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

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

Affiliations:

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

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

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

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

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

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

†These authors contributed equally to this work.

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

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

Main Text: The formation of every lake or island represents a fresh opportunity for colonization, proliferation and diversification of living forms. In some cases, the ecological opportunities presented by underutilized habitats facilitate adaptive radiation - rapid and extensive diversification of the descendants of the colonizing lineages[1-3]. Adaptive radiations are thus exquisite examples of the power of natural selection, as seen for example in Darwin's Galapagos finches[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 and of identifying sister species remains challenging due both to retention of large amounts of ancestral genetic polymorphism (i.e. incomplete lineage sorting) and to evidence suggesting gene flow between taxa[12,14-19].

Initial genome assemblies of cichlids from East Africa suggest that an increased rate of gene duplications together with accelerated evolution of some regulatory elements and protein coding genes may have contributed to the radiations[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.

Previous work on the phylogeny of Lake Malawi haplochromine cichlid radiation, mainly based on mitochondrial DNA (mtDNA), has divided the species into six groups with differing ecology and morphology[20]: 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 benthic substrate; 6) a diverse group of benthic species, mainly found in shallow non-rocky habitats. In addition, *Astatotilapia calliptera* is a closely related generalist that inhabits shallow weedy margins of Lake Malawi, and other lakes and rivers in the catchment and more widely. To characterize the genetic diversity, species relationships, and signatures of selection across the whole radiation, we obtained paired-end Illumina whole-genome sequence data from 134 individuals of 73 species distributed broadly across the seven groups (Fig. 1a)[21]. This includes 102 individuals at ~15× coverage and 32 additional individuals at ~6× (Table S1).

Low genetic diversity and species divergence

Sequence data were aligned to the *Metriaclima zebra* reference assembly version 1.1[11], followed by variant calling restricted to the 'accessible genome' (the portion of the genome where variants can be determined confidently with short read alignment), which comprised 653Mb or 91.5% of the assembly excluding gaps[21]. Average divergence from the reference per sample was 0.19% to 0.27% (Fig. S1). Across all samples, after filtering and variant refinement, we obtained 30.6 million variants of which 27.1 million were single nucleotide polymorphisms (SNPs) and the rest were short insertions and deletions (indels)[21]. Unless otherwise indicated, the following analyses are based on SNPs.

To estimate nucleotide diversity (π) within the sampled 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[22], and there appears to be little relationship between π and the rate of speciation: individuals in the species-rich mbuna and shallow benthic groups show levels of π comparable to the relatively species-poor utaka, *Diplotaxodon*, and *Rhamphochromis* (Fig. S1).

Despite their extensive phenotypic differentiation, all species within the Lake Malawi haplochromine cichlid radiation are genetically closely related[23,24]. However, genome-wide genetic divergence has never been quantified. To examine the extent of genetic differentiation between species across the radiation we calculated the average pairwise differences between all pairs of sequences from different species (d_{XY}). A comparison of d_{XY} against heterozygosity reveals that the two distributions partially overlap (Fig. 1b). Thus, the sequence divergence within a single diploid individual is sometimes higher than the divergence between sequences found in two distinct species. The average d_{XY} is 2.0×10^{-3} with a range between 1.0 and 2.4×10^{-3} per bp. The maximum d_{XY} is approximately one fifth of the divergence between human and chimpanzee[25]. The low ratio of divergence to diversity means that most genetic variation is shared between species. On average both alleles are observed in other species for 82% of heterozygous sites within individuals, consistent with the expected and previously observed high levels of incomplete lineage sorting (ILS) between Lake Malawi species[24].

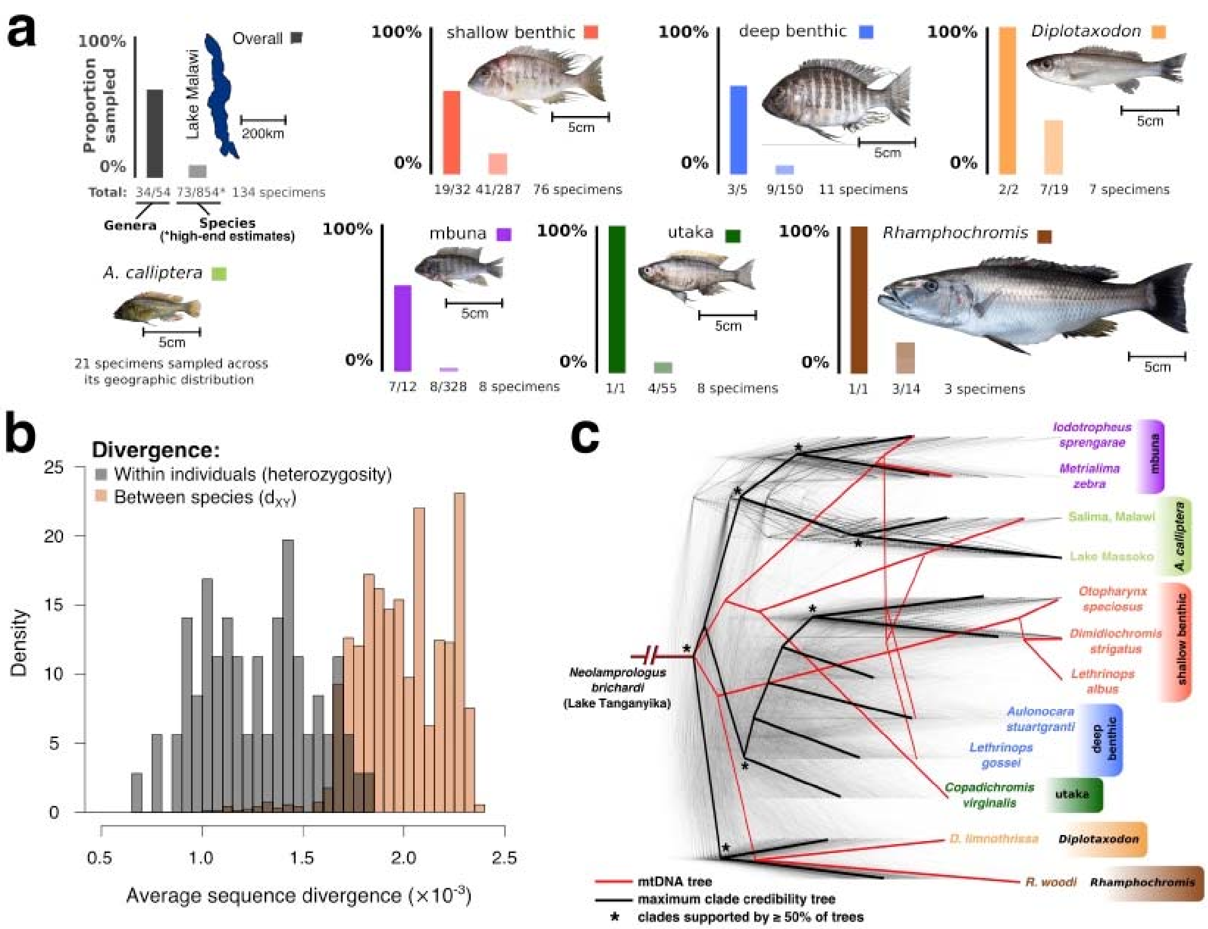

Fig. 1: The Lake Malawi haplochromine cichlid radiation. a, The sampling coverage of this study: overall and for each of the seven main eco-morphological groups within the radiation. A representative specimen is shown for each group (Diplotaxodon: *D. limnothrissa*; shallow benthic: *Lethrinops albus*; deep benthic: *Lethrinops gossei*; mbuna: *Metriaclima zebra*; utaka: *Copadichromis virginalis*; Rhamphochromis: *R. woodi*). Numbers of species and genera are based on ref. 26. **b,** The distributions of genomic sequence diversity within individuals (heterozygosity; $\hat{\pi}$) and of divergence between species (d_{XY}). **c,** A set of 2543 Maximum Likelihood (ML) phylogenetic trees for non-overlapping regions along the genome. Branch lengths were scaled for visualization so that the total height of each tree is the same. The local trees were built with 71 species sequenced to 15x coverage and then subsampled to

12 individuals representing the eco-morphological groups. The maximum clade credibility tree was built from the region phylogenies with 12 individuals. A maximum likelihood mitochondrial phylogeny is shown for comparison.

