Supplementary Information

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

Milan Malinsky, Hannes Svardal, Alexandra M. Tyers, Eric A. Miska, Martin J. Genner, George F. Turner, and Richard Durbin
correspondence to: millanek@gmail.com (MM), rd@sanger.ac.uk (RD)

Table of contents:

Supplementary Methods 2
Supplementary Note 7
I. Sample selection
II. Simulations and treemix results
III. Gene selection and cichlid-specific genes
IV. Non-green cone opsins are not involved in shared depth adaptation
Supplementary Figures 11
Supplementary Tables 46
Additional References 54

Supplementary Methods

DNA extraction and sequencing:
DNA was extracted from fin clips using the PureLink® Genomic DNA extraction kit (Life Technologies). Genomic libraries for paired-end sequencing on the Illumina HiSeq 2000 machine were prepared according to the Illumina TruSeq HT protocol to obtain paired-end reads with mean insert size of 300-500bp. As detailed in Supplementary Table 1, we used either Illumina HiSeq v3 chemistry (generating 100bp paired-end reads) or Illumina HiSeq v4 reagents (125bp paired-end reads). Low coverage (~6x) samples with v4 reagents were multiplexed 12 per lane. High coverage (~15x) v4 samples were multiplexed four per lane. For high coverage (~15x) v3 samples, a multiplexed library with 8 samples was sequenced over three lanes. The nine trio samples were multiplexed across eight lanes using the v3 chemistry, delivering approximately 40x coverage per individual. Raw data have been deposited at the NCBI Sequence Read Archive under BioProject PRJEB1254; sample accessions are listed in Supplementary Table 4.

Alignment:
Reads were aligned to the *Metriaclima zebra* reference assembly version 1.1[11] (Supplementary Table 5) using the `bwa-mem v.0.7.10` algorithm[82] with default options. For the trio parent samples that were used in the main analysis, we aligned data from only three of the eight lanes, aiming to have the same genome coverage for the trios as for the remaining (15x) samples. For each sample, 96-98% of reads could be aligned to the reference. Duplicate reads were marked on both per-lane and per sample basis using the `MarkDuplicates` tool from the `Picard` software package with default options (http://broadinstitute.github.io/picard). Local realignment around indels was performed on both per lane and per sample basis using the `IndelRealigner` tool from the GATK v3.3.0 software package[114].

Variant calling, filtering, and genotype refinement:
Briefly, SNP and short indel variants against the *M. zebra* reference were called independently using GATK v3.3.0 haplotype caller[83] and samtools/bcftools v.1.1[84]. Variant filtering was then performed on the GATK variant calls using hard filters based on overall depth, quality by depth, excess missingness, excess of reads with zero mapping quality, strand/mapping bias, and inbreeding coefficient (see below). After filtering the GATK dataset, we performed an intersection of GATK and samtools sites and kept only variant sites present in both datasets. If the GATK and samtools alleles differed at a particular locus, we kept the GATK allele. At this point, multiallelic sites were excluded and we used genotype likelihoods output by GATK sites to perform genotype refinement, imputation, and phasing in BEAGLE v.4.0[85]. Indels were retained for the genotype refinement step, but later excluded from analyses using `vcftools v0.1.12b` option `--remove-indels`.

The particular commands/parameters used were:
GATK haplotype caller (per sample):

```
java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R REFERENCE.fa --
emitRefConfidence GVCF --variant_index_type LINEAR --variant_index_parameter
128000 -I SAMPLEn.bam -o GATK_SAMPLEn.g.vcf
```

Haplotype caller per-sample files were combined using the GATK `GenotypeGVCFs` tool with the `--includeNonVariantSites` option so that every basepair of the assembly was represented in the multisample VCF file.

Hard filters were applied to the following GATK annotations:

```
Minimal inbreeding coefficient: 'InbreedingCoeff < -0.6'
Minimum overall read depth: 'DP < 1000'
Maximum overall read depth: 'DP > 3000' (except for mtDNA: scaffolds 747,2036)
Max phred-scaled p-value from Fisher's exact test to detect strand bias: 'FS > 40.0' (except for mtDNA: scaffolds 747,2036)
QualityByDepth: 'QD < 2.0'
Excess Missingness: 'NCC > 32' (>16 individuals with missing data)
More than 10% of reads have mapping quality zero: '(MQ0/(1.0*DP)) > 0.10'
Low mapping quality: 'MQ < 40.0'
```

samtools calling (multisample):

```
samtools mpileup -t DP,DPR,INFO/DPR -C50 -pm2 -F0.2 –ugf REFERENCE.fa SAMPLE1.bam SAMPLE2.bam ... | bcftools call -vmO z -f GQ -o samtools_VARIANTS.vcf.gz
```

The consensus GATK and samtools call set was obtained using the `GATK`:

```
java -Xmx10000m -jar GenomeAnalysisTK.jar -T SelectVariants -R REFERENCE.fa --variant onlyVariants_filtered.vcf.gz -o onlyVariants_filtered_concord.vcf.gz --concordance samtools_unfiltered.vcf.gz
```

BEAGLE genotype refinement (per scaffold) - specifying the trio relationships:

```
java -jar beagle.r1398.jar gl=onlyVariants_filtered_concord_sc${sc}.vcf.gz ped=../Malawi_trios.pedind nthreads=8 ibd=true ibdtrim=200 phase-its=8 impute-its=8 out=beagle_onlyVariants_filtered_concord_sc_${sc}
```

Accessible genome

The accessible genome was defined by masking every basepair of the genome assembly (including invariant sites) where any of the filters applied to the GATK dataset or where a GATK variant was absent in the samtools dataset. Excluding gaps, the reference assembly contains 713.6Mb of sequence. Our approach masked out just under 61Mb of sequence, resulting in an 'accessible genome' of 653Mb, or 91.5% of the reference sequence excluding gaps.

Shapeit phasing

To further improve the quality of haplotype phasing, we re-phased the BEAGLE output data for scaffolds larger than 1Mb (scaffolds 0 to 201) using the `shapeit v2.r790` haplotype phasing method [86] including the use of phase-informative reads[87].

The specific commands (with full trios included) were as follows:

```
extractPIRs --bam all_alignment_files_sc_${i}.txt --vcf onlyVariants_filtered_concord_sc${sc}.vcf.gz --out sc_${i}_PIRlist
shapeit -assemble --input-ped beagle_wTrios_sc${sc}.ped beagle_wTrios_sc${sc}.map --input-pir sc_${i}_PIRlist -O sc_${i}_phased_trio_ped --thread 4 --window 0.5
```

To check phasing accuracy, we also ran the above phasing process including only unrelated individuals. Then we compared switch error rates in the trio samples between the two datasets using the `vcftools v0.1.12b` option `--diff-switch-error` (Supplementary Fig. 25).

Variant calls for outgroup *Astatotilapia*

A separate variant call set was generated using short reads from 19 individuals from seven outgroup *Astatotilapia* species (Supplementary Table 2), 13 individuals representing the Lake Malawi eco-morphological groups (*C. afra*, *M. zebra*, *A. calliptera* from Salima and Lake Massoko, *O. lithobates*, *T. nigriventer*, *C. likomae*, *A. geoffreyi*, *L. gossei*, *D. limnothrissa*, *D. ngulube*, *R. longiceps*, and *R. woodi*) and the more distantly related Lake Tanganyika cichlid *N. brichardi.* The *N. brichardi* short read data (100bp reads, ~15x coverage) were downloaded from the NCBI Short Read Archive (Accession: SRR077327). Because of the greater divergence between all these samples (compared with the within-Malawi dataset), we choose to map the reads to the *Oreochromis niloticus* (Nile Tilapia; version Orenil1.1) genome assembly which is equally distant from all the samples. The technical details of mapping and variant calling were as for the within-Malawi dataset (see above), except that we did not generate samtools calls (only GATK haplotype caller).

Because of mapping to a more distant reference, we applied slightly more stringent filtering on the variant calls compared with the within-Malawi dataset. Filters were applied to the following GATK annotations:

```
Minimal inbreeding coefficient: 'InbreedingCoeff < -0.5'
Minimum overall read depth: 'DP < 550'
Maximum overall read depth: 'DP > 950'
Fisher's exact test to detect strand bias: 'FS > 40.0'
QualityByDepth: 'QD < 2.0'
Excess Missingness: 'NCC > 1' (any individuals with missing data)
More than 10% of reads have mapping quality zero: '(MQ0/(1.0*DP)) > 0.10'
Low mapping quality: 'MQ < 40.0'
Excess heterozygosity: 'ExcessHet >= 20'
```

In addition, we masked out all sites in the reference where any overlapping 50-mers (sub-sequences of length 50) could not 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`).

Finally, we removed multiallelic sites and indels and masked sites within +/- 3bp of indels.

As separate accessible genome mask was produced for this dataset (see above).

Whole-genome alignments

Pairwise alignments of cichlid genome assemblies listed in Supplementary Tables 5 and 6 were generated using `lastz v1.0` [115], with the following parameters:

For cichlid vs. cichlid alignments:

```
B=2 C=0 E=150 H=0 K=4500 L=3000 M=254 O=600 Q=human_chimp.v2.q T=2 Y=15000
```

For cichlid vs. other teleost alignments:

```
B=2 C=0 E=30 H=0 K=3000 L=3000 M=50 O=400 T=1 Y=9400
```

This was followed by using Jim Kent's `axtChain` tool with `-minScore=5000` for cichlid-cichlid and `-minScore=3000` for cichlid-other teleost alignments. Additional tools with default parameters were then used following the UCSC whole-genome alignment paradigm[90] (http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto) in order to obtain a contiguous pairwise alignment.

Details on selection of multispecies coalescent approaches

The advantages of the SNAPP package[36] are that it aims to estimate the species tree from unlinked biallelic SNPs (there is no need to identify regions free from ILS), and that it is a representative of the Bayesian framework. However, application of SNAPP to the full set of 70+ species and hundreds of thousands of biallelic markers was not computationally feasible.

SVDquartets[37,38] also uses SNPs and does not require identifying regions free from ILS. However, due to its algebraic approach, it is much faster than SNAPP, allowing it to handle the full dataset. SVDquartets first calculates matrices of site pattern frequencies for quartets of samples. The authors have shown that the expected rank of the matrix that corresponds to the correct quartet tree under ILS is 10, and use singular value decomposition to estimate which of the possible quartet matrices is closest to having rank 10. The best of the three four-taxon trees is chosen. After all quartets have been evaluated, the `PAUP*` version of the QFM algorithm[103] is used to search for the overall tree that minimizes the number of quartets that are inconsistent with it. Advantages of this approach include robustness to site specific rate variation[38] and to gene-flow between sister taxa[39]

Finally, `ASTRAL`[40] is a representative of yet another approach: a 'summary method' which attempts to reconstruct the species tree from a set of local trees (also referred to as 'gene trees', although the alignments on which they are based do not necessarily correspond to genes). We choose ASTRAL for its speed, being able to handle the full genome-wide dataset, and because of its accuracy in reconstructing trees under ILS[40,41].

The ABBA-BABA related *f* statistics

The *f* 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]. Assume we have four populations with given phylogeny (((A, B), C), O). If we denote:

$$S(A,B,C,O) = \sum_{loci} (1-\hat{p}_A)\,\hat{p}_B\hat{p}_C(1-\hat{p}_O) - \hat{p}_A(1-\hat{p}_B)\hat{p}_C(1-\hat{p}_O)$$

then an estimator of the *f*-statistic is given by:

$$\hat{f} = \frac{S(A,B,C,O)}{S(A,C_1,C_2,O)}$$

where C_1 and C_2 designate two subsamples of population C. In a scenario where genetic material introgressed from population C into population B, $\hat{f}$ estimates the genomic proportion in B tracing its ancestry through the introgression event from C. Note that in theory C_1 should be sampled from the population ancestral to C that introgressed into the ancestral population of B. However, in practice we use random subsamples of our sample for C.

Neolamprologus brichardi was used as the outgroup O for the reported calculations, however, using *Pundamilia nyererei* instead yielded very similar results. We computed the *f* statistic for all trios ((A, B), C) in the ASTRAL* tree (Supplementary Fig. 7). Statistical significance was assessed by computing block-jackknifes on windows of 60k SNPs in the original VCF. Multiple testing was accounted for by computing Bonferroni-Holm family-wise error rate (FWER).

The branch-specific $f_b(C)$ statistic

f (or D) statistics calculated for different sets of species are not independent as soon as they share drift (that is, internal or external branches) (Supplementary Fig. 26). At the same time, these correlations can be informative about the timing of introgression, i.e. about which (internal) branch was subject to ancestral structure or gene flow (Supplementary Fig. 26A,B). To obtain a set of less correlated f statistics, we calculate for each branch b

$$f_b(C) = median_A\left[min_B[f(A,B,C,O)]\right]$$

where B runs over all clades that are descendants of b, and A over all clades that are descendants of $b's$ sister branch a (Supplementary Fig. 26A). Thus $f_b(C)$ measures the relative excess of allele sharing between the descendants of branch b with the clade C, compared with allele sharing of the descendants of branch a with the clade C. Note that negative f scores, i.e. those where A and C share excess alleles relative to B, are set to zero in this calculation (but are accounted for in the score of sister branch a).

The idea behind taking the minimum across B is that for any B, a descendent of the branch b, $f(A,B,C)$ measures excess allele sharing due to introgression (or ancestral structure) affecting b (Supplementary Fig. 26B), but introgression events that affect a branch b' descendent of b would only be picked up by $f(A,B,C)$ if B is descendent of b', but not if B is not descendent of b' (Supplementary Fig. 26C). Hence, taking the minimum of $f(A,B,C,O)$ across B gives a conservative estimate of the amount of excess allele sharing of b with C (relative to A) that is attributable to events involving b (rather than its descendent branches).

The rationale for taking the median across descendants of b's sister branch a rather than the minimum is that additional introgression events from C (or a relative) into the descendent branches of a can dilute the signal of introgression from C into b since such events decrease $f(A,B,C,O)$. Conversely, introgression of an outgroup of A, B, C into a A would increase $f(A,B,C,O)$. Hence, by taking the median across A we reduce the chances of $f_b(C)$ being confounded by events on the branches between a and A. We note however, that there remains a general identifiability problem between events involving a or b. Results are qualitatively similar if the minimum is taken across A rather than the median.

