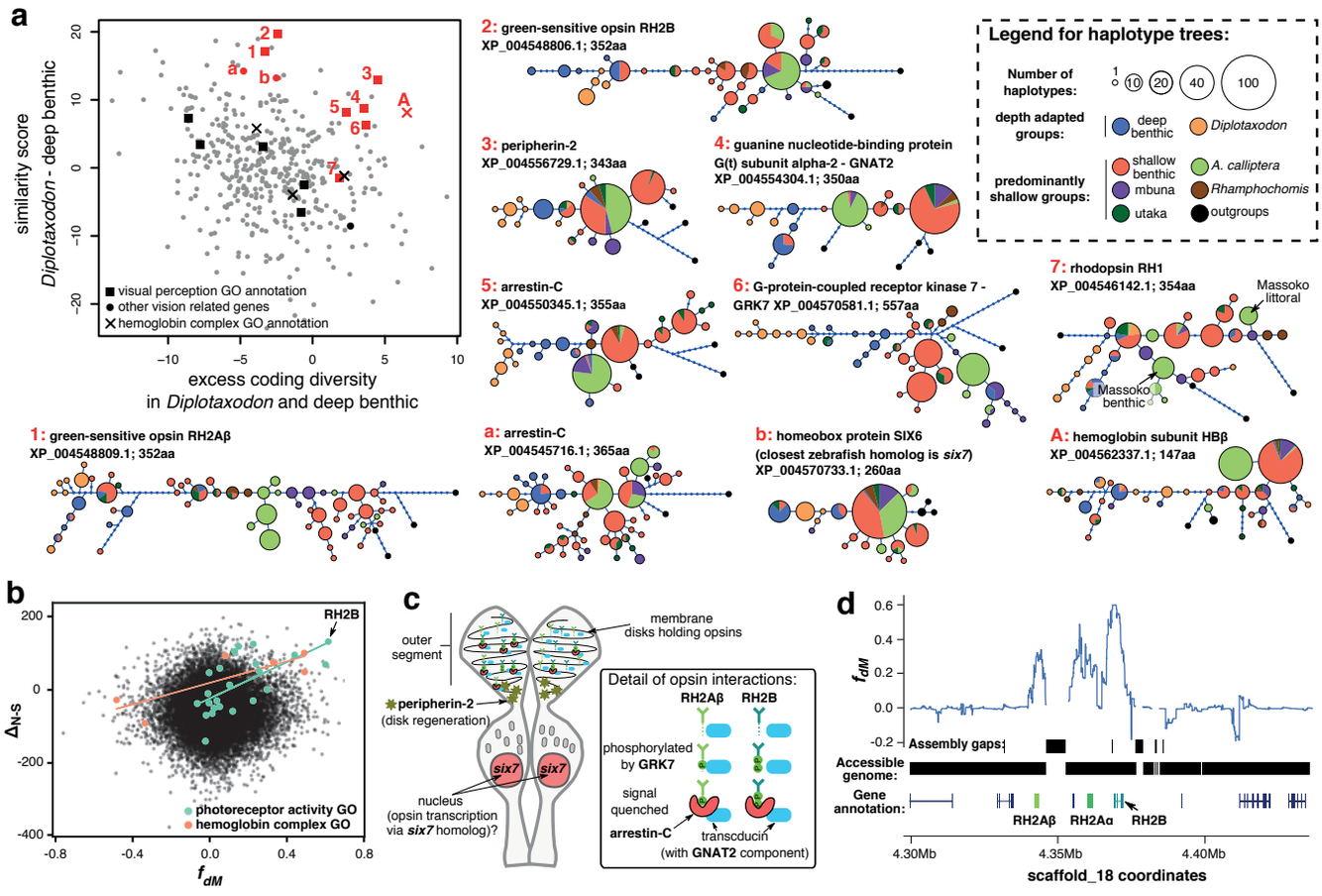

a
100%
0%
Proportion sampled
overall
Lake Malawi
200km
Total: 34/54 73/854* 134 specimens
Genera
Species
(*high-end estimates)
A. calliptera
5cm
21 specimens sampled across its geographic distribution
shallow benthic
19/32 41/287 76 specimens
deep benthic
3/5 9/150 11 specimens
Diplotaxodon
2/2 7/19 7 specimens
mbuna
7/12 8/328 8 specimens
utaka
1/1 4/55 8 specimens
Rhamphochromis
1/1 3/14 3 specimens
b
Divergence:
Within individuals (heterozygosity)
Between species (d_{XY})
Density
Average sequence divergence ($\times 10^{-3}$)
c
mbuna
A. calliptera
shallow benthic
C. cf. trewavasae
C. trimaculatus
Rhamphochromis
utaka
A. stuartgranti
A.steveni
deep benthic
Diplotaxodon
PC2
4.2% of genetic variation
PC1
7.9% of genetic variation