Low per-generation mutation rate

It has been suggested that the species richness and morphological diversity of teleosts in general and of cichlids in particular might be explained by elevated mutation rates compared to other vertebrates[27,28]. To obtain a direct estimate of the per-generation mutation rate, we reared offspring of three species from three different Lake Malawi groups (*Astatotilapia calliptera*, *Aulonocara stuartgranti*, and *Lethrinops lethrinus*). We sequenced both parents and one offspring of each to high coverage (40x), applied stringent quality filtering, and counted variants present in each offspring but absent in both its parents (Fig. S2)[21]. There was no evidence for significant difference in mutation rates between species. The overall mutation rate (μ) was estimated at 3.5×10^{-9} (95%CI: 1.6×10^{-9} to 4.6×10^{-9}) per bp per generation, approximately three to four times lower than the rate obtained in similar studies using human trios[29], 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. (ref. 28) 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. 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 $N_e = \pi/4\mu$). Given previous estimates of the age of the radiation of the order of a million years[12,30,31] (max estimate 4.63 million years[32]), and the hundreds of species present, this result suggests that alleles at a locus will only rarely coalesce within the time between successive speciation events, consistent with high sharing of heterozygosity and ILS. This is because both the mean and standard deviation in the time to the common ancestor of a pair of alleles are expected to be in the order of 2Ne generations, or hundreds of thousands of years.

Between-species relationships

To obtain a first estimate of between-species relationships we divided the genome into 2543 non-overlapping windows, each comprising 8000 SNPs (average size: 274kb), and constructed a Maximum Likelihood (ML) phylogeny separately for 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 (Figs. 1C, S4)[21]. Despite extensive variation among the individual trees, it is apparent that there is some general consensus. Individuals from the previously identified eco-morphological groups tend to cluster together, with *Rhamphochromis*, *Diplotaxodon*, mbuna and *A. calliptera* forming well supported (in >80% of trees) apparently reciprocally monophyletic groups (Fig. S4B), while individuals from the utaka, and deep and shallow benthic were clustered within their respective groups, but with lower support. In the subset of 12 individuals shown in Fig. 1c, the pelagic clades *Diplotaxodon* and *Rhamphochromis* tend to cluster together and form a sister group to the rest of the radiation. Perhaps surprisingly, the majority of the trees place the widely-distributed lake/river-dwelling *A. calliptera* as the sister taxon to the specialized rocky-shore mbuna group in a position that is nested within the radiation (Figs. 1c, S4B). The overall sub-structuring is also apparent from patterns of linkage disequilibrium (LD). Mean LD decays within a few hundred base-pairs across the set of all species, in a few kilobases (kb) for subsets of species from within eco-morphological subgroups (mbuna, *Diplotaxodon* etc.), and extends beyond 10kb within species (Fig. S3)[21]. Because the extent of LD is substantially shorter than the window size of the ML phylogenies, we expect extensive ILS also within the windows, and it is established that ML may be inappropriate for building trees from large regions with extensive ILS[34-36]. Therefore we note that even in the absence of introgression the ML tree results above would be merely illustrative of overall genetic similarities rather than providing definitive reconstructions of the branching order of taxa.

It is also clear that the mtDNA phylogeny is an outlier, being substantially different both from the MCC summary and from the majority of the individual trees (Figs. 1c, S5). Discordances between mtDNA and nuclear phylogenies in Lake Malawi has been reported previously in refs. [15,19], and interpreted as a signature of past hybridization events. However, large discrepancies between mitochondrial and nuclear phylogenies have been shown previously in a large number of other systems, reflecting both that mtDNA as a single locus is not expected to reflect the consensus under ILS, and that it often does not evolve neutrally (e.g. refs. 37-39). In particular, the high incidence of mitochondrial selection underlines the importance of evaluating the Lake Malawi radiation from a genome-wide perspective rather than drawing conclusions regarding species relationships based on mtDNA signals alone.

If the observed high levels of phylogenetic incongruence were due to ILS alone, it should still be possible to resolve a cleanly bifurcating species tree by applying either the multispecies coalescent model[40,41], or the distance based Neighbor-Joining method which has been shown to be a statistically consistent and accurate species tree estimator under ILS[42,43]. We constructed a whole genome NJ tree using the Dasarathy et al. algorithm (Fig. 2A). A similar overall topology was produced from applying the Bayesian multispecies coalescent method SNAPP[44] to the subset of 48,922 unlinked SNPs in 12 individuals representing the eco-morphological groups (Fig. S6; analysis of the full dataset was not computationally feasible using this method). Overall, the resulting phylogenies share many similarities that reflect features of previous taxonomic assignment, but some currently-recognized genera are clearly polyphyletic, including *Placidochromis*, *Lethrinops*, and *Mylochromis*.

Violations of the species tree concept

Previous studies 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[12,14,16,17] and Malawi[15,19]. Consistent with this, we found some variation in species tree topologies depending on which inference method was used and also within the Bayesian MCMC samples (Figs. 2A, S4B, S6). Furthermore, we contrasted the pairwise genetic distances used to produce the raw NJ tree (Fig. 2B; above the diagonal) against the distances between samples along the tree branches (Fig. 2A), calculating the residuals (Fig. 2B; below the diagonal). If the tree were able to perfectly capture all the genetic relationships in our sample, we would expect the residuals to be zero. However, we found a large number of differences, with some standout cases. The fact that we are using over 25 million variable sites, and the high bootstrap values on all branches of the NJ tree (mean support is 90%, median is 100%; Fig. 2C), suggest these differences are not due to sampling noise, but may reflect violations of the bifurcating species tree model.

The patterns of differences between the true genetic distance and the NJ tree distance affect both groups of species and individual species. Among the strongest signals on individual species, we see that: 1) *Copadichromis trimaculatus*, which in the NJ tree clusters within the shallow benthic clade, is genetically much closer to other utaka, and in particular to *C. quadrimaculatus*, than its placement in the tree would suggest; 2) *Placidochromis* cf. *longimanus* is genetically closer to the deep benthic clade and to a subset of the shallow benthic (mainly *Lethrinops* species) than the tree suggests; and 3) our sample of *Otopharynx tetrastigma* is much closer to *Astatotilapia calliptera* from Lake Kingiri (and to a lesser degree to other *A. calliptera*) than would be expected from the tree. The *O. tetrastigma* specimen comes from Lake Ilamba, a satellite crater lake of Lake Malawi that also harbours a population of *A. calliptera* and is geographically close (3.2 km) to Lake Kingiri.