Finally, we would like to point out that while the $f_b(C)$ statistic accounts for the inter-dependence of f statistics for different A and B, the same introgression event can still lead to correlated signals for different C, either because introgression involved the common ancestor of different C, or because of the fact that related C show correlated allele frequencies. We do not see a straightforward way to correct for these correlations, but suggest that a careful visual inspection of the data (Fig. 3) allows one to draw conclusions on likely introgression patterns.

As with the f scores, we also summarize block-jackknifing Z scores to get a branch-specific measure of significance

$$Z_b(C) = median_A\left[min_B[Z(A,B,C,O)]\right].$$

Supplementary Note

I. Sample selection

Our sample selection provides broad coverage of all the major lineages of the rapidly radiating cichlid tribe Haplochromini in Lake Malawi, with specimens from:

1. Eight species from the 'mbuna' group of mainly shallow-water rock-dwelling cichlids. Our specimens represent much of the diversity of this group, covering 6 out of 10 genera defined by Ribbink *et al.* in their detailed classification of 196 Malawian mbuna species [116].
2. Nine species of 'deep water' benthic haplochromines: many of these species are found at depths >50m, a 'twilight' zone with very little visible light; a few inhabit shallower water but are often crepuscular feeders, residing among rocks by day. These include members of the genera *Alticorpus* and *Aulonocara* (characterized by greatly enlarged sensory openings of their heads and lateral lines) and several of the species currently assigned to the genus *Lethrinops*.
3. Forty-one species of cichlids found predominantly in shallow waters close to the shore (like 'mbuna'), but on sandy or muddy lake floor and the transition zones between sandy and rocky habitats. This is a very diverse group of cichlids with hundreds of described species[29], including for example large (over 35cm) predators such as *Buccochromis nototaenia*, the invertebrate picker *Placidochromis johstoni*, molluscivores such as *Chilotilapia rhoadesii* and *Mylochromis anaphyrmus*, and cichlids that filter the sandy sediment like *Ctenopharynx nitidus*. We refer to this group collectively as 'shallow benthics'.
4. Four species of zooplankton-feeding, shoaling cichlids which are commonly referred to as 'utaka'. These species feed in the water column, but their distribution is limited to locations close to the shore[29]. We use the term utaka *sensu* Bertram et al.[117] to refer to the species that were later assigned to the genus *Copadichromis*[50] (see also discussion in ref. 29, p. 120).
5. Three out of eight described species from the genus *Rhamphochromis*[50] - large pelagic (open-water) piscivores, feeding mainly on lake sardines[29]. Cichlids are primarily bottom-dwellers, but members of this group [and *Diplotaxodon* (see below)] have undergone extensive changes in morphology and behavior to invade the substantial pelagic habitat in Lake Malawi.
6. Three out of seven scientifically described species of *Diplotaxodon* (ref. [118], p. 198), and three undescribed species: *D. macrops* 'black dorsal'[119], *D. macrops* 'ngolube' (ref. [118] p. 239), and *D.* 'white back similis' (a new undescribed species). Like *Rhamphochromis*, *Diplotaxodon* are pelagic cichlids, but are typically found at greater depths (>50m) and their diet consists mainly of zooplankton, rather than fish. We include in this group also the species *Pallidochromis tokolosh,* which is morphologically intermediate between the two genera and is a slightly more benthic form (ref. [118], pp. 198-199), but its genetic affinities are more to the genus *Diplotaxodon*, as we show in this manuscript.
7. *Astatotilapia calliptera* - one of only two haplochromine cichlids found in Lake Malawi able to cross the lake-river barrier. *A. calliptera* is a versatile, relatively small cichlid (~10-15cm), common in the rivers throughout the Lake Malawi catchment. In Lake Malawi, *A. calliptera* frequents shallow sheltered bays with muddy sediment and aquatic plants, often feeding on snails (ref. 29, p. 281). It has been suggested that it may be

related to the ancestral lake-river generalist species that seeded most or perhaps all of the Lake Malawi haplochromine radiation[29]. Therefore, we sampled *A. calliptera* genetic variation extensively, including 16 specimens (four from Lake Malawi itself and twelve more from the broader Lake Malawi catchment).

The other species able to cross the lake-river barrier is *Serranochromis robustus*: a large predator often seen in very shallow water near river estuaries (ref. 29, p. 277) and common in rivers to the south-west of Lake Malawi, including the Zambezi system[120]. However, *S. robustus* is not a part of the Lake Malawi haplochromine radiation, but instead belongs to a distant group sometimes referred to as the 'Congo clade'[121]. For a list of all samples and details of our assignment of Lake Malawi species into these seven lineages see Supplementary Table 1.

II. Simulations and treemix results

We tested the behaviour of our 'tree violation' inference methods (the f_b statistic and our approach to calculating the residuals of neighbour-joining trees) on data simulated using the software `ms` [122]. Four models were evaluated, with parameters compatible with those inferred for the Lake Malawi cichlid radiation, as depicted in Supplementary Fig. 27. They include a phylogeny without gene flow, and three independent simulations each with a directional gene flow event between two non-terminal branches, where the strength of the gene flow was 10%, 30%, and 50%.

First, we estimated species trees from the simulated data using the Neighbour-Joining (NJ) algorithm. The inference is very accurate in the absence of gene flow (Supplementary Fig. 28A) or if gene flow is relatively weak (f = 10%, Supplementary Fig. 28B). However, for simulations with more substantial gene flow (f = 30%, and f = 50%) the inferred trees do not correspond to either of the two alternative topologies (i.e. neither the standard one nor the one following the gene flow event). Instead, the inferred topologies are 'intermediate', with details dependent on the strength of the gene flow (Supplementary Fig. 28C,D).

Calculating residuals between the pairwise genetic distance matrix and the inferred NJ tree can help to capture inconsistencies of the bifurcating tree model (Supplementary Fig. 29, cf. Supplementary Fig. 12). However, the magnitude of the residuals does not increase linearly with the strength of gene flow. This is because the inferred tree topology changes with increasing gene flow in a way that keeps residuals lower than they would be if the original topology were enforced. This effect may be weaker in a larger tree which has more constraints.

Next, we computed all 56 f tests consistent with the inferred NJ tree for each model. The number of significant f statistics were 0, 24, 28, and 28, for f = 0%, 10%, 30%, and 50%, respectively (FWER<0.001). Calculating f_b on the simulated data, yielded 0, 7, 8, and 10, significant f_b, respectively (Supplementary Fig. 30). Thus it is clear that while using the f_b statistic reduces the correlation between different tests, the correlations are not removed entirely. A single gene flow event can still produce multiple significant f_b values. However, we suggest that an additional benefit of the f_b statistic is that it is branch specific and can be intuitively interpreted as identifying "problematic" branches in the phylogeny, in the sense that the phylogenetic model is violated at these branches. For example, Supplementary Fig. 30C shows that the ancestral branch of the species A1 and A2 shows excess allele sharing with the clade containing C1, C2, D1, and D2, and most strongly so with C1 and C2. At the same time, B1 show excess allele sharing with

A1 and A2, compared to B2. While these observations do not correspond to true gene-flow events, they are expected given the difference between the true relatedness and the inferred tree, and thus hint at the "correct" species relationships.

Note that the inferred magnitudes of f and f_b are substantially lower than the actual gene flow proportion. We suggest that this is due to drift on the terminal branches that leads to a decrease of excess allele sharing between A and C. A possible explanation for the large values of f observed in the real data is the occurrence of gene flow over an extended period.

Finally, we applied the software `treemix`[52] `v1.12`, which is widely used to infer gene-flow events, to the simulated data (Supplementary Figs. 31-33), with the parameter `-k` 5000 - i.e. using blocks of 5000 SNPs. The results were plotted using the R function `plot_tree` which is included with `treemix`. We found that `treemix` was able to correctly infer the gene flow edge in the model with relatively weak gene flow (f = 10%) when the parameter `m` is set to one (see Supplementary Fig. 31C; the parameter `m` indicates to `treemix` how many gene flow events to expect and must be supplied by the user). However, increasing the strength of gene flow in the simulated models or increasing the `m` parameter value leads to unexpected results. First, when the `m` parameter value is increased, the correct event is still inferred, but higher migration weights are given to other, incorrect migration events (Supplementary Figs. 32C, 33C). Second, as soon as gene flow is strong enough that the maximum likelihood tree computed by treemix does not correspond to either of the alternative topologies (as in the simulations with f = 30% and 50%), `treemix` infers gene flow edges which do not seem to bear any connection to the true event. In particular, in such cases `treemix` tends to infer strong gene flow into the outgroup.

In conclusion, we suggest that while the number of significant f_b should not be interpreted as the number of independent gene flow events, f_b correctly tags problematic branches and shows excess allele sharing, even in cases of strong gene flow where the inferred phylogeny is likely wrong, while `treemix` seems to give misleading results in such cases.

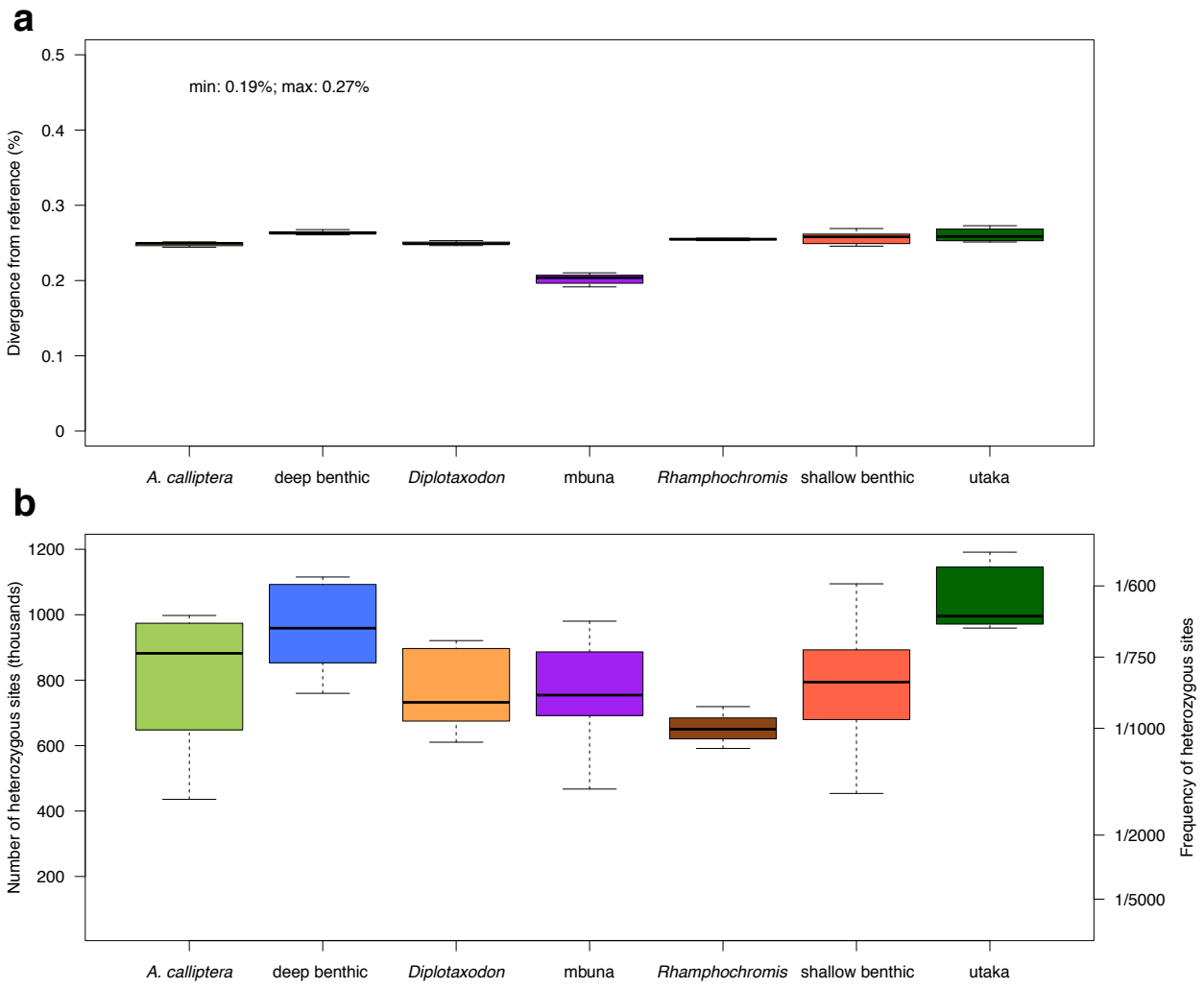
III. Gene selection and cichlid-specific genes

This candidate genes identified by the Δ_{N-S} are highly enriched for genes without homologs in either 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 showed significant differences ($p < 2.2\times10^{-16}$, Mann-Whitney test; Supplementary Fig. 34), with both a substantial component of genes with low $\bar{p}_N$, reflecting purifying selection, and also an excess of genes with high $\bar{p}_N$ (Supplementary Fig. 35).