a
Proportion of trios (%)
0 20 40 60 80 100
within mbuna
within A. calliptera
within shallow benthic
within deep benthic
within Diplotaxodon
2 group comparisons
3 group comparisons
0 5 10 15 20
Dmin (%)
b
Iodotropheus sprengarae
Metriaclima zebra
mbuna
Salima, Malawi
Lake Massoko
A. calliptera
Otopharynx speciosus
Dimidiochromis strigatus
Lethrinops albus
shallow benthic
N. brichardi
Aulonocara stuartgranti
Lethrinops gossei
deep benthic
Copadichromis virginalis
utaka
D. limnothrissa
Diplotaxodon
R. woodi
Rhamphochromis
mtDNA tree
maximum clade credibility tree
clades supported by ≥ 50% of trees
c
method
mean distance
0% 20% 40% 60% 80% 100%
Normalised Robinson-Foulds pairwise tree distance
Figure where the tree is displayed
main trees
12 sample trees
NJ 31.0 Sup. Fig. 9
NJ* 30.9 9 Sup. Fig. 9
ASTRAL 30.6 30 28 Sup. Fig. 6
ASTRAL* 25.3 15 21 13 Sup. Fig. 6
SVDQ 38.3 41 36 34 34 Sup. Fig. 8
SVDQ* 32.8 37 38 32 33 6 Sup. Fig. 8
MCC 50.5 57 54 48 49 57 53 Sup. Fig. 5b
mtDNA RAxML 83.9 89 88 89 88 90 89 91 Sup. Fig. 5a
MCC 18.2 10 11 10 0 22 12 30 70 Fig. 2b
SNAPP.t1 17.7 0 12 0 0 25 14 22 67 17 Sup. Fig. 7
SNAPP.t2 25.0 22 12 22 0 38 14 44 78 0 20 Sup. Fig. 7
NJ NJ* ASTRAL ASTRAL* SVDQ SVDQ* MCC mtDNA RAxML MCC SNAPP.t1 SNAPP.t2
* trees that exclude samples with intermediate group assignment: C. trimaculatus, A. steveni, and A. stuartgranti


excess allele sharing (C)
branch affected (b)
$f_b(C)$
0.135
0.120
0.105
0.090
0.075
0.060
0.045
0.030
0.015
0.000
R. longiceps
R. esox
R. woodi
D. limnotrissa
D. 'macrops black dorsal'
D. macrops
D. 'macrops ngulube'
D. 'white back similis'
D. greenwodi
P. tokolosh
G. mento
C. axelrodi
C. afra
I. sprengerae
T. tropheops
M. zebra
P. genalutea
A. c. Chizumulu
A. c. Luwawa
A. c. Bua
A. c. Salima
A. c. Chitimba
A. c. South Rukuru
A. c. Enukweni
A. c. North Rukuru
A. c. Kyela
A. c. Songwe River
A. c. Kingiri
A. c. Mbaka River
A. c. Massoko littoral
A. c. Massoko benthic
C. likomae
C. virginalis
C. quadrimaculatus
A. geoffreyi
L. 'longimanus redhead'
L. 'oliveri'
L. gossei
A. 'yellow'
A. minutus
A. macroclethrum
L. lethrinus
L. albus
L. auritus
T. praeorbitalis
T. macrorhynchus
T. furcicauda
P. electra
M. anaphyrmus
C. intermedius
C. nitidus
P. cf. longimanus
O. tetrastigma
O. speciosus
B. rhoadesii
B. notataenia
T. holotaenia
T. nigriventer
C. caeruleus
C. cf. frewavasae
S. modestus
S. guttatus
O. 'brooksi nkhata'
M. melanotaenia
T. placodon
N. linni
N. livingstoni
N. polystigma
D. kiwinge
D. compressiceps
D. strigatus
D. dimidiatus
H. oxyrhynchus
M. ericotaenia
O. lithobates
P. milomo
P. ornatus
C. rhoadesii
H. spilopterus
P. johnstoni