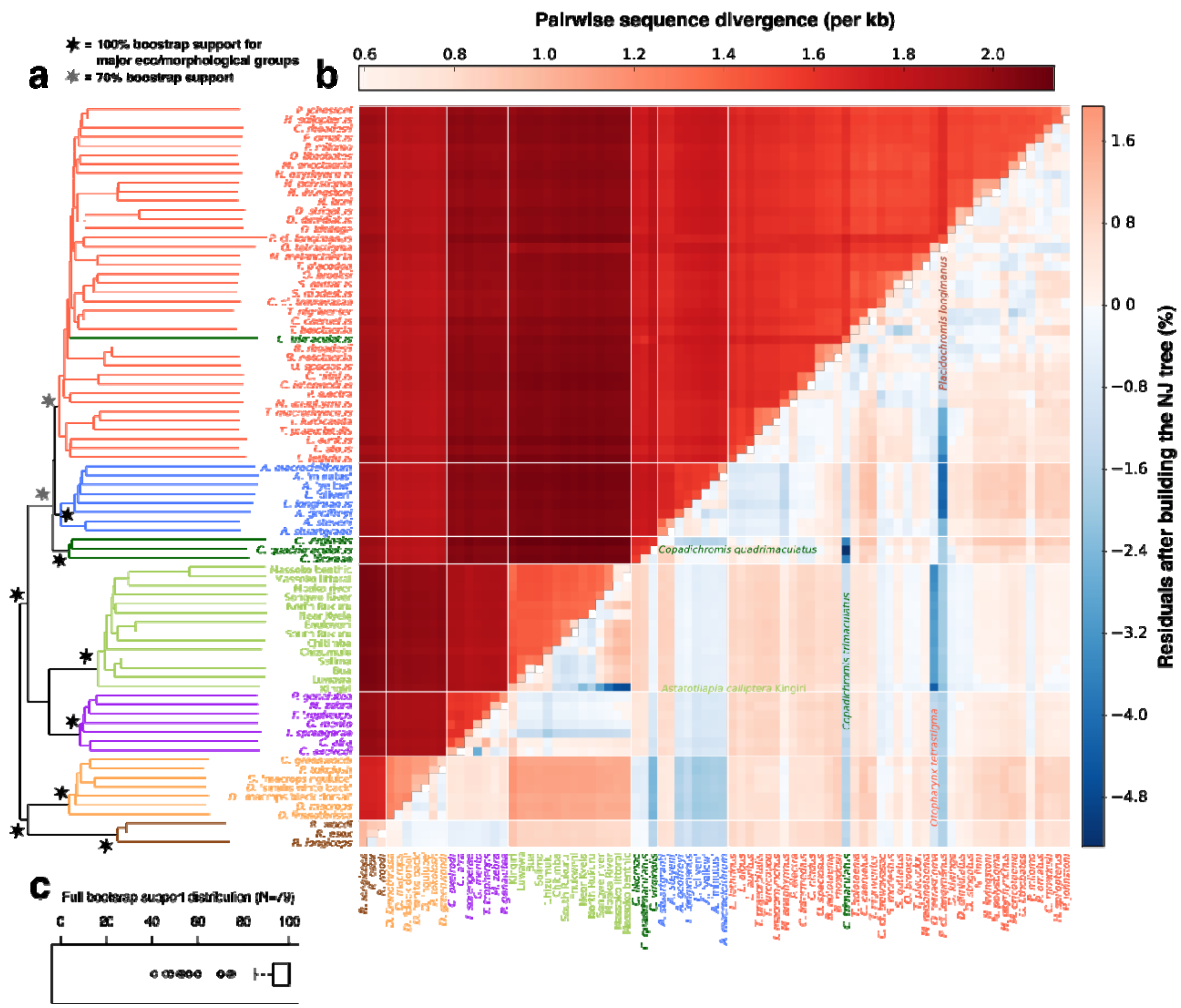

Fig. 2: Species relatedness. a, Neighbor-joining tree of pairwise differences. Long terminal long terminal branches reflect the high ratio of within-species to between-species variation. **b,** Pairwise genetic differences (above diagonal) and residuals of pairwise difference and tree distance (below diagonal). The residuals for each pair of individuals are calculated as: (sequence distance - tree distance)/sequence distance. Blue cells beneath the diagonal indicate pairs of samples that share more alleles than expected according to the tree. **c,** The distribution of block-bootstrap support values for the NJ tree.

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[45] and calculated a 'co-ancestry matrix' of all species[21] - a summary of nearest neighbour (therefore recent) haplotype relationships in the dataset. We found that the Lake Ilamba *O. tetrastigma* and Lake Kingiri *A. calliptera* stand out in this analysis by showing a strong signature of recent gene flow between distantly related species (Fig. S7). The other tree-violation signatures described above are also visible on the haplotype sharing level but are less pronounced, consistent with being older events perhaps involving the common ancestors of multiple present-day species. However, the chromopainter analysis reveals numerous other examples of excess co-ancestry between species that do not cluster immediately together (e.g. the utaka *C. virginalis* with *Diplotaxodon*; more highlighted in Fig. S7). Furthermore, the clustering based on recent co-ancestry is different from any tree generated using phylogenetic methods; in particular a number of shallow benthics including *P.* cf. *longimanus* cluster next to the deep benthics. Related to this, principal component analysis (PCA), while generally separating the major groups, shows an extension of utaka and benthic samples towards *Diplotaxodon* (Fig. 3a), a pattern that is typical for admixed populations (e.g. ref. 46). In particular among the benthics, deeper water species are closer to the deep water *Diplotaxodon*, an observation we will return to below in the context of shared mechanisms of depth adaptation.

To gather further evidence regarding potential gene flow between the 'tree-violating' taxa identified above, we computed the f_4 admixture ratio[47-49] (f statistic in the following). The f statistic is closely related to Patterson's D (ABBA-BABA test)[47], and when elevated due to introgression is expected to be linear in relation to the proportion of introgressed material. For each of the three cases discussed in the analysis of Figure 2, we found strong signals of non-tree-like relatedness (Fig. 3B). Specifically, there are two very high f statistics involving *C. trimaculatus* and shallow benthic and utaka species; it is also notable that the position of *C. trimaculatus* within the shallow benthic group is an unusual feature of the NJ tree in Fig. 2a; it clusters with the other utaka in all other phylogenies. We interpret the evidence as suggesting that the gene-pool of *C. trimaculatus* is a product of hybridization between an utaka lineage and a shallow benthic lineage.

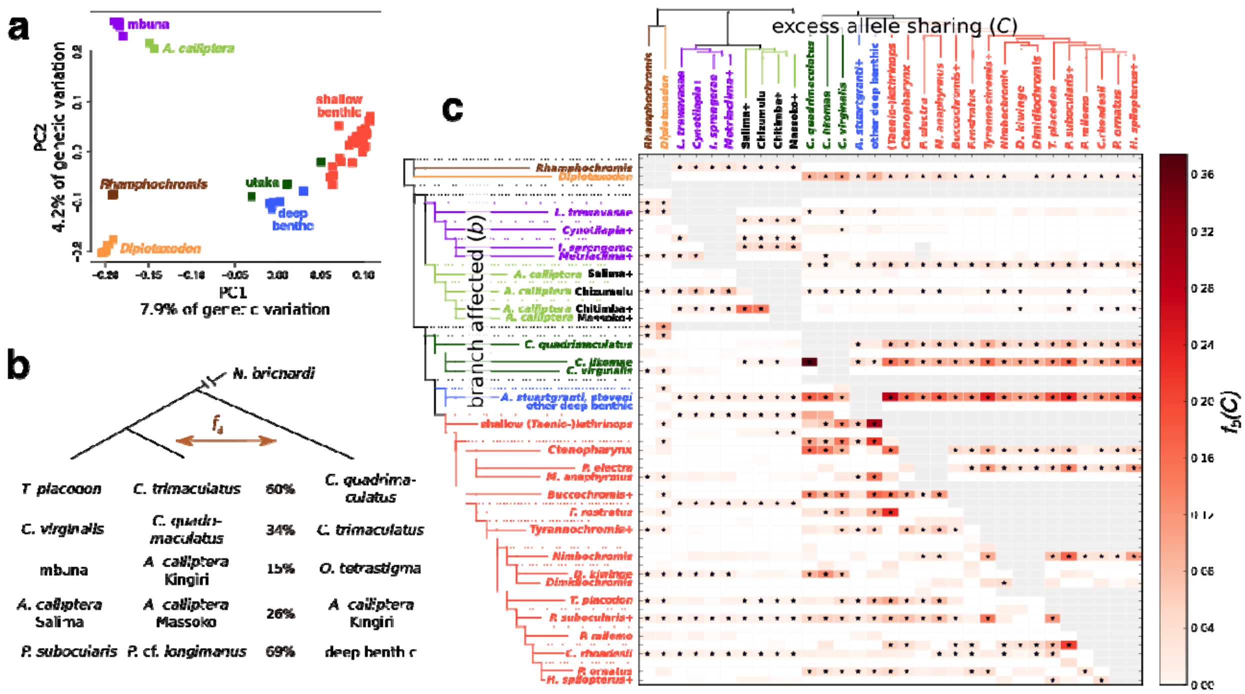

Fig. 3: Evidence for gene flow and non-tree like ancestral structure. Colors correspond to eco/morphological groups as in Fig. 1. **a,** PCA analysis. **b,** Selected strong f_4 admixture ratios (block jack-knifing p-values < 10-300). More comparisons can be found in Table S8. **c,** Branch-specific statistic f_b amongst 31 subgroups. The ++ signs in labels signify that the subgroup includes multiple or additional taxa. For a full list of samples corresponding to each subgroup see Fig. S10. The * sign denotes block jack-knifing significance at $|Z|>3.17$ (Holm-Bonferroni FWER<0.001). Grey data points in the matrix correspond to tests that are not consistent with the NJ phylogeny.