IV. Non-green cone opsins are not involved in shared depth adaptation

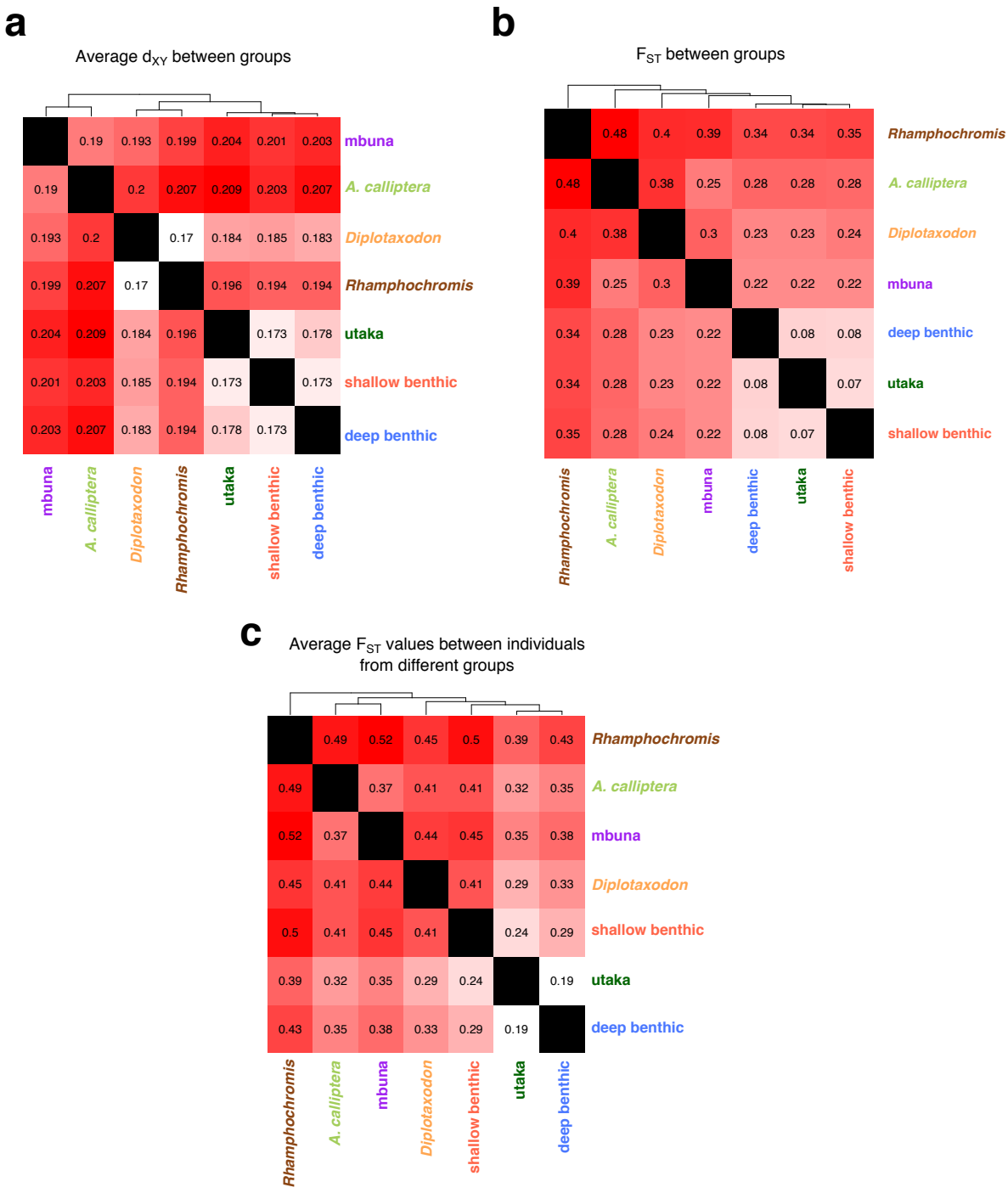
Among the non-green cone opsins, long wavelength, red-sensitive opsin (LWS) has been shown to play a role in speciation along a depth gradient in Lake Victoria[123]. While it did not receive high similarity or shared diversity scores in *Diplotaxodon* and deep benthics, it may still play some role in their depth adaptation. *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 (Supplementary Fig. 21). The short-wavelength opsin SWS1 is among the genes with high $\Delta_{N\text{-}S}$ scores but it does not exhibit any 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 Δ_{N-S} scores in our Lake Malawi dataset and thus are not among the candidate genes.



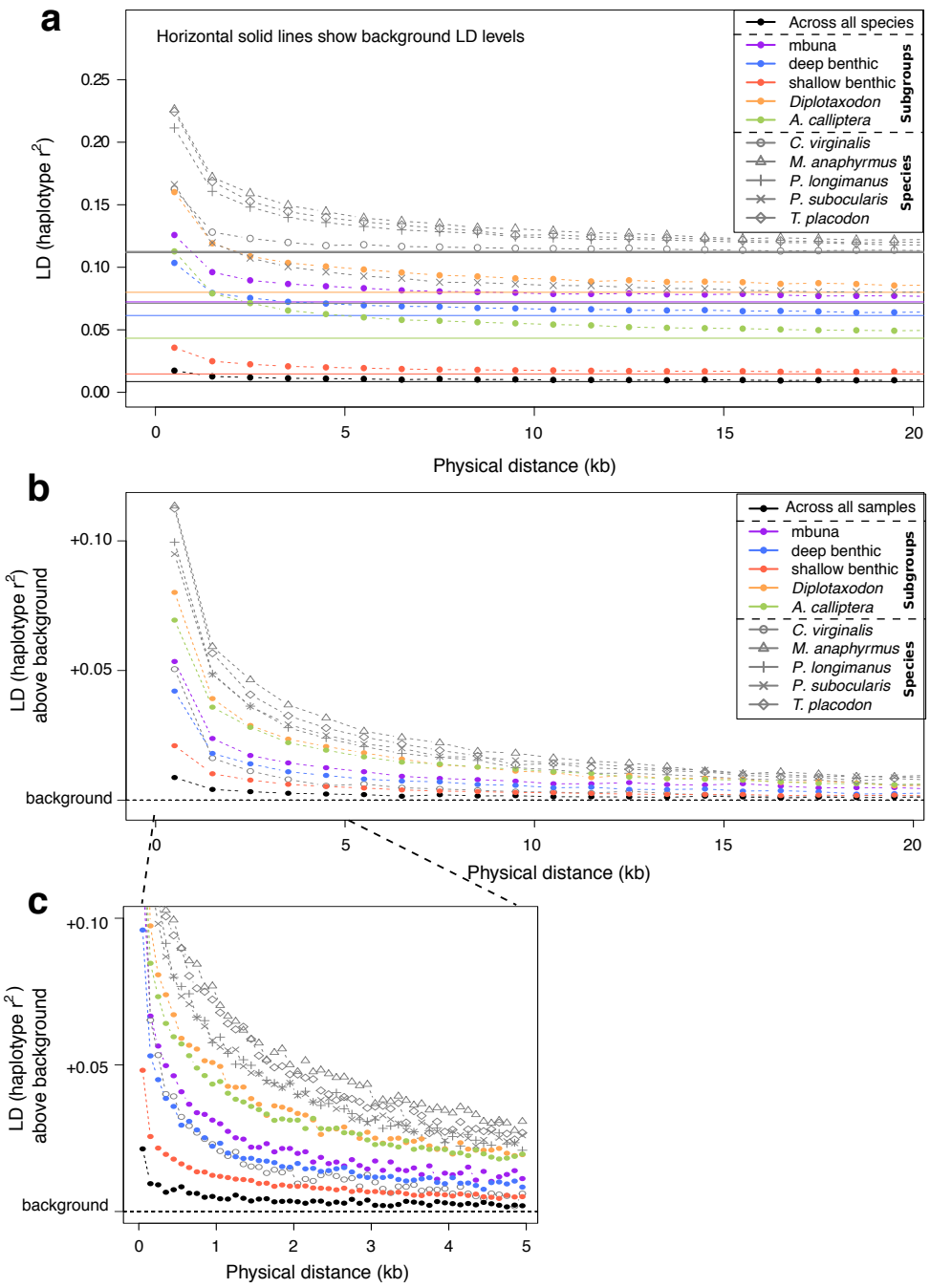
Supplementary Figure 1

Genome-wide variant statistics summarized per cichlid lineage. a, Variants called against *M. zebra* genome in Lake Malawi samples. **b,** The number and frequency (per bp) of heterozygous sites.



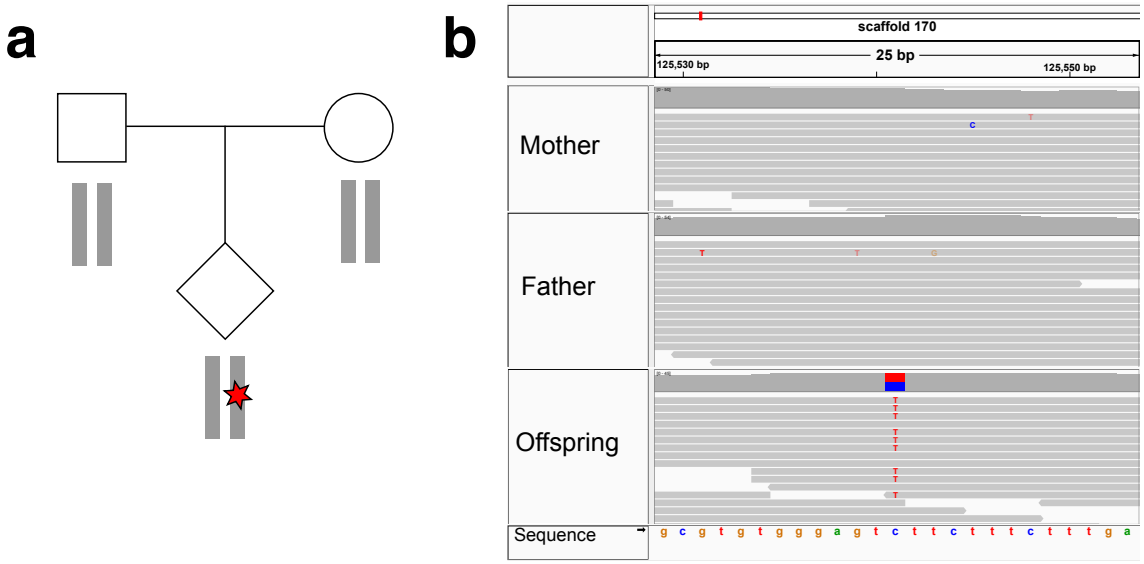
Supplementary Figure 2

Genetic divergence between the seven major eco-morphological groups. a, Average sequence divergence (d_{XY}). **b,** Between group F_{ST}. **c,** The values in **b** depend strongly on the proportion of sampled genetic variation within the groups. Therefore, we also calculated the average F_{ST} between individuals, which should be much less dependent on the particular set of species sampled from each group. See Methods for details of F_{ST} calculations.



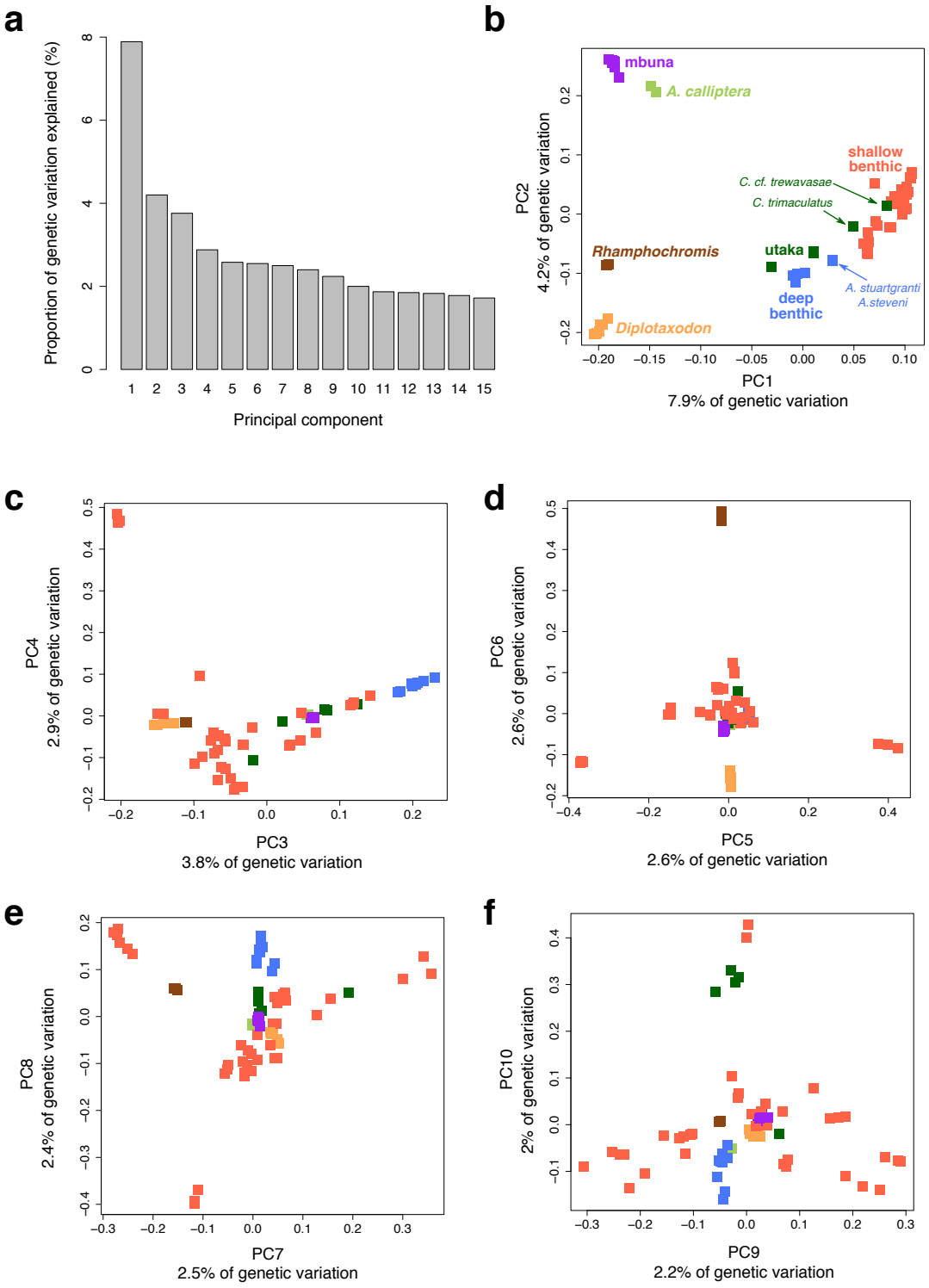
Supplementary Figure 3

Mean linkage disequilibrium (LD) decay. Three levels of detail are combined in the plots. across the whole radiation mean LD decays within a few hundred base-pairs. Within broad eco-morphological groups LD decays in a few kilobases (kb), and within-species it extends beyond 10kb. Only species and groups for which we had more than six individuals to calculate LD are shown (*Rhamphochromis* and utaka are therefore missing). **a,** Average haplotype r^2 values in 1kb windows over 20kb. **b,** As in panel a but with background LD (LD between variants mapping to different chromosomes) subtracted. **c,** Subtracted LD in 100bp widows over the first 5kb.