a
Lake Malawi radiation
incl. A. calliptera
A. bloyeti
A. 'rukwa'
A. 'rufiji blue'
A. 'ruaha 1'
A. tweddlei
A. 'ruaha 2'
A. burtoni
Outgroup Astatotilapia
N. brichardi
0.001
b
mbuna
Lake Kingiri
Massoko benthic
Massoko littoral
Mbaka River
Songwe River
North Rukuru
Kyela
Enukweni
South Rukuru
Chitimba
Chizumulu Island, Lake Malawi
Luwawa Dam
Bua River
Salima, Lake Malawi
Kitai Dam
Upper Rovuma
Lake Chidya
Rovuma headwaters
Lake Chilwa
shallow benthic
deep benthic
utaka
Diplotaxodon
Rhamphochromis
N. brichardi
c
Lake Malawi
Indian Ocean
250km
Catchments:
Lake Malawi
Indian Ocean
Zambezi
Lake Rukwa
Lake Tanganyika
(Congo)
d
N. brichardi
f4
A. calliptera
Indian Ocean
catchments (IOC)
A. calliptera
Malawi
catchment
30%
mbuna
e
17.7% of measured morphological variation
eigenvector 2
0.05
0.00
-0.05
-0.10
-0.05
0.00
0.05
eigenvector 1
34.6% of measured morphological variation
Outgroup Astatotilapia
Lake Malawi endemic
A. calliptera
5cm
f
5cm
pelagic radiation
shallow and deep benthic + utaka radiation
mbuna radiation
A. calliptera
Astatotilapia type ancestor?
1.07
460
(350,990)
0.96
410
(320,900)
0.92
390
(300,860)
dXY – heterozygosity (×10-3)
inferred time (ka)
95% confidence intervals (ka)


a
Actual Δ N-S distribution
Null distribution
top 5% cutoff
Number of genes
0
200
400
600
800
photoreceptor activity:
hemoglobin complex:
cytokine receptor binding:
piRNA metabolic process:
−300
−200
−100
0
100
Δ N-S between Malawi species
b
p = 5.8×10-6
p = 0.18
Proportion of candidate genes
0.00
0.02
0.04
0.06
243/ 12216
18/ 812
61/ 2494
31/ 458
1–1
1–N
N–1
N–N
gene copies
Malawi reference - zebrafish
c
iron-sulfur cluster assembly
2 iron, 2 sulfur cluster binding
photoreceptor outer segment
protein-chromophore linkage
regulation of peptidyl-tyrosine phosphorylation
hemoglobin complex
iron ion binding
photoreceptor activity
phototransduction
piRNA metabolic process
oxygen transporter activity
heme binding
cellular response to light stimulus
visual perception
2'-deoxyribonucleotide metabolic process
oxygen binding
intrinsic component of membrane
deoxyribonucleoside triphosphate catabolic process
oxygen transport
response to virus
cellular response to xenobiotic stimulus
deoxyribonucleotide catabolic process
calcium ion binding
chemokine-mediated signaling pathway
epiboly involved in gastrulation with mouth forming second
inflammatory response
cytokine activity
peptidase inhibitor activity
cytokine receptor binding
extracellular region
enrichment p-values
10-2
10-8


a
similarity score
Diplotaxodon - deep benthic
excess coding diversity
in Diplotaxodon and deep benthic
visual perception GO annotation
other vision related genes
hemoglobin complex GO annotation
2: green-sensitive opsin RH2B
XP_004548806.1; 352aa
3: peripherin-2
XP_004556729.1; 343aa
4: guanine nucleotide-binding protein
G(t) subunit alpha-2 - GNAT2
XP_004554304.1; 350aa
5: arrestin-C
XP_004550345.1; 355aa
6: G-protein-coupled receptor kinase 7 -
GRK7 XP_004570581.1; 557aa
7: rhodopsin RH1
XP_004546142.1; 354aa
Massoko littoral
Massoko benthic
Legend for haplotype trees:
Number of haplotypes:
depth adapted groups:
deep benthic
Diplotaxodon
predominantly shallow groups:
shallow benthic
mbuna
utaka
A. calliptera
Rhamphochomis
outgroups
1: green-sensitive opsin RH2Aβ
XP_004548809.1; 352aa
a: arrestin-C
XP_004545716.1; 365aa
b: homeobox protein SIX6
(closest zebrafish homolog is six7)
XP_004570733.1; 260aa
A: hemoglobin subunit HBβ
XP_004562337.1; 147aa
b
RH2B
photoreceptor activity GO
hemoglobin complex GO
c
outer segment
membrane disks holding opsins
peripherin-2
(disk regeneration)
six7
six7
nucleus
(opsin transcription
via six7 homolog)?
Detail of opsin interactions:
RH2Aβ
RH2B
phosphorylated by GRK7
signal quenched
arrestin-C
transducin
(with GNAT2 component)
d
Assembly gaps:
Accessible genome:
Gene annotation:
RH2Aβ
RH2Aα
RH2B
4.30Mb
4.35Mb
4.40Mb
scaffold_18 coordinates