To investigate cross-species gene flow and violation of a tree-like model of species relationships even more exhaustively across the radiation, we extended the *f* analysis to all trios of species that fit the relationships ((A, B), C). Out of 85311 computed *f* statistics 56062 were significant at FWER < 0.001 (Fig. S8). However, a single gene flow event can lead to multiple significant *f* statistics: the values calculated for different combinations of ((A, B), C) groups are not independent as soon as they share internal or external branches. Therefore, we sought to obtain branch-specific estimates of excess allele sharing that would be less correlated. Building on the logic employed to understand correlated gene flow signals in ref. 50, we developed $f_b(C)$: a summary of *f* scores that captures excess allele sharing between a clade *C* and a branch *b* compared to the sister branch of *b*[21]. The $f_b(C)$ score can be interpreted as a measure of how well a tree-like branching pattern captures the genetic relationships of samples descending from against the rest of the phylogeny. This method reduces the number of calculated *f* statistics to 11452. Of these, 1725 scores are still significantly elevated (at FWER<0.001), while 105 of the

160 branches in the phylogeny show significant excess allele sharing with at least one other group *C* (Fig. S9). To summarise these results for discussion below, we partitioned the species into 31 subgroups chosen to represent the majority of the signal in the dataset, excluding the strongly admixed taxa for whom *f* was already calculated in Fig. 3B (*C. trimaculatus, A. calliptera* Kingiri, *O. tetrastigma, P.* cf. *longimanus*), and present $f_b(C)$ scores for these in Fig. 3C.

The *f* statistic tests are robust to the occurrence of incomplete lineage sorting, in the sense that ILS alone can not generate a significant test result[48]. 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[48,51]. Nevertheless, to be able to explain many of the patterns reported here, ancestral population structure would have needed to segregate through multiple 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.

Not only are there many significant *f* scores, they also are unusually large: 519 out of the 1725 significant scores (4.5% of the total 11452; Fig. S10) are larger than 3%, corresponding to inferred human-neanderthal introgression[47]. Across the subgroups in Fig. 3C, the strongest signal of $f_b(C)$ = 37% points to excess allele sharing of the branch leading to *C. likomae* with *C. quadrimaculatus* relative to *C. virginalis* and suggests that the non-treelike relationships of utaka extend beyond the clearly admixed status of *C. trimaculatus* reported above. Underlining their complicated relationships, while *C. virginalis* and *C. quadrimaculatus* are sister species in Fig. 2A, removing *C. trimaculatus* leads to *C. likomae* becoming the sister species to *C. virginalis* with *C. quadrimaculatus* basal, as seen in Fig. 3C.

Several highly significant $f_b(C)$ scores point to multiple genetic exchanges between the deep and shallow benthic groups. *Aulonacara stuartgranti* and *A. steveni*, which overall cluster with deep benthics and have enlarged lateral line sensory apparatus like many of those, share excess derived alleles with all shallow benthic subgroups [max $f_b(C)$ = 28% relative to other deep benthics], reflecting that they are typically found in shallower water[26]. Conversely, shallow benthic species of the genera *Lethrinops* and *Taeniolethrinops* show excess allele sharing with the remaining deep benthic taxa [$f_b(C)$ = 30% relative to other shallow benthics].

On a broader scale, there are strong signals of genetic exchange between major ecological groups. Most prominently, we infer excess allele sharing of all the utaka and benthics with *Diplotaxodon*, one of the two pelagic clades [$f_b(C)$ = 10% relative to mbuna and *A. calliptera*], as previously suggested by the PCA plot (Fig 3A). Furthermore, there is evidence for additional *Diplotaxodon* ancestry in utaka [$f_b(C)$ = 7%] and sub-clades of the benthics [most strongly for deep benthics at $f_b(C)$ = 3%] relative to their sister clades, which could be explained either by additional more recent gene flow events or by differential fixation of introgressed material, possibly due to selection. Reciprocally, *Diplotaxodon* shows excess allele sharing with all utaka and benthics relative to *Rhamphochromis*. On the other hand, while ref. 19 suggested gene flow between the deep benthic and mbuna groups on the basis of mtDNA phylogeny, our genome-wide analysis did not find any signal of substantial genetic exchange between these groups.

Inference using the software treemix[52] also suggests strong evidence for gene flow within and between the major clades, mainly involving *Diplotaxodon*, deep and shallow benthics and utaka (Fig. S11). However, the overall topology and specific gene flow events in treemix results depend heavily on the value of a parameter specifying the number of allowed migration events. Using simulations, we have shown that the accuracy of treemix results can be extremely sensitive to erroneous inferences of the initial phylogeny, whereas the interpretation of $f_b(C)$ scores as a measure of violation of a tree-like branching model is more robust in this respect[21].

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

Origins of the radiation

Astatotilapia calliptera, although showing strong preference for shallow weedy habitats, is abundant and widespread in both flowing and still waters, and is thus considered a generalist. It has often been referred to as the 'prototype' for the closely-related endemic genera of Lake Malawi cichlids[26]. Therefore, discussions concerning the origin of the Malawi radiation often rely on ascertaining the relationship of this species to the rest of the Malawi radiation[15,53]. Previous phylogenetic analyses, using mtDNA and small numbers of nuclear markers, showed inconsistencies with respect to *A. calliptera* (compare refs. 15 and 19). On the other hand, our whole genome data indicate a clear and consistent position of all the *A. calliptera* individuals from the Lake Malawi catchment as members of a sister group to the mbuna, consistent with the nuclear DNA phylogeny in ref. 19, although that study included only a single *A. calliptera* specimen.

To explore the origins of the Lake Malawi radiation in greater detail, we obtained whole genome sequences from 19 individuals from seven *Astatotilapia* species not found in Lake Malawi (Table S2) and generated new variant calls[21]. In addition, we sequenced five more *A. calliptera* specimens from Indian Ocean catchments, thus covering most of the geographical distribution of the species. We constructed NJ trees based on genetic distances and found that even with these additional data all the *A. calliptera* (including samples from outside the Lake Malawi catchment) continue to cluster as a single group at the same place nested within the radiation, whereas the other *Astatotilapia* species clearly branched off well before the lake radiation (Fig. 4a,b,c).

Joyce et al. (ref. 15) reported that the mtDNA haplogroup of *A. calliptera* from the Indian Ocean catchment clustered with mbuna (as we see for mtDNA tree in Fig. S12) and concluded that the phylogenetic discordances between mtDNA and nuclear markers can be explained by gene-flow. They suggested there had been repeated colonization of Lake Malawi by independent *Astatotilapia* lineages with different mitochondrial haplogroups, the first founding the entire species flock, and the second, with the mtDNA haplogroup common in the Indian Ocean catchment, introgressing into the Malawi radiation and contributing strongly to the mbuna group[15]. This hypothesis predicts that among the *A. calliptera*, the Indian Ocean catchment individuals should be closer to mbuna than the individuals sampled within the Malawi catchment. However, using the *f* statistics, we found a strong signal in the opposite direction across the nuclear genome (*f*= 30%; Fig. 4d). Therefore, the Joyce et al. hypothesis based on the mtDNA phylogeny in not supported by genome-wide data.