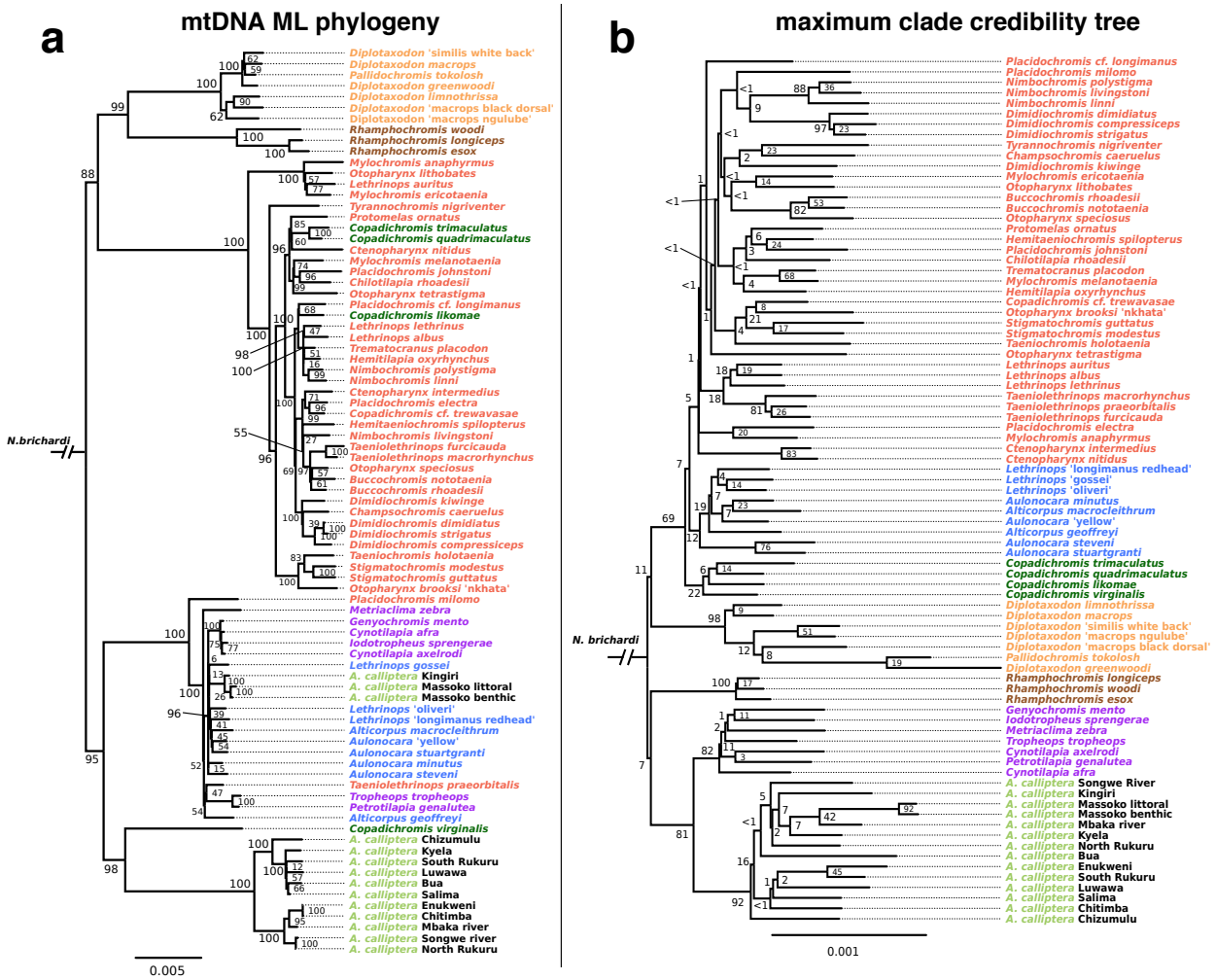
Supplementary Figure 4

Direct *de novo* mutation rate estimation from parent-offspring trios. a, Illustrating the approach: in each trio we looked for mutations in the child that were not present in either of the parents. **b,** A genome browser screenshot with an example of *de novo* mutation in *Lethrinops lethrinus* offspring on scaffold 170.



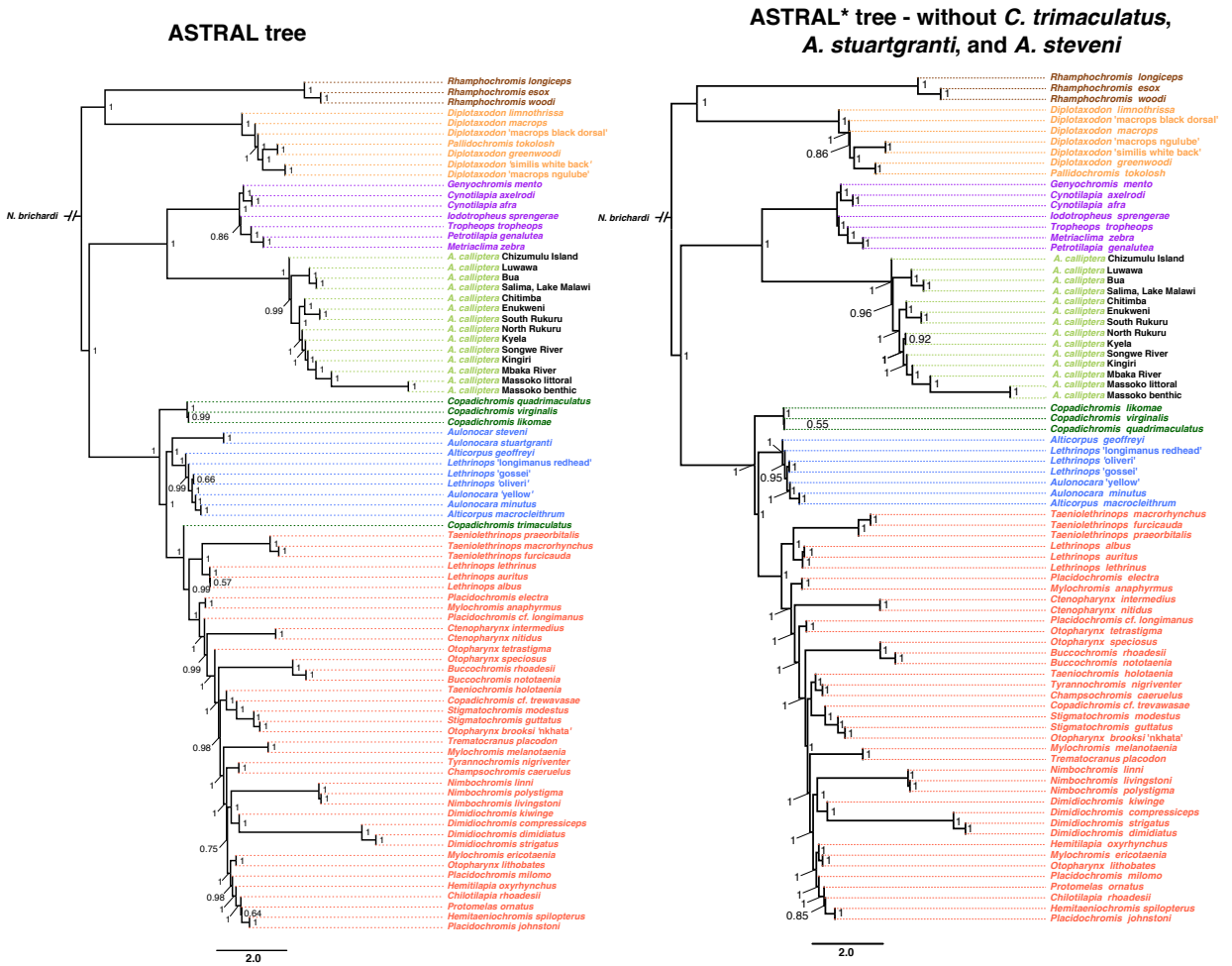
Supplementary Figure 5

Principal component analysis. a, The proportion of genetic variance explained by each of the first 15 principal components (PCs). **b-f)** The positions of individual samples along the first 10 PCs. Colours correspond to ecomorphological group assignment as indicated in panel **b**.



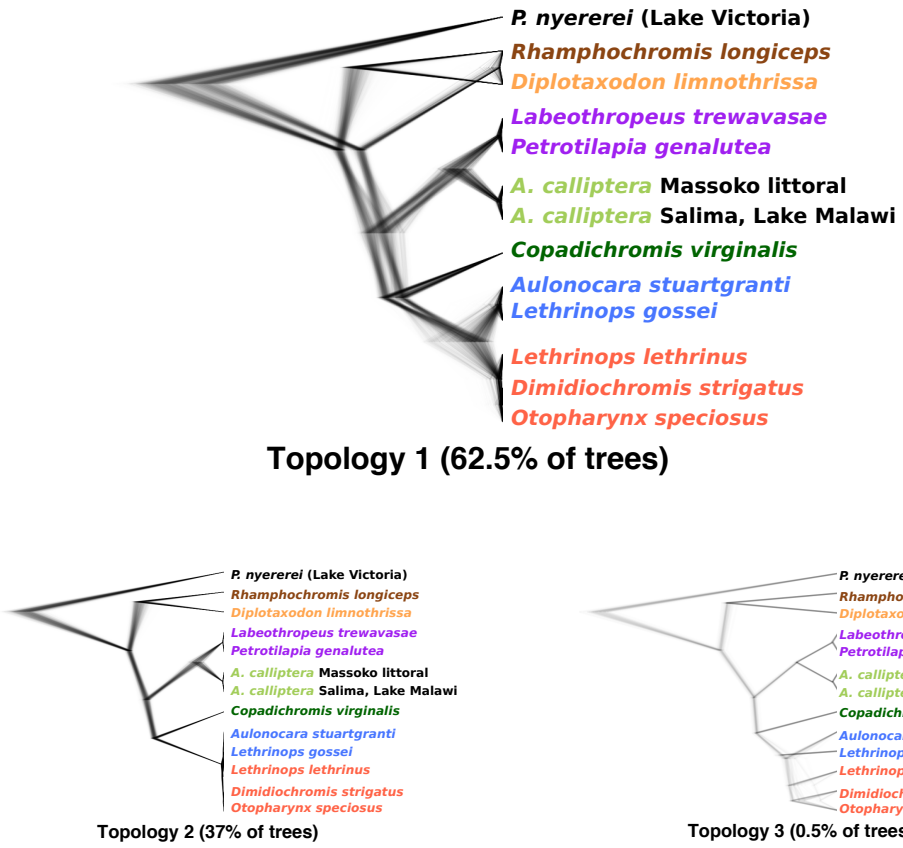
Supplementary Figure 6

mtDNA and maximum clade credibility (MCC) phylogenies. Scales are given in number of SNPs per bp. **a,** A maximum likelihood (ML) phylogeny based on the full mtDNA sequence. Node labels show bootstrap support based on 200 replicates. **b,** An MCC summary of 2543 ML trees based on non-overlapping genomic windows. Node labels show the percent prevalence of each clade.



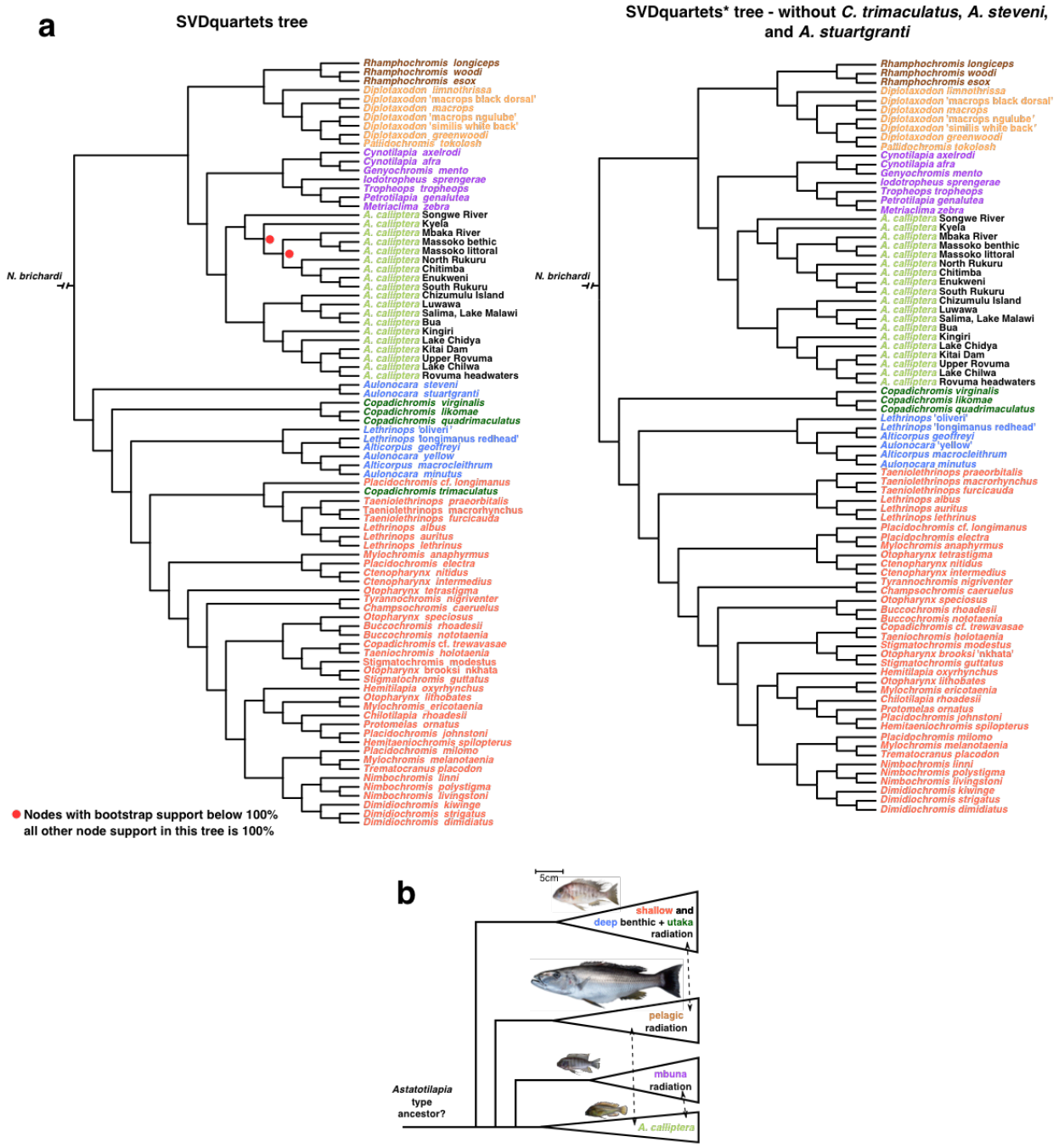
Supplementary Figure 7

Species tree inference under the multispecies coalescent - ASTRAL trees. Support values are local posterior probabilities. Internal branch lengths are inferred by ASTRAL in coalescent units. Terminal branch lengths cannot be inferred with only one individual per species. *A. calliptera* populations were analysed as separate species. In the ASTRAL* tree, the individuals that are genetically intermediate between eco-morphological groups have been removed