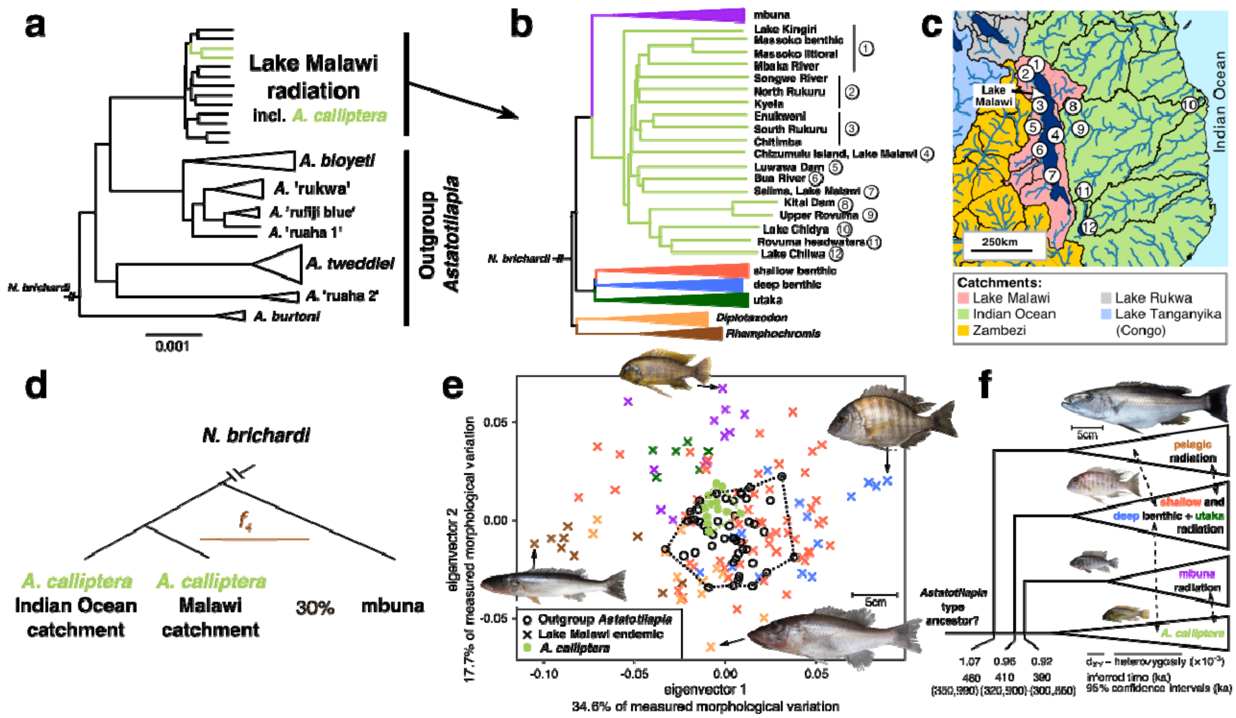

Fig. 4: Origins of the radiation and the role of *A. calliptera*. a, An NJ phylogeny showing the Lake Malawi radiation in the context of other East African *Astatotilapia* taxa. **b,** A Lake Malawi NJ phylogeny with expanded view of *A. calliptera*, with all other groups collapsed. **c,** Approximate *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, translation of this value into a mean time estimate, and 95% CI for the time estimate reflecting the statistical uncertainty in mutation rate. Dashed lines with arrows indicate likely instances of gene flow between major groups; their absolute timing (position along the x axis) is arbitrary.

It has been repeatedly suggested that *A. calliptera* may be the direct descendant of the riverine-generalist lineage that seeded the Lake Malawi radiation [e.g. refs. 7,15,53-55]. This hypothesis is lent further support by our geometric morphometric analysis. Using 17 homologous body shape landmarks[21] we established that, despite the relatively large genetic divergence, *A.*

calliptera is nested within the morphospace of the other more distantly related but ecologically similar *Astatotilapia* species (Figs. 4a,e), and these together have a central position within the morphological space of the Lake Malawi radiation (Figs. 4e, S13).

Therefore, we propose here a model which reconciles the nested phylogenetic position of *A. calliptera* in Fig. 4b with its ancestral riverine phenotype. We suggest a model in which the Lake Malawi species flock consists of three separate radiations splitting off the ancestral lineage leading to *A. calliptera*; the pelagic radiation was seeded first, then the benthic + utaka, and finally the rock-dwelling mbuna, all in a relatively quick succession, followed by subsequent gene flow as described above (Fig. 4f). Using our per-generation mutation rate we obtained mean separation time estimates for these lineages between 460 thousand years ago (ka) [95%CI: (350ka to 990ka)] and 390ky [95%CI: (300ka to 860ka)] (Fig. 4f), assuming three years per generation as in ref. 56. The point estimates all fall within the second most recent prolonged deep lake phase as inferred from the Lake Malawi paleoecological record[30] while the upper ends of the confidence intervals cover the third deep lake phase. We also note that split times estimated from sequence divergence are likely to be reduced by subsequent gene-flow, leading to underestimates. Therefore we conclude that the data are consistent with the previous reports based on fossil time calibration which put the origin of the Lake Malawi radiation at 700-800ka[12].

Focusing on the *A. calliptera* individuals, we found they cluster by geography (Fig. 4b,c), except for the specimen from crater Lake Kingiri, whose position in the clade is likely a result of the admixture signals shown in Fig. 3b. Indeed, the Kingiri individual clusters according to geography with the specimens from the nearby crater Lake Massoko and Mbaka River if a NJ tree is built with *A. calliptera* samples only (Fig. S14). Applying the same logic, 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 the *A. calliptera* position were affected by hybridization with mbuna). However, we found that the position of *A. calliptera* is not affected by the removal of mbuna (Fig. S15), suggesting that the nested position is not due to later hybridization. The f statistic analysis in Fig. 3c further supports this claim, because the signals involving the whole mbuna or *A. calliptera* groups are modest and do not suggest erroneous placement of these whole groups in all phylogenetic analyses.

Furthermore, the nested position of *A. calliptera* is also supported by the vast majority of the genome. We specifically searched for the basal branch in a set of 2638 local ML phylogenies for non-overlapping genomic windows and found results that are consistent with the whole genome NJ tree: the most common basal branches are the pelagic groups *Rhamphochromis* and *Diplotaxodon* (in 42.12% of the genomic windows). In comparison, *A. calliptera* (including all of the Indian Ocean catchment samples) were found to be basal only in 5.99% of the windows (Fig. S16).

Finally, we note that all the *A. calliptera* have a relatively recent common ancestor, with divergence at ~75% of the most distinct species in the Malawi radiation and corresponding to 340ka [95%CI: (260ka, 740ka)], suggesting that the Lake Malawi population has been a reservoir that has repopulated the river systems and more transient lakes following dry-wet transitions in East African hydroclimate[30,57]. 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 that is it not likely that the many strong phenotypic affinities of *A. calliptera* to the basal *Astatotilapia* [see refs. 58,59; Fig. 4e], would be reinvented from a lacustrine species. After all, *A. calliptera* is widespread and abundant in its preferred shallow weedy habitat throughout present-day Lake Malawi, despite the presence of hundreds of closely-related endemic lacustrine specialist species.

Signatures and consequences of selection

To gain insight into the functional basis of diversification and adaptation in Lake Malawi cichlids, we next turned our attention to protein coding genes. We compared the between-species levels of non-synonymous variation $\bar{p}_N$ to synonymous variation $\bar{p}_S$ in over 20,000 genes and calculated the difference between these two values ($\delta_{N-S} = \bar{p}_N - \bar{p}_S$)[21]. Overall, coding sequence exhibits signatures of purifying selection: the average between-species $\bar{p}_N$ was 54% lower than in a random matching set of non-coding regions. Interestingly, the average between-species synonymous variation $\bar{p}_S$ in genes was slightly but significantly higher than in non-coding controls (13% lower mean; $p < 2.2 \times 10^{-16}$, one tailed Mann-Whitney test). 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.

Average per-gene non-synonymous excess variation (δ_{N-S}) calculated between Lake Malawi species correlates only relatively weakly with δ_{N-S} for genes in the five cichlid genome assemblies presented by Brawand *et al.*[11] which represent one from each of the major lineages and radiations of East African cichlids, and are approximately an order of magnitude more divergent than Malawi cichlids (Spearman $\rho_S = 0.32$; Fig. S17A). Thus the majority of genic selection within the Malawi radiation appears distinct from selection acting on these longer timescales between radiations. On the other hand, Malawi between-species δ_{N-S} correlates substantially more with δ_{N-S} calculated between *A. calliptera* populations ($\rho_S = 0.49$; Fig. S17B). Assuming that present day *A. calliptera* populations segregate more ancestral alleles than phenotypically more derived species, this result would suggest that within-Malawi selection was influenced by the diversity of alleles present in the generalist ancestor of the radiation. An alternative hypothesis is that the different *A. calliptera* populations may be acquiring these alleles through gene flow from the derived Lake Malawi species.

To control for statistical effects stemming from variation in gene length and sequence composition we normalized the δ_{N-S} values per gene by taking into account the variance in $p_N - p_S$ across all pairwise sequence comparisons for each gene, deriving the non-synonymous excess score (Δ_{N-S})[21]. The genes with highly positive Δ_{N-S} are likely to be under positive selection. We focus below on the top 5% of the Δ_{N-S} distribution ($\Delta_{N-S} > 40.2$, 1034 candidate genes; Fig. 5a).

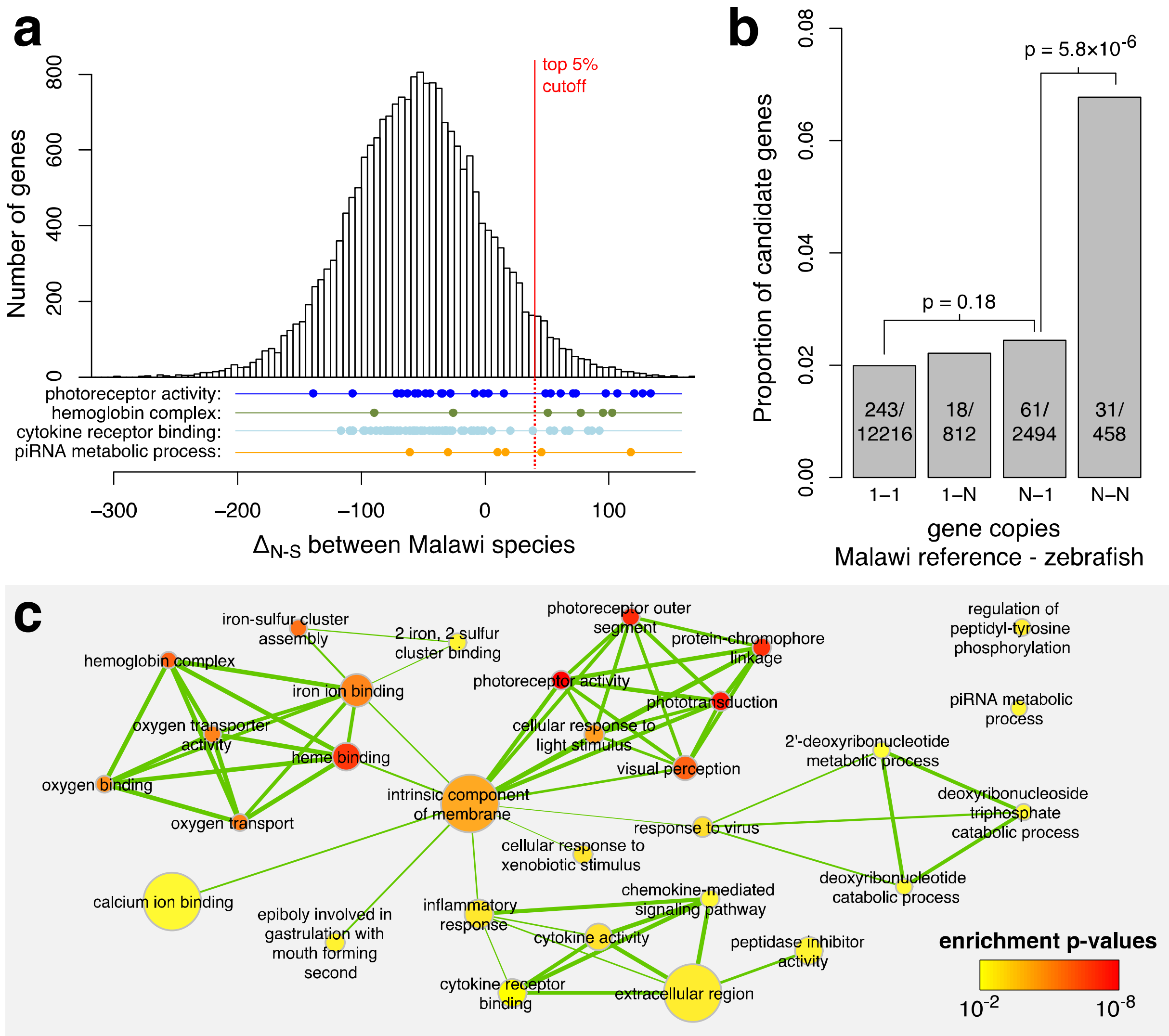

Figure 5: Gene selection scores, copy numbers, and ontology enrichment. **a,** The distribution of the non-synonymous variation excess scores ($\boldsymbol{\Delta_{N-S}}$) highlighting the 5%FDR cutoff, and the distributions of genes in selected Gene Ontology (GO) categories. **b,** The relationship between the probability of $\boldsymbol{\Delta_{N-S}}$ being positive and in the top 5% and the relative copy numbers of genes in the Lake Malawi reference (*M. zebra*) and zebrafish. The p-values are based on χ^2 tests of independence. 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 color of each node indicates the p-value for the term and the size of the node is proportional to the number of genes annotated with that GO category.

This candidate gene set is highly enriched for genes for which no homologs were found in any of Medaka, Stickleback, Tetraodon or Zebrafish (other teleosts) when examined in ref. 11 (606 out of 4,190 without vs. 428 out of 16,472 with homology assignment; χ^2 test $p < 2.2\times10^{-16}$). Genes without homologs tend to be short (median coding length is 432bp) and some of the signal may be explained by a component of gene prediction errors. However a comparison of short genes (≤450bp) without homologs to a set of random noncoding sequences (Fig. S18) showed significant differences ($p < 2.2\times10^{-16}$, Mann-Whitney test), with both a substantial component of genes with low $\bar{p}_N$, reflecting genes under purifying selection, and also an excess of genes with high $\bar{p}_N$ (Fig. S19).

Cichlids have an unexpectedly large number of gene duplicates and it has been suggested that this phenomenon has contributed to their extensive adaptive radiations[3,11]. To investigate the extent of divergent selection on gene duplicates, we examined how the non-synonymous excess scores are related to gene copy numbers in the reference genomes. Focusing on homologous genes annotated both in the Malawi reference (*M. zebra*) and in the zebrafish genome, we found that the highest proportion of candidate genes was among genes with 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) relative to 1 - 1 genes, is of a much lesser magnitude and is not significant (χ^2 test $p < 0.18$), suggesting that selection is occurring more often within ancient multi-copy gene families, rather than on genes with cichlid-specific duplications.

Next we used Gene Ontology (GO) annotation of zebrafish homologs to test whether candidate genes are enriched for particular functional categories. We found significant enrichment for 30 GO terms [range: $1.6\times10^{-8} < p < 0.01$, `weigh` algorithm[21,60]; Table S3], 10 in the Molecular Function (MF), 4 in the Cellular Component (CC), and 16 in Biological Process (BP) category. Combining the results from all three GO categories in a network connecting terms with high overlap (i.e. sharing many genes) revealed clear clusters of enriched terms related to (i) haemoglobin function and oxygen transport; (ii) phototransduction and visual perception; and (iii) the immune system, especially inflammatory response and cytokine activity (Fig. 5c). It has been previously suggested that evolution of genes in these functional categories has contributed to cichlid radiations (as discussed below); it is nevertheless interesting to see that these categories stand out in an unbiased analysis of all 20,000+ genes in the genome.

Shared mechanisms of depth adaptation

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

Sharply decreasing levels of dissolved oxygen and low light intensities with narrow short wavelength spectra are the hallmarks of the habitats at below ~50 meters to which the deep benthic and pelagic *Diplotaxodon* groups have both adapted, either convergently or in parallel[61]. Signatures of selection on similar haplotypes in the same genes involved in vision and in oxygen transport would therefore point to shared molecular mechanisms underlying this ecological parallelism.

To obtain a quantitative measure of shared molecular mechanisms, we calculated for each gene a similarity score for deep benthic and *Diplotaxodon* amino acid sequences and also compared the amounts of non-coding variation in these depth-adapted groups against the rest of the radiation[21]. Both measures are elevated for candidate genes in the 'visual perception' category (Fig. 6a; p=0.007 for similarity, p=0.08 for shared diversity, and p=0.003 when similarity and diversity scores are added; all p-values based on Mann-Whitney test). The measures are also elevated for the 'haemoglobin complex' category, although due to the small number of genes the differences are not statistically significant in this case. Furthermore, the level of excess allele sharing between *Diplotaxodon* and deep benthic [measured by the local f statistic f_{dM}[49,56]] is strongly correlated with the $\Delta_{\text{N-S}}$ selection score for genes annotated with photoreceptor activity and haemoglobin complex GO terms ($\rho_S = 0.63$ and 0.81, $p = 0.001$ and $p = 0.051$, respectively, Fig. 6b).