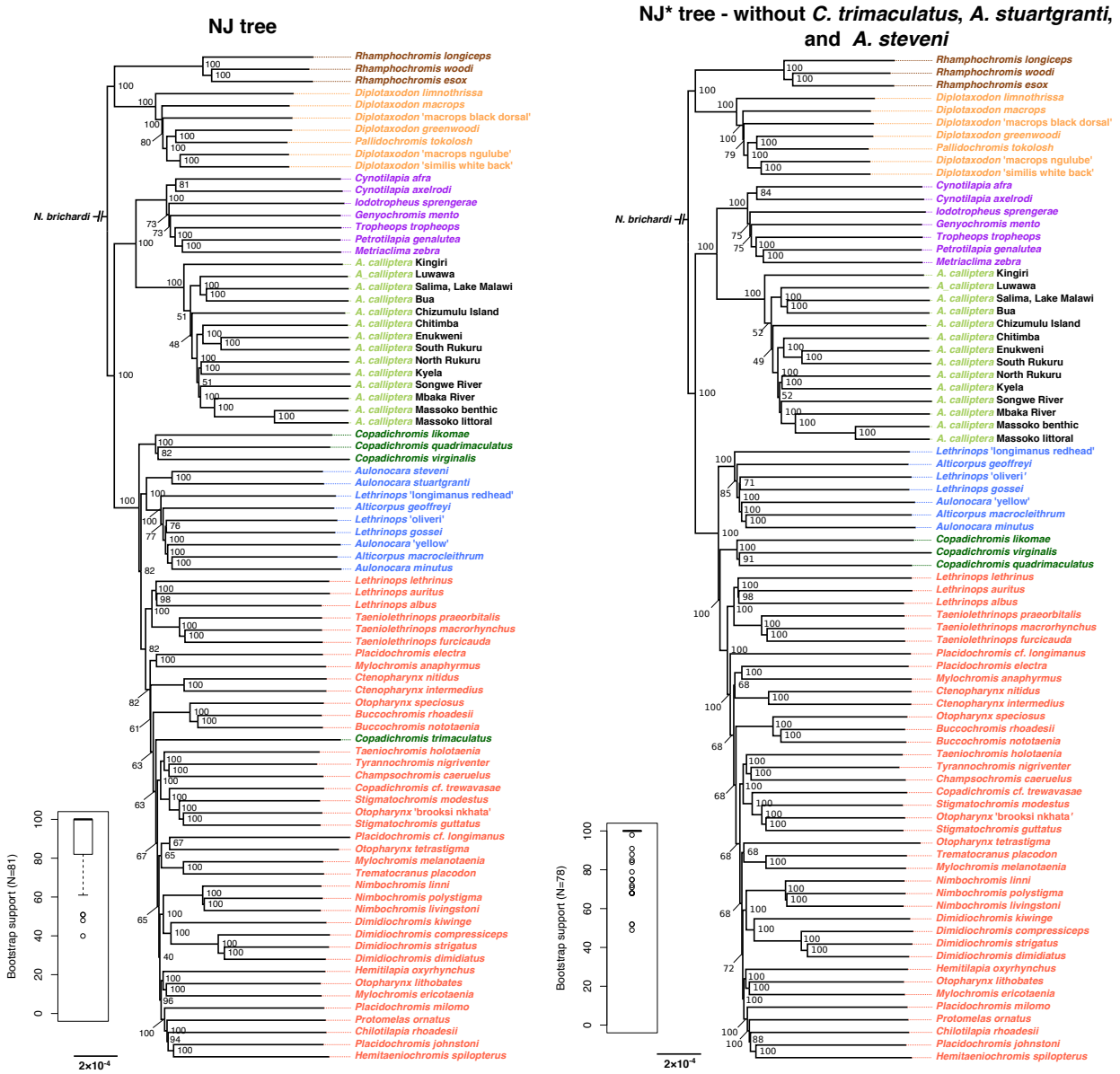
Supplementary Figure 8

Species tree inference under the multispecies coalescent - SNAPP trees. 30617 SNAPP trees sampled from three MCMC inference chains at every 500 steps. This inference includes 12 Lake Malawi species representing the major lineages (plus the Lake Victoria species *Pundamilia nyererei* as root), and is based on a random subsample of 48922 unlinked SNPs. Three distinct topologies were observed among the sampled trees. Topology 2 differs from topology 1 in that *Aulonocara stuartgranti* and *Lethrinops gossei* no longer cluster together (instead the *Aulonocara* clusters with *L. lethrinus*) ; the branch lengths in the whole benthic group (deep + shallow) are extremely short. Topology 3 differs from topology 1 in that *Lethrinops gossei* is basal to the remainder of the benthic group.



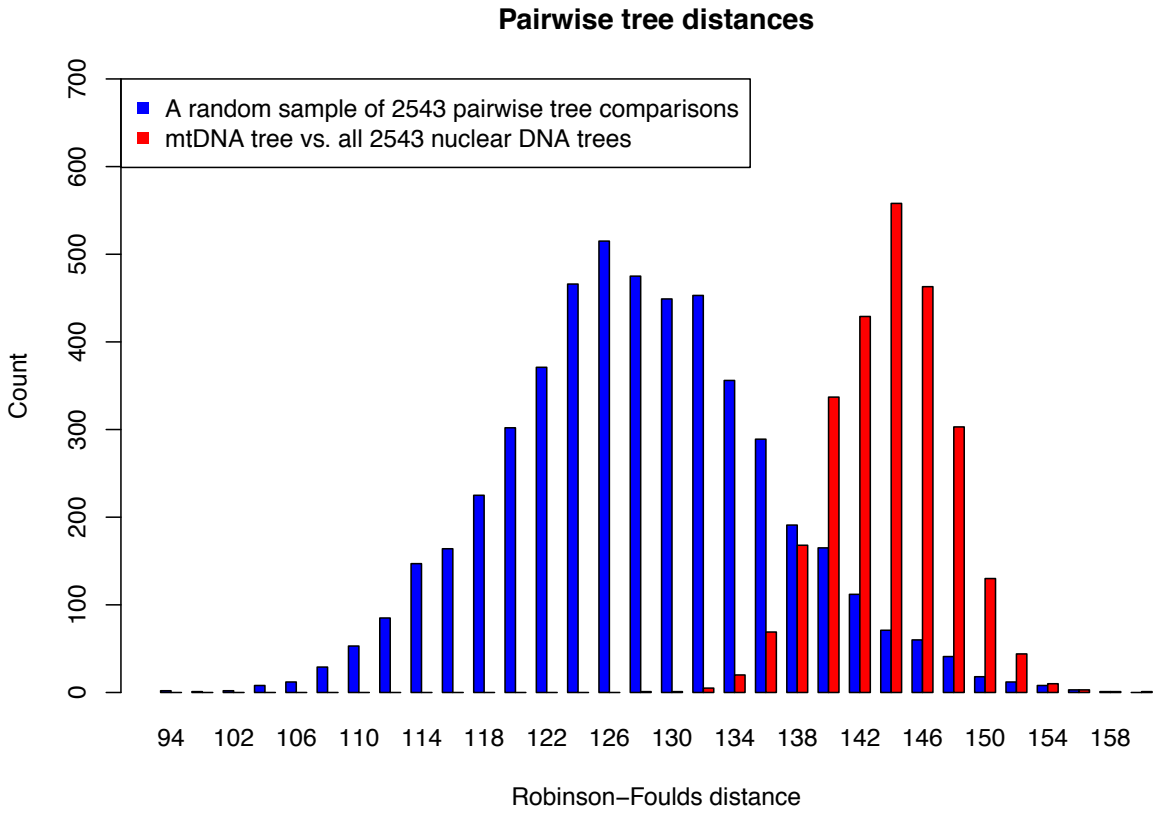
Supplementary Figure 9

Species tree inference under the multispecies coalescent - SVDquartet trees. a, The inferred species tree based on all SNPs. The branch lengths do not have any meaning as SVDquartets infers only tree topology. In the SVDquartets* tree, the individuals that are genetically intermediate between eco-morphological groups have been removed. **b,** Implications of the inferred branching order between the major groups for the model presented in Fig. 4f. The SVDquartets tree suggests that the benthic + utaka lineages branched off first, then the pelagic lineages, and finally the mbuna.



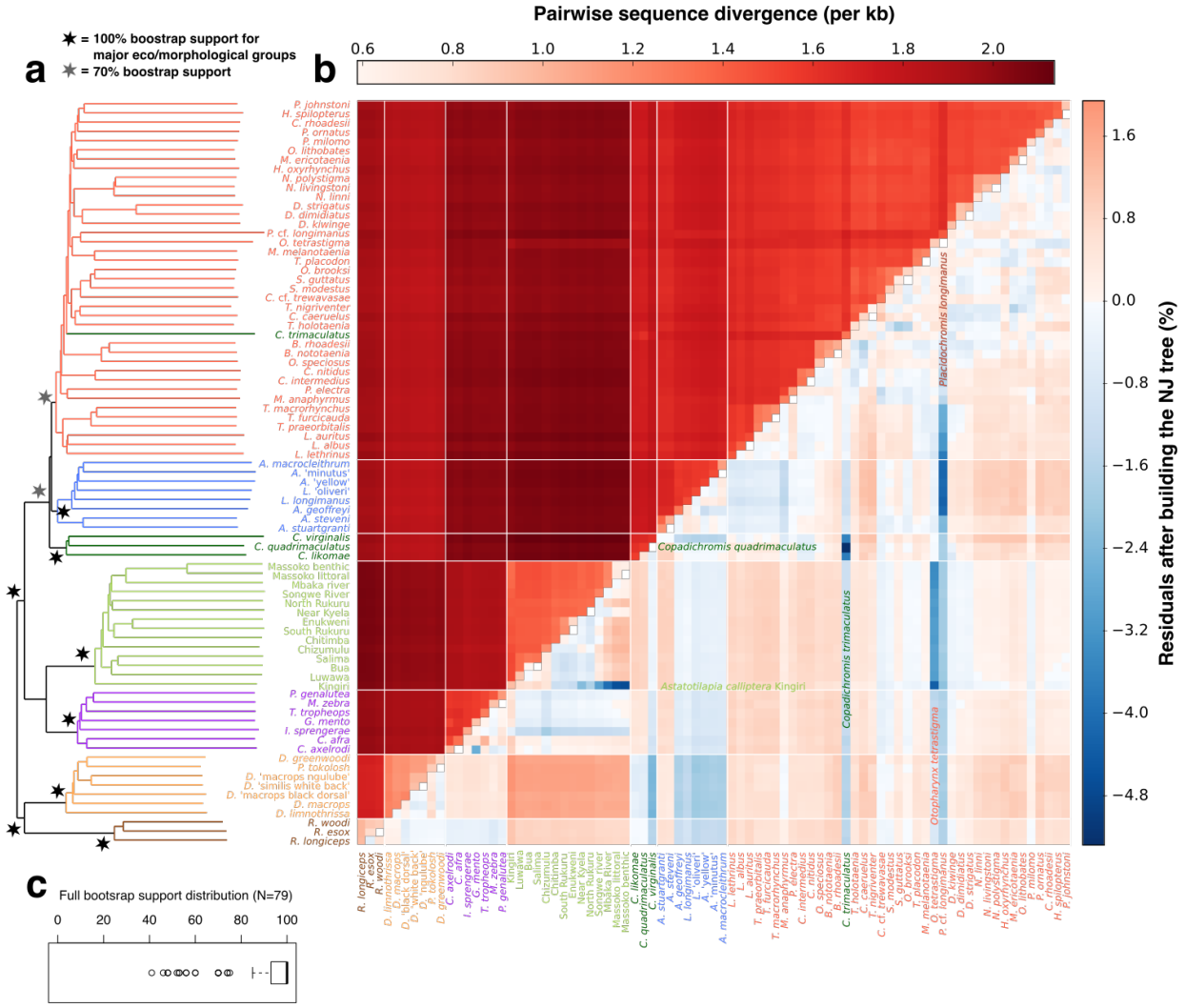
Supplementary Figure 10

Neighbour-joining trees based on pairwise differences. Long terminal long terminal branches reflect the high ratio of within-species to between-species variation. Scale given in substitutions per bp. In the NJ* tree, the individuals that are genetically intermediate between eco-morphological groups have been removed.