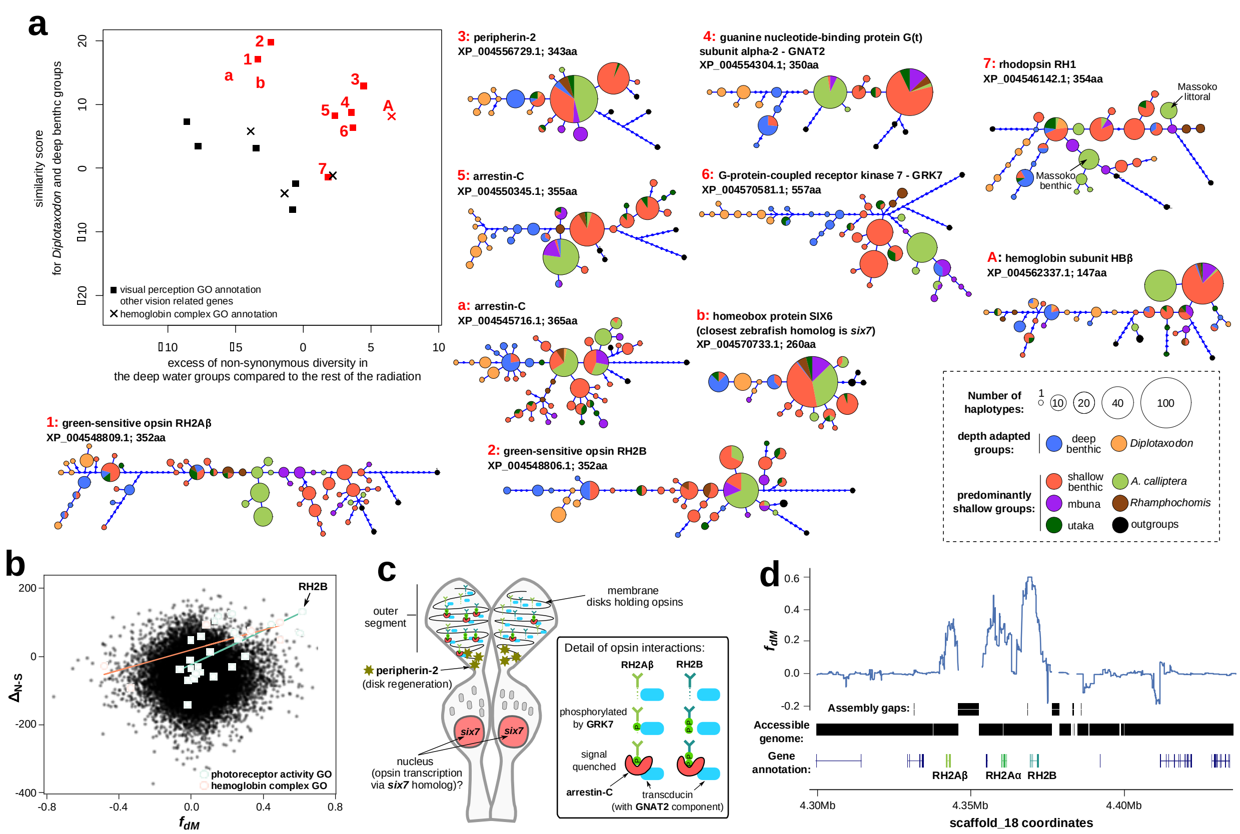

Fig. 6: Shared selection between the deep water adapted groups *Diplotaxodon* and deep benthic. a, The scatterplot shows the distribution of genes with high Δ_{N-S} scores (candidates for positive selection) along axes reflecting shared selection signatures. Only genes with zebrafish homologs are shown. Amino acid haplotype genealogies, 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 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 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.

Vision genes with high similarity and diversity scores for the deep benthic and *Diplotaxodon* groups include three opsin genes: the green sensitive opsins RH2Aβ, RH2B, and rhodopsin (Figs. 6a, S20A). The specific residues that distinguish the deep adapted groups from the rest of the radiation differ between the two RH2 copies, with only one shared mutation out of a possible fourteen (Fig. S20B). RH2Aβ and RH2B are located within 40kb from each other on the same chromosome (Fig. 6c); a third paralog, RH2Aα, is located between them, but it has very little coding diversity specific to deep benthic and *Diplotaxodon* (Fig. S21). This finding is consistent with previous reports suggesting functional divergence between RH2Aα and RH2Aβ following the duplication of RH2A early in the cichlid lineage[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[56]. The haplotype genealogy presented here for the broader radiation strongly suggests that the Massoko alleles did not originate by mutation in that lake but were selected out of ancestral variation (Fig. 6a).

The long wavelength, red-sensitive opsin (LWS) has been shown to play a role in speciation along a depth gradient in Lake Victoria[65]. While it is not particularly diverse in *Diplotaxodon* and deep benthics, it is interesting to note that *Diplotaxodon* have haplotypes that are clearly distinct from those in the rest of the radiation, while the majority of deep benthic haplotypes are their nearest neighbours (Fig. S21). The short-wavelength opsin SWS1 is among the genes with high $\Delta_{N\text{-}S}$ scores but it does not exhibit shared selection between *Diplotaxodon* and deep benthics - it is most variable within the shallow benthic group. Finally, the short-wavelength opsins SWS2A and SWS2B have negative $\Delta_{N\text{-}S}$ scores in our Lake Malawi dataset and thus are not among the candidate genes.

There have been many previous studies of selection on opsin genes in fish [e.g. reviewed in [66-68]], 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 the genealogies of six other vision associated candidate genes: a homolog of the homeobox protein *six7*, the G-protein-coupled receptor kinase GRK7, two copies of the retinal cone arrestin-C, the α-subunit of cone transducin GNAT2, and peripherin-2 (Fig. 6a). The functions of these genes suggest a prominent role of cone cell vision in depth adaptation. The homeobox protein *six7* governs the expression of RH2 opsins and is essential for the development of green cones in zebrafish[69]. One of the variants in this gene that distinguishes deep benthic and *Diplotaxodon* is just a residue away from the DNA binding site of the HOX domain, while another is located in the SIX1_SD domain responsible for binding with the transcriptional activation co-factor of *six7*[70] (Fig. S22C). The kinase GRK7 and the retinal cone arrestin-C genes have complementary roles in photoresponse recovery, where arrestin produces the final shutoff of the cone pigment following phosphorylation by GRK7, thus determining the temporal resolution of motion vision[71]. Note that bases near to the C-terminus in RH2Aβ mutated away from serine (S290Y and S292G), thus reducing the number of residues that can be modified by GRK7 (Fig. S22B). 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[72]. The final gene, peripherin-2, is essential to the development and renewal of the membrane system in the outer cell segments that hold the opsin pigments in both rod and cone cells[73]. Cichlid green-sensitive opsins are expressed exclusively in double-cone photoreceptors and the wavelength of maximum absorbance in cells expressing a mixture of RH2Aβ with RH2B (λ_{max} = 498nm) corresponds to the part of light spectrum that transmits the best into deep water in Lake Malawi[68]. Figure 6C illustrates the possible interactions of all the above genes in a double-cone photoreceptor of the cichlid retina.

Haemoglobin genes in teleost fish are located in two separate chromosomal locations: the minor 'LA' cluster and the major 'MN' cluster[74]. The region around the LA cluster has been highlighted by selection scans among four *Diplotaxodon* species by Hahn et al.[75], 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 (Fig. S22). It appears that shared patterns of depth selection may be particular to the β-globin genes (Fig. S22B), although this hypothesis must remain tentative due to the highly repetitive nature of the MN cluster limiting our ability to confidently examine variation in all the haemoglobin genes in the region.

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 genealogies and f_{dM} statistics summarizing local patterns of excess allele sharing, there is evidence for each of these processes acting on different genes. The haplotype genealogies for rhodopsin and HBβ have outgroup taxa appearing at multiple locations on their haplotype networks, while *A. calliptera* specimens also appear at divergent positions (Fig. 6a). This suggests that the haplotype diversity of these genes may reflect ancient differences in the founders. In contrast, networks for the green cone genes show patterns more consistent with the Malawi radiation all being derived with respect to outgroups (or with us not having sampled a source of ancestral variation) and we found substantially elevated f_{dM} scores extending for around 40kb around the RH2 cluster (Fig. 6d), consistent with adaptive introgression in a pattern reminiscent of mimicry loci in *Heliconius* butterflies[76]. In contrast, the peaks in f_{dM} scores around peripherin-2 and one of the arrestin-C genes are relatively narrow, with boundaries that correspond almost exactly to the gene boundaries. Furthermore, these two genes have elevated f_{dM} scores only for non-synonymous variants Fig. S23, while synonymous variants do not show any excess allele sharing between *Diplotaxodon* and deep benthics. Due to the close proximity of non-synonymous and synonymous sites within the same gene, this suggests that for these two genes there may have been independent selection on the same *de novo* mutations.

Discussion

Genome sequences form the substrate for evolution. Here we have described genome variation at the full sequence level across the Lake Malawi haplochromine cichlid radiation. We focused on ecomorphological diversity, representing more than half the genera from each major group rather than obtaining deep coverage of any particular group. Therefore, we have more samples from the morphologically highly diverse benthic lineages than, for example, the mbuna where many species are largely recognised by colour differences.