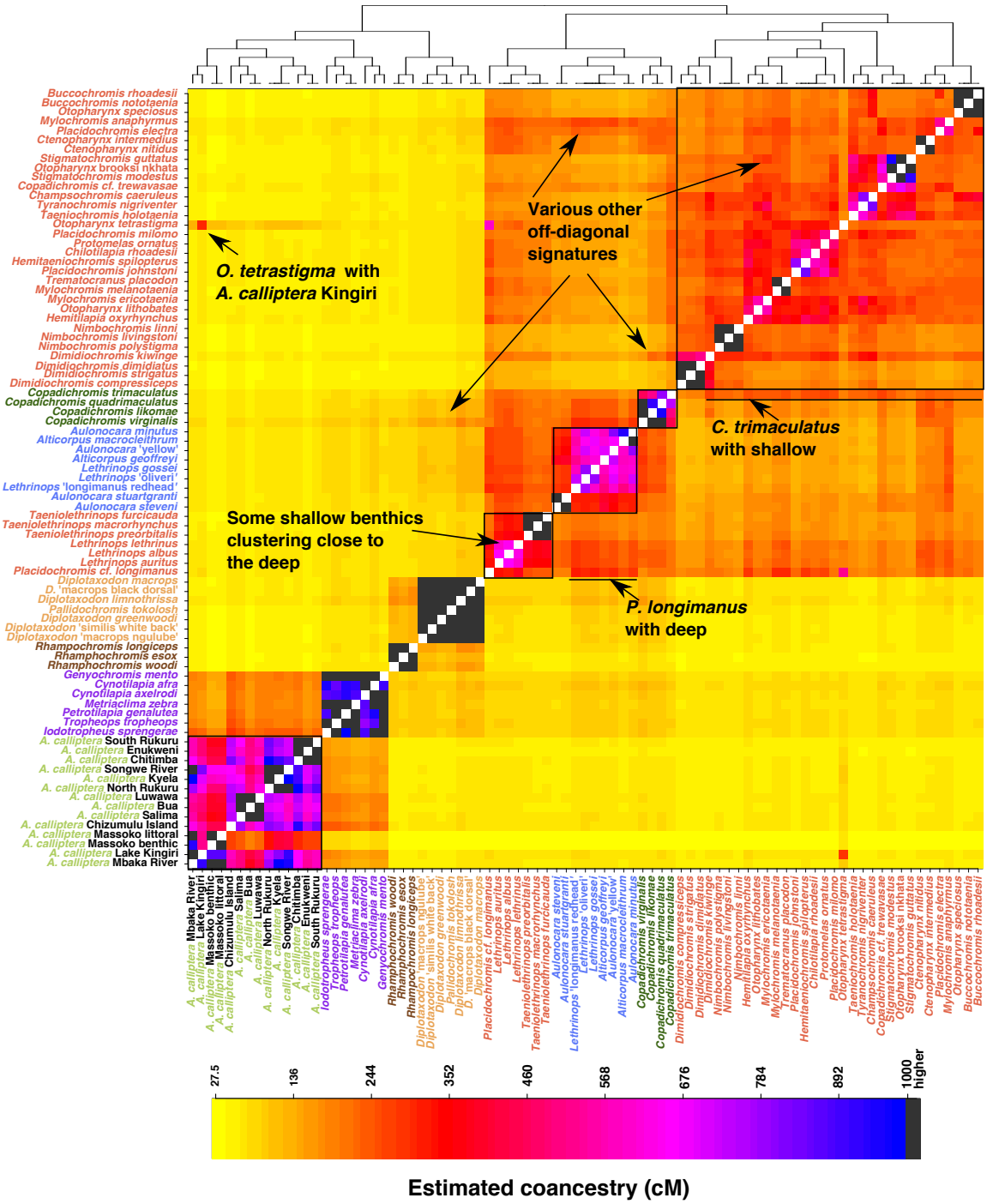
Supplementary Figure 11

Pairwise tree distances. Differences in topology (as measured by the Robinson-Foulds distance) tend to be much greater between the mtDNA tree and the local ML phylogenies based on genomic windows in nuclear DNA than when two random genomic windows are compared against each other.



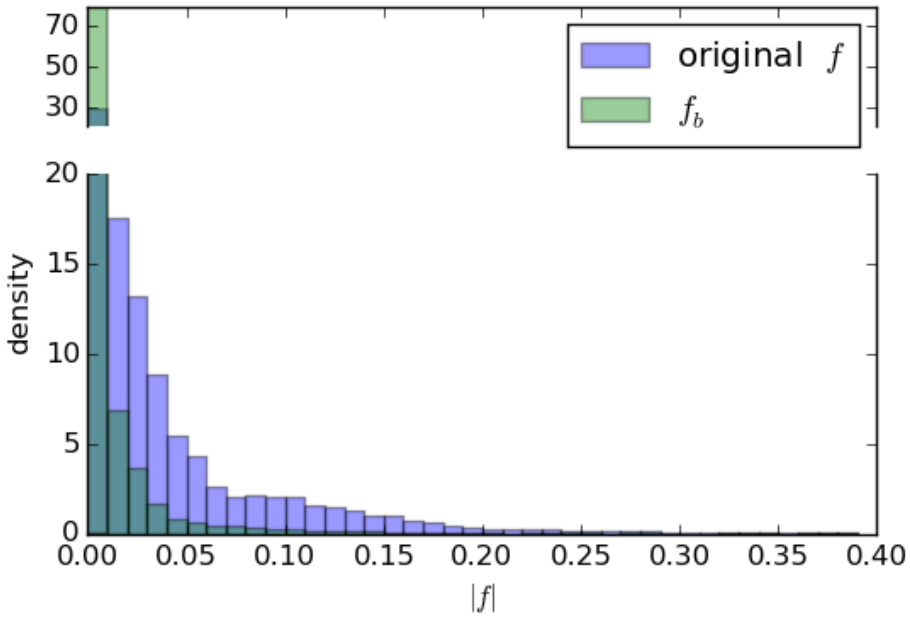
Supplementary Figure 12

Tree residuals provide additional insights. a, An NJ tree as in Supplementary Fig. 9a. **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.



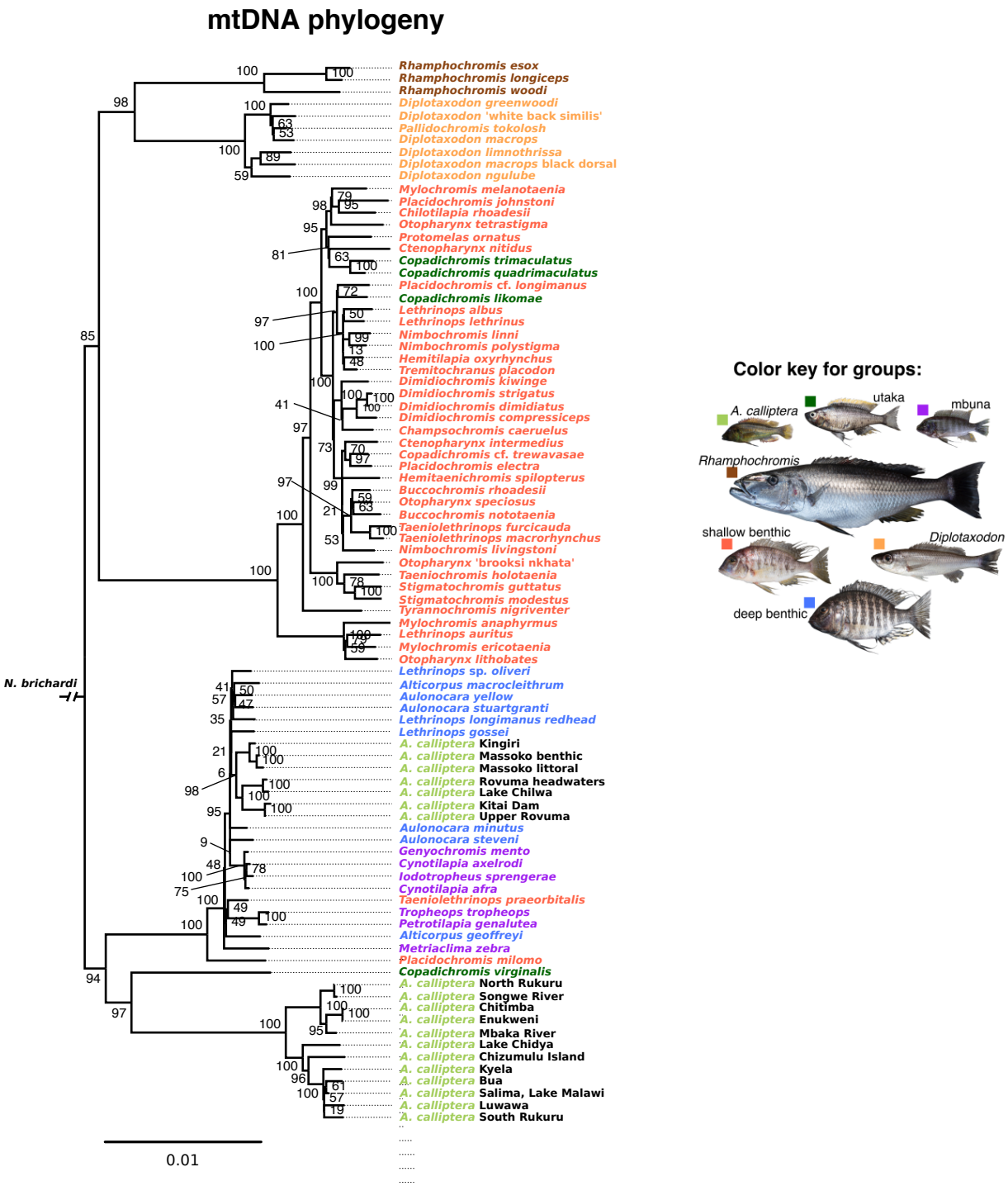
Supplementary Figure 13

Co-ancestry between Lake Malawi species measured by the `Chromopainter` software. Each row corresponds to a 'recipient' and each column to a 'donor'. Thus the values indicate the total length (in cM) along which a 'donor' haplotype was inferred to be the closest relative of a 'recipient' haplotype. Clusters corresponding to the major morphological groups are highlighted by rectangles (species name colours as in Fig. 1). A number of interesting cases of excess 'nearest neighbour' haplotype sharing are indicated by arrows, including those discussed in the main text and some additional examples that stand out.



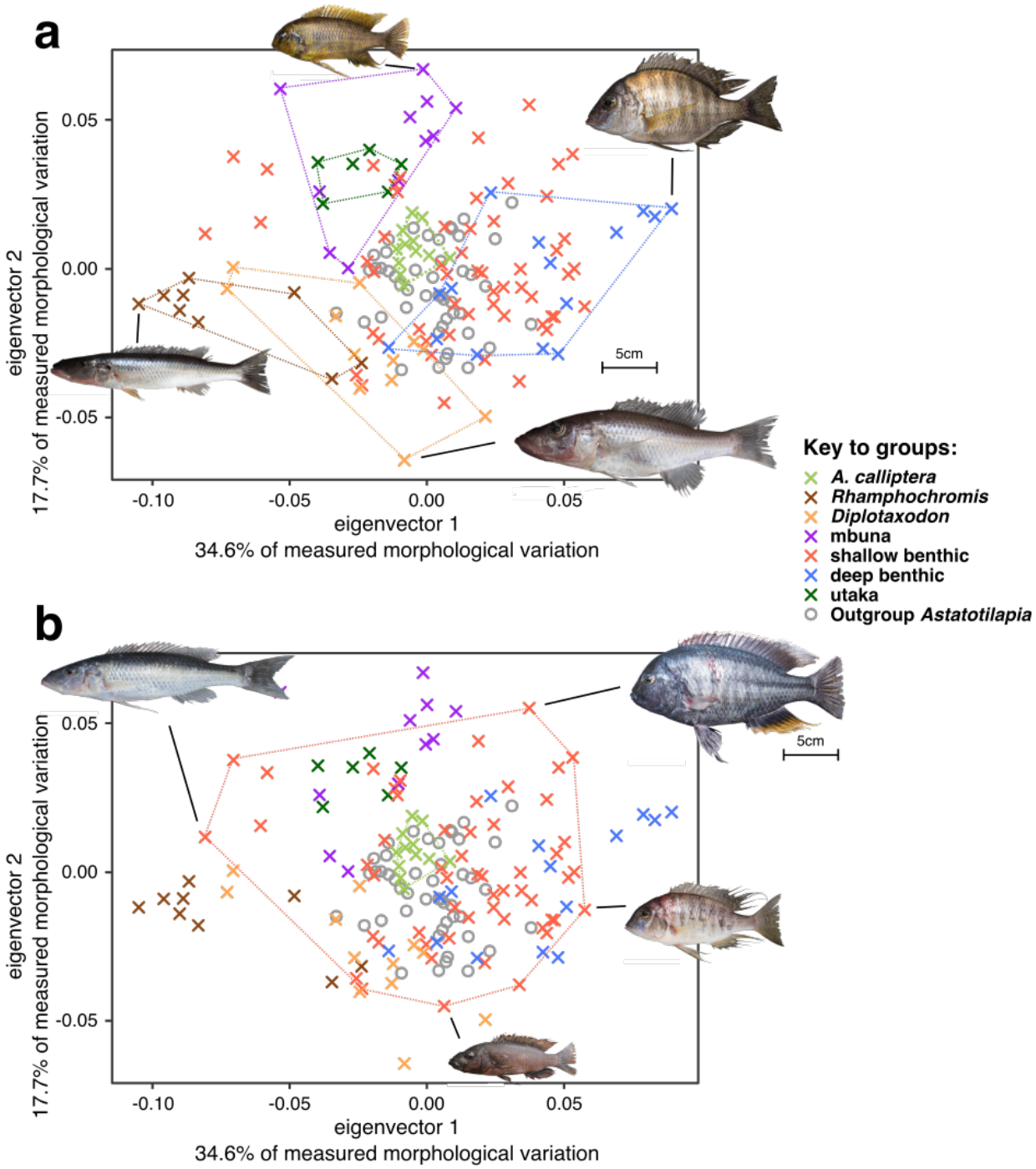
Supplementary Figure 14

Overlaid density histograms of original f scores (170640 tests), and the reduced branch-specific scores f_b (85311 tests).



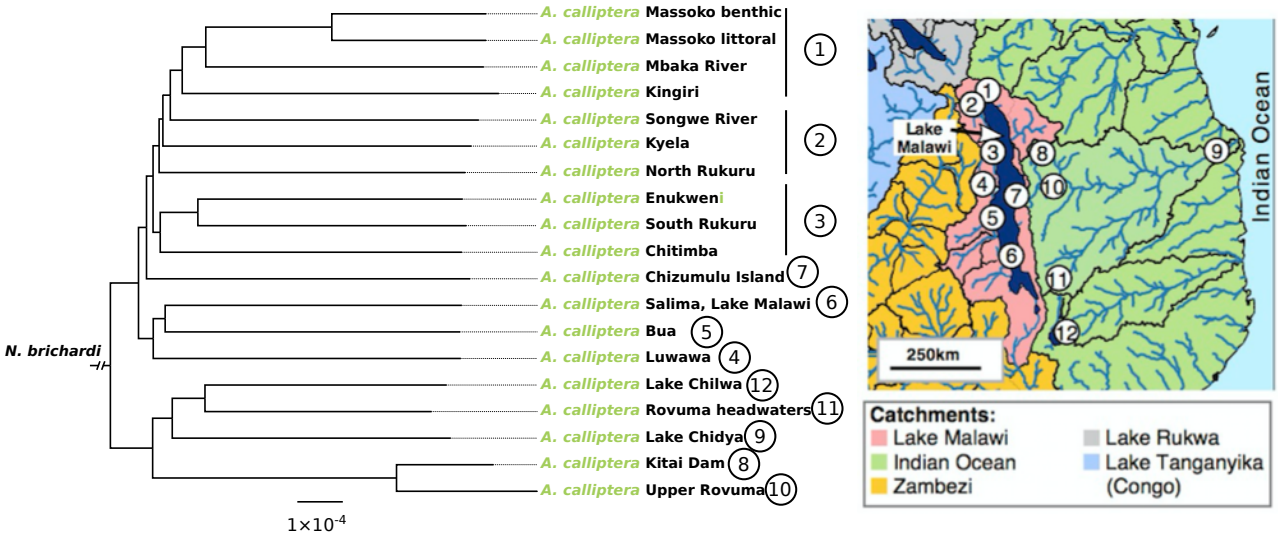
Supplementary Figure 15

mtDNA phylogeny with Indian Ocean catchment A. calliptera included. Scale is given in number of SNPs per bp. Node labels show bootstrap support based on 200 replicates.



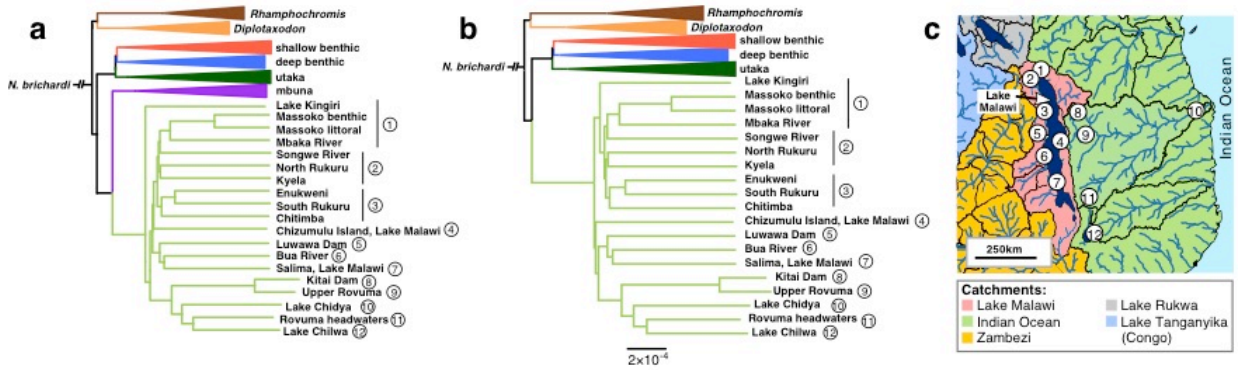
Supplementary Figure 16

Morphological diversity in sequenced Lake Malawi cichlids. PCA of body shape variation obtained from geometric morphometric analysis. **a,** *A. calliptera* individuals occupy a central position along the first two eigenvectors, and there is very little overlap and there is no overlap other Malawi groups except for shallow benthic. **b,** Although the extremes of the morphospace are occupied by individuals from other groups, it is notable that the body shape diversity within the shallow benthic group alone approaches the diversity of the full radiation.



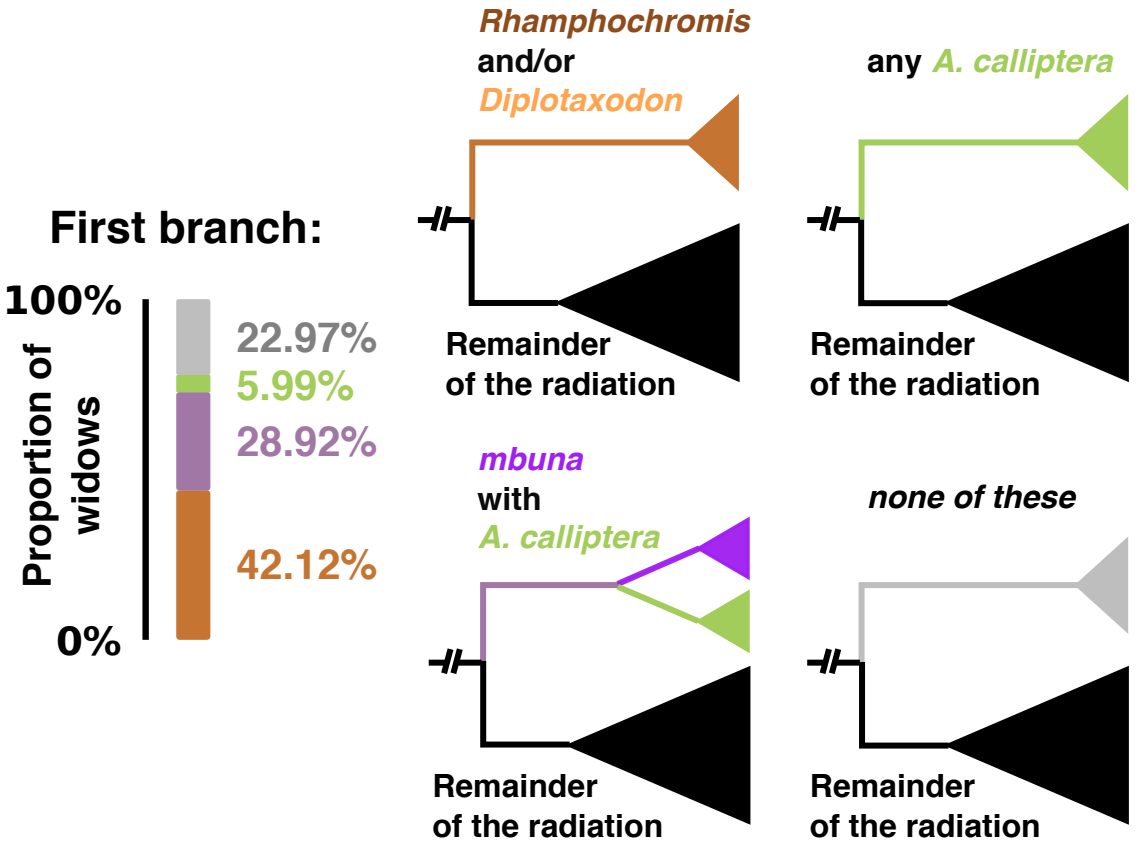
Supplementary Figure 17

A Neighbour Joining tree of *A. calliptera* from all sampled locations. The scale is given in SNPs per bp. All samples, including the one from Lake Kingiri cluster according to geography.



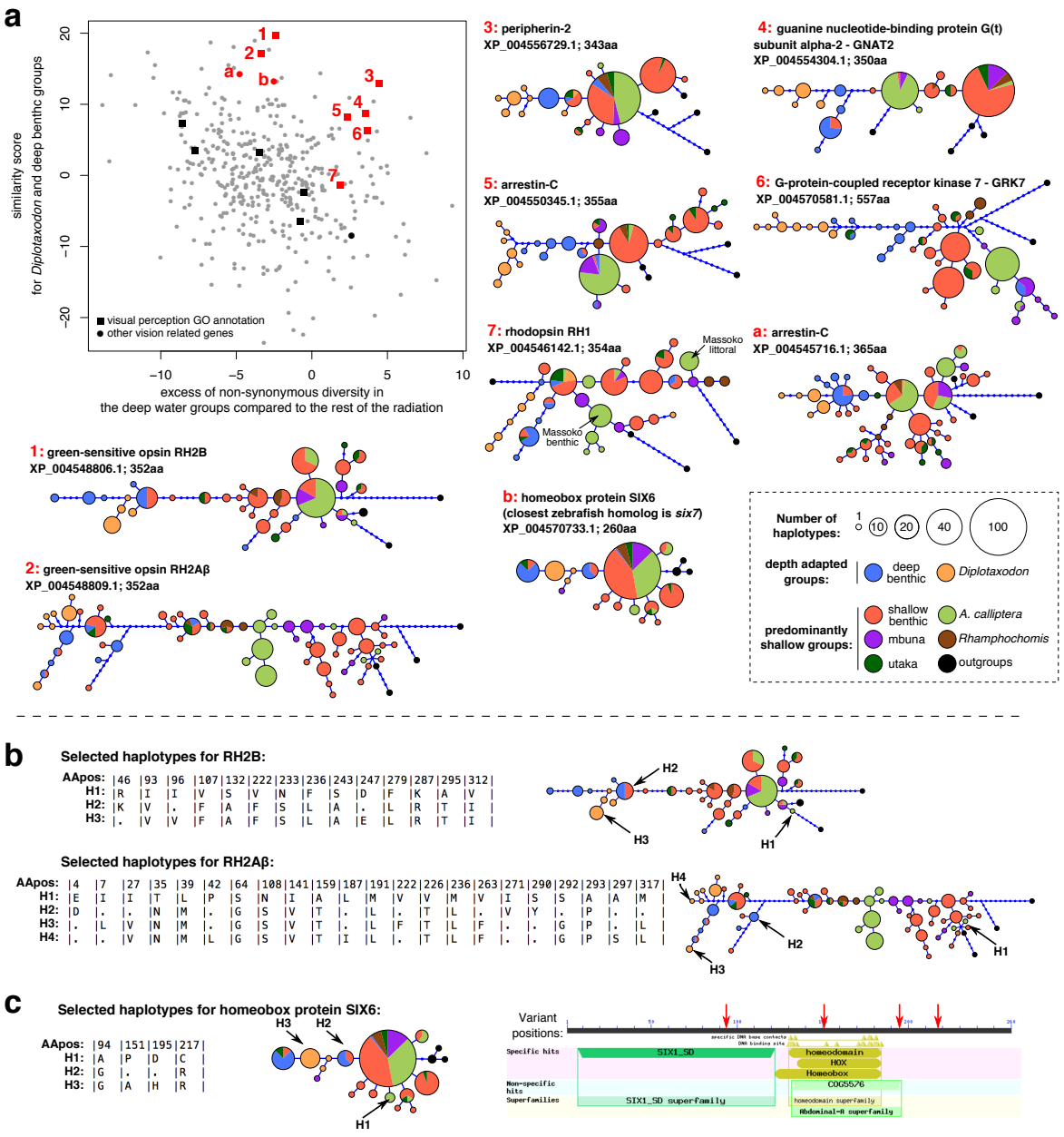
Supplementary Figure 18

Removing all mbuna individuals does not change the NJ tree topology. Specifically, the position of the A. calliptera group remains unchanged. **a,** NJ tree as in Fig 4b. **b,** NJ tree with the mbuna group removed. **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.



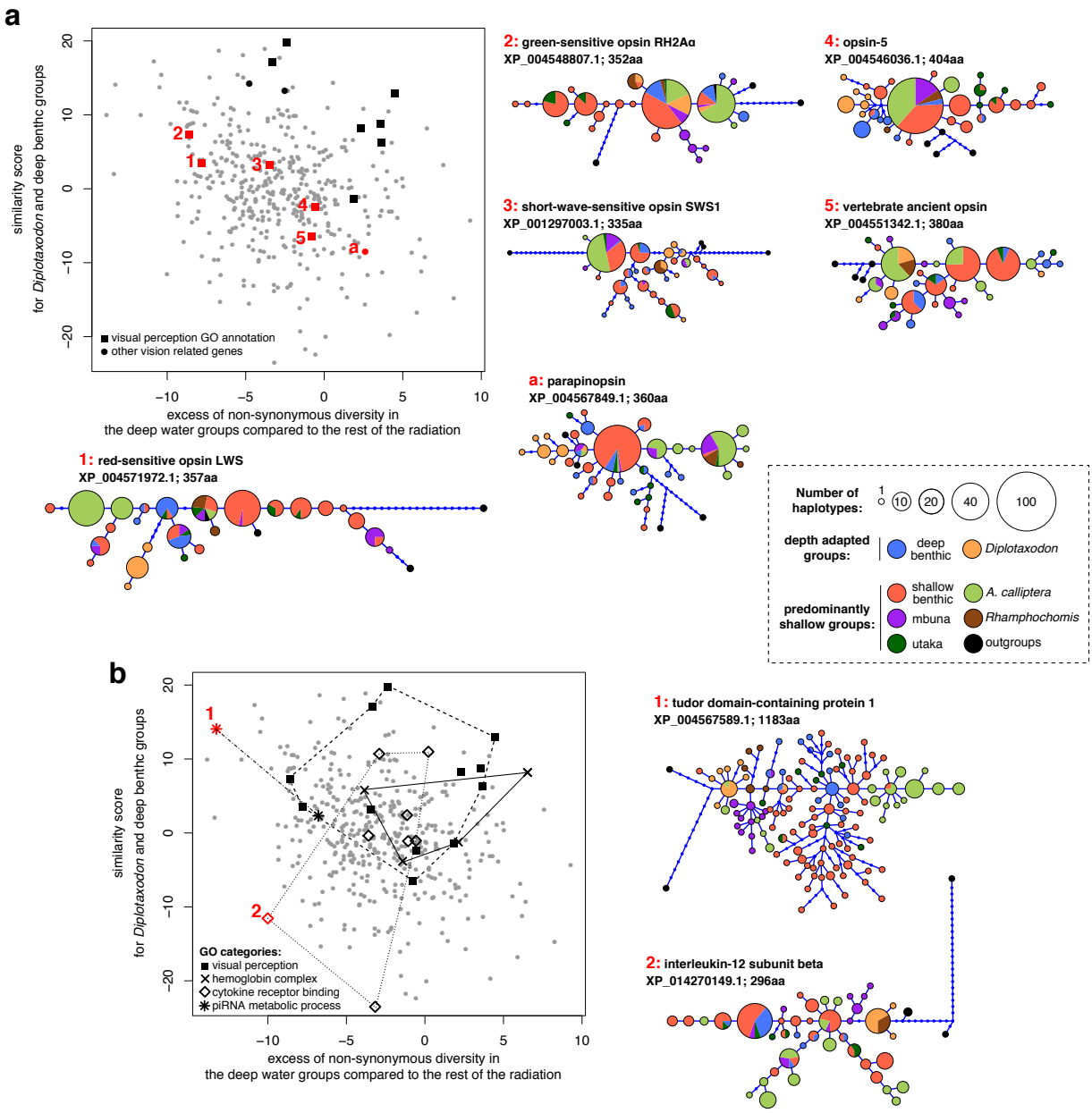
Supplementary Figure 19

Variation in the basal (or 'first') branch among ML phylogenies for 2638 non-overlapping genomic windows of 8000SNPs each. This result is consistent with all the NJ, ASTRAL, and SNAPP trees which place the pelagic *Diplotaxodon* and *Rhamphochromis* together as the sister group to the rest of the radiation.