The observation that cichlids within an African Great Lake radiation are genetically very similar is not new[77], but we now quantify the relationship of this to within-species variation, and the consequences for variation in local phylogeny across the genome. The observation of within-species diversity being relatively low for vertebrates, at around 0.1%, suggests that low genome-wide nucleotide diversity levels do not necessarily limit rapid adaptation and speciation. This contrasts with the suggestion that high diversity levels may have been important for rapid adaptation in Atlantic killifish[78]. One possibility is that in cichlids repeated selection has maintained diversity in adaptive alleles for a range of traits that support ecological diversification, as appears to be the case for sticklebacks[79].

We provide evidence that gene flow during the radiation, although not ubiquitous, has certainly been extensive. Overall, the numerous violations of the bifurcating species tree model suggest that full phylogenomic resolution of interspecies relationships in this system will require network approaches (see e.g. ref. 41; section 6.2). However, the majority of the signals affect groups of species, suggesting events involving their common ancestors, or are between closely-related species within the major ecological groups. We see only one strong and clear example of recent gene flow between individual more distantly-related species, not within Lake Malawi itself but between *Otopharynx tetrastigma* from crater Lake Ilamba and local *A. calliptera*. Lake Ilamba is very turbid and the scenario is reminiscent of cichlid admixture in low visibility conditions in Lake Victoria[80]. 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[30].

Our suggested model of the early stages of radiation in Lake Malawi (Fig. 4f) is broadly consistent with the model of initial separation by major habitat divergence[24], although we propose a refinement in which there were three relatively closely-spaced separations from a generalist *Astatotilapia* type lineage, initially of pelagic genera *Rhamphochromis* and *Diplotaxodon*, then of shallow- and deep-water benthics and utaka [a clade which includes Kocher's sand dwellers[24,26]], and finally of mbuna (Fig. 4f). Thus, we suggest that Lake Malawi contains three separate haplochromine cichlid radiations, stemming from the *Astatotilapia calliptera* riverine generalist lineage, and interconnected by subsequent gene flow.

The finding that cichlid-specific gene duplicates do not tend to diverge particularly strongly in coding sequences (Fig. 5b) suggests that other mechanisms of diversification following gene duplications may be more important in cichlid radiations. Divergence via changes in expression patterns has been illustrated and discussed in ref. 11. Future studies that address larger scale structural variation between cichlid genomes will be able to assess the contribution of differential retention of duplicated genes.

The evidence concerning shared adaptation of the visual and oxygen-transport systems to deep-water environments between deep 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[81] and has a similar light attenuation profile but a steeper oxygen gradient than Lake Malawi[61].

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

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. Joyce, D. A. *et al.* Repeated colonization and hybridization in Lake Malawi cichlids. *Curr. Biol.* **21,** R108–9 (2011).
16. 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).
17. 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
18. 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).
19. 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).
20. Moran, P., Kornfield, I. & Reinthal, P. N. Molecular Systematics and Radiation of the Haplochromine Cichlids (Teleostei: Perciformes) of Lake Malawi. *Copeia* **1994,** 274 (1994).
21. See supplementary materials.
22. Leffler, E. M. *et al.* Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* **10,** e1001388 (2012).
23. 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).
24. Kocher, T. D. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.* **5,** 288–298 (2004).
25. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
26. Konings, A. *Malaŵi Cichlids in Their Natural Habitat*. (Cichlid Press, 2007).
27. Ravi, V. & Venkatesh, B. Rapidly evolving fish genomes and teleost diversity. *Curr. Opin. Genet. Dev.* **18,** 544–550 (2008).
28. 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).

29. 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).
30. 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).
31. Malinsky, M. & Salzburger, W. Environmental context for understanding the iconic adaptive radiation of cichlid fishes in Lake Malawi. *Proceedings of the National Academy of Sciences* **113,** 11654–11656 (2016).
32. Genner, M. J. *et al.* Age of cichlids: new dates for ancient lake fish radiations. *Mol. Biol. Evol.* **24,** 1269–1282 (2007).
33. Heled, J. & Bouckaert, R. R. Looking for trees in the forest: summary tree from posterior samples. *BMC Evol. Biol.* **13,** 221 (2013).
34. Mendes, F. K. & Hahn, M. W. Gene Tree Discordance Causes Apparent Substitution Rate Variation. *Syst. Biol.* **65,** 711–721 (2016).
35. Roch, S. & Steel, M. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor Popul Biol* **100C,** 56–62 (2014).
36. Springer, M. S. & Gatesy, J. The gene tree delusion. *Mol. Phylogenet. Evol.* **94,** 1–33 (2016).
37. Ballard, J. W. O. & Whitlock, M. C. The incomplete natural history of mitochondria. *Mol Ecol* **13,** 729–744 (2004).
38. Toews, D. P. L. & Brelsford, A. The biogeography of mitochondrial and nuclear discordance in animals. *Mol Ecol* **21,** 3907–3930 (2012).
39. 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).
40. Edwards, S. V. Is a new and general theory of molecular systematics emerging? *Evolution* **63,** 1–19 (2009).
41. Edwards, S. V. *et al.* Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* **94,** 447–462 (2016).
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. 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).
45. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLoS Genet.* **8,** e1002453 (2012).
46. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
47. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
48. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28,** 2239–2252 (2011).

49. 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).
50. Martin, S. H. *et al.* Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* **23,** 1817–1828 (2013).
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. Genner, M. J., Ngatunga, B. P., Mzighani, S., Smith, A. & Turner, G. F. Geographical ancestry of Lake Malawi's cichlid fish diversity: Figure 1. *Biol. Lett.* **11,** 20150232 (2015).
54. Eccles, D. H. & Trewavas, E. *Malawian Cichlid Fishes*. (Lake Fish Movies, 1989).
55. 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).
56. Malinsky, M. *et al.* Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* **350,** 1493–1498 (2015).
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. Seehausen, O. *et al.* Speciation through sensory drive in cichlid fish. *Nature* **455,** 620–626 (2008).
66. Bowmaker, J. K. & Hunt, D. M. Evolution of vertebrate visual pigments. *Current Biology* **16,** R484–R489 (2006).
67. 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).
68. 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).
69. 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).
70. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45,** D200–D203 (2017).
71. 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).
72. 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).
73. Boesze-Battaglia, K. & Goldberg, A. F. X. Photoreceptor renewal: a role for peripherin/rds. *Int. Rev. Cytol.* **217,** 183–225 (2002).
74. 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).
75. Hahn, C., Genner, M. J., Turner, G. T. & Joyce, D. A. The genomic basis of adaptation to the deep water'twilight zone'in Lake Malawi cichlid fishes. *bioRxiv* (2017).
76. Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487,** 94–98 (2012).
77. 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).
78. Reid, N. M. *et al.* The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science* **354,** 1305–1308 (2016).
79. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484,** 55–61 (2012).
80. Seehausen, O. Cichlid Fish Diversity Threatened by Eutrophication That Curbs Sexual Selection. *Science* **277,** 1808–1811 (1997).
81. Konings, A. *Tanganyika cichlids in their natural habitat*. (Cichlid Press, 2015).
82. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org* **q-bio.GN,** (2013).
83. DePristo, M. A. M. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
84. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
85. 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).
86. 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).

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

88. 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).

89. Harris, R. S. Improved pairwise alignment of genomic DNA. (PhD Thesis: The Pennsylvania State University, 2007).

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

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

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

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

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

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

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

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

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

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

100. 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).

101. 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).

102. Rohlf, F. J. tpsDig2.

103. 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).

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

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

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

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

108. 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).

109. Ribbink, A. J., Marsh, B. A., Marsh, A. C., Ribbink, A. C. & Sharp, B. J. A preliminary survey of the cichlid fishes of rocky habitats in Lake Malawi. *S. Afr. J. Zool.* **18,** 149–310 (1983).
110. *The cichlid diversity of Lake Malawi/Nyasa/Niassa.* (Cichlid Press, 2004).
111. Genner, M. J. *et al.* Reproductive isolation among deep-water cichlid fishes of Lake Malawi differing in monochromatic male breeding dress. *Mol Ecol* **16,** 651–662 (2007).
112. Joyce, D. A. *et al.* An extant cichlid fish radiation emerged in an extinct Pleistocene lake. *Nature* **435,** 90–95 (2005).
113. Schwarzer, J. *et al.* Repeated trans-watershed hybridization among haplochromine cichlids (Cichlidae) was triggered by Neogene landscape evolution. *Proceedings of the Royal Society B: Biological Sciences* **279,** 4389–4398 (2012).
114. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* **12,** e1004842 (2016).

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