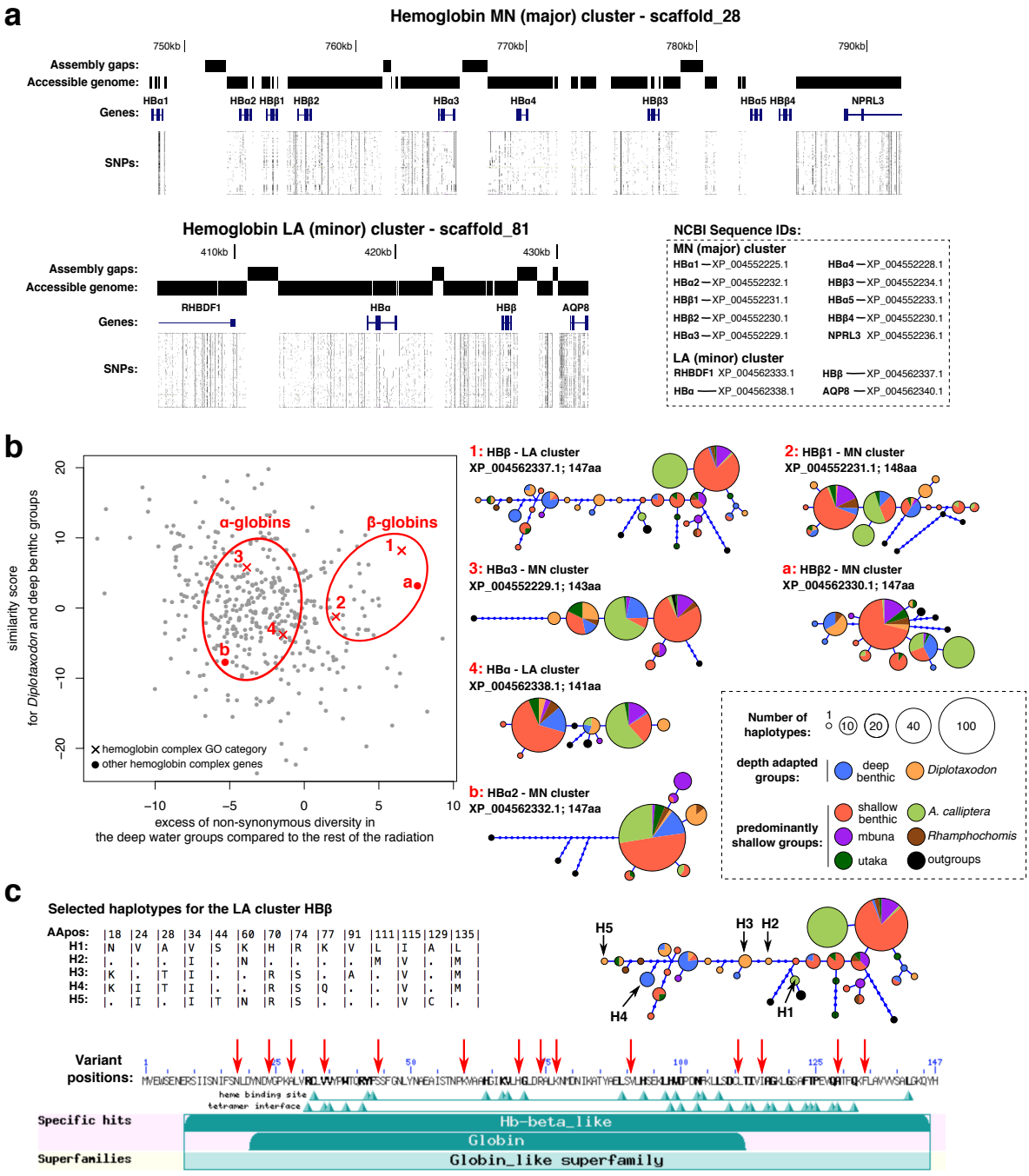
Supplementary Figure 20

Vision-related candidate genes with signatures of shared selection in deep benthic and *Diplotaxodon*. Outgroups include *Oreochromis niloticus*, *Neolamprologus brichardi*, *Astatotilapia burtoni*, and *Pundamilia nyererei*. **a,** Amino acid haplotype trees for genes that have high similarity between and/or excess diversity in deep benthic and *Diplotaxodon*. The homeobox gene is annotated as SIX6 in the NCBI database, but a BLAST search revealed that its closest zebrafish homolog is clearly the *six7* gene. **b,** Amino acid positions of mutations in the two copies of RH2. **c,** Amino acid positions of mutations in the homolog of the homeobox protein *six7*. One variant that distinguishes deep benthic and *Diplotaxodon* is just a residue from the DNA binding site of the HOX domain, while another is in the SIX1_SD domain responsible for binding with a transcriptional activation co-factor[124]



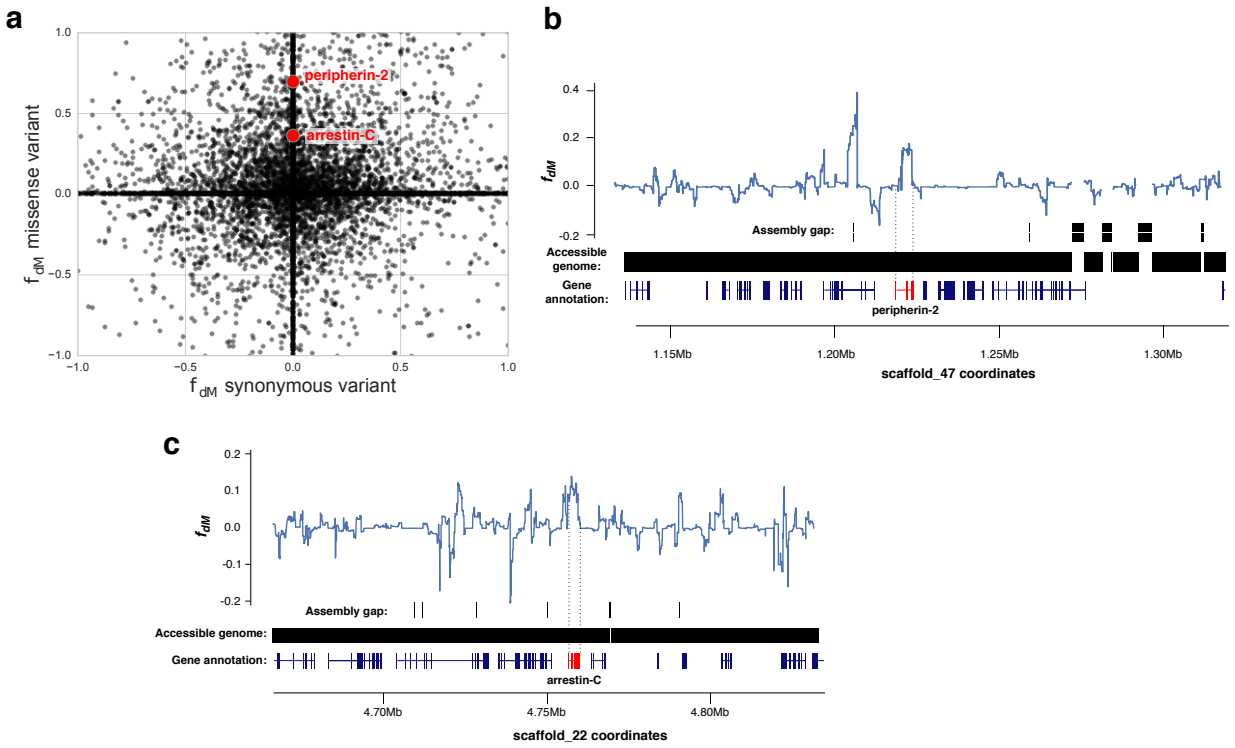
Supplementary Figure 21

Genes with low deep benthic - *Diplotaxodon* scores. The genealogies suggest deep benthic and *Diplotaxodon* groups do not share selection patterns for these genes. Outgroups include *Oreochromis niloticus*, *Neolamprologus brichardi*, *Astatotilapia burtoni*, and *Pundamilia nyererei*. **a,** Amino acid haplotype trees for candidate vision-related genes. **b,** Examples of genes from other GO categories. The legend is shared for both panels a and b.



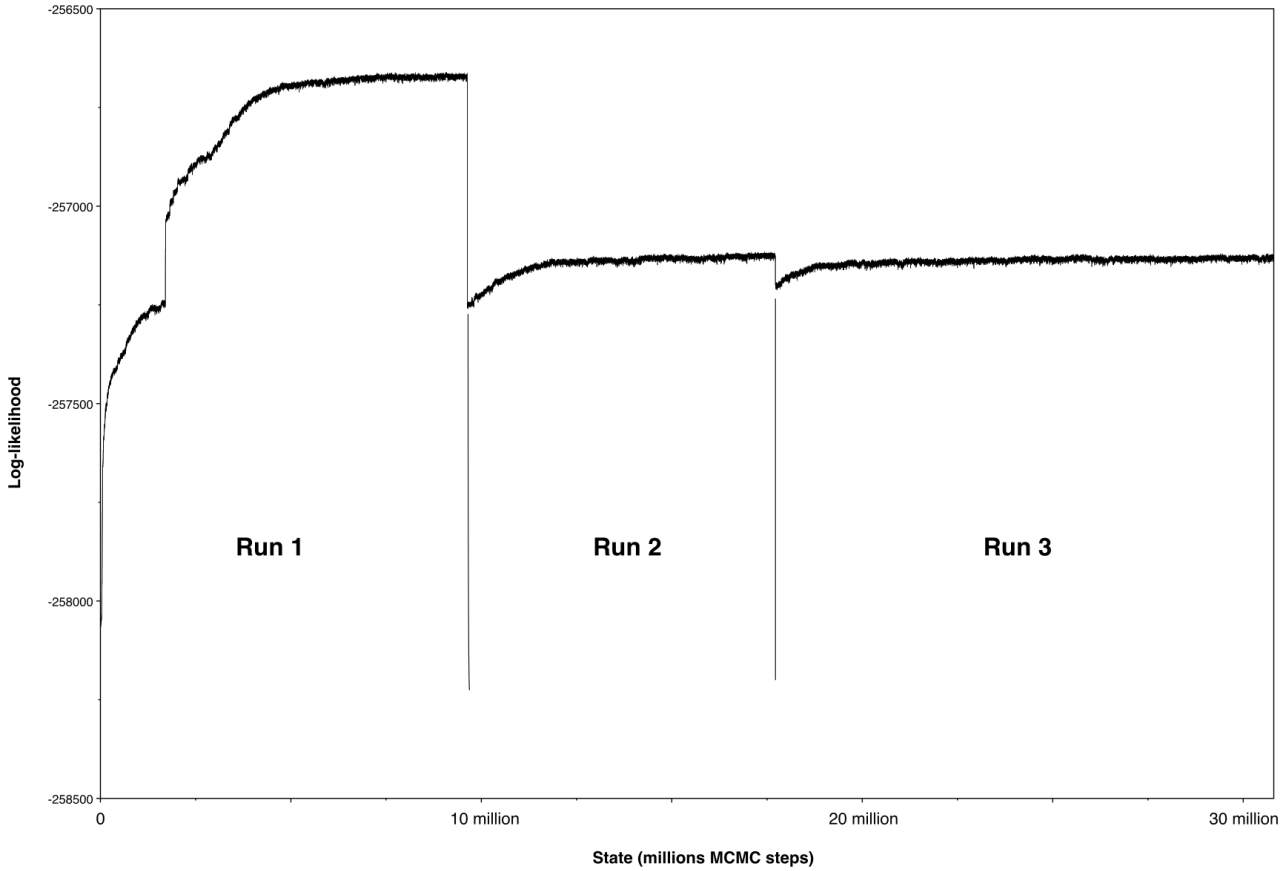
Supplementary Figure 22

Haemoglobin complex genes. a, Genome browser screenshots of the haemoglobin clusters with genes from the BROADMZ2 annotation. Note that the MN cluster genes HBα1, HBα5, HBβ3, and HBβ4 are either entirely or mostly outside of the 'accessible genome'. Due to the highly repetitive nature of the MN locus, we were unable to confidently call variants in these genes. **b,** Amino acid haplotype trees of the genes with high selection scores, with scatterplot showing measures of shared selection signatures between *Diplotaxodon* and deep benthic. **c,** Amino acid positions of mutations in HBβ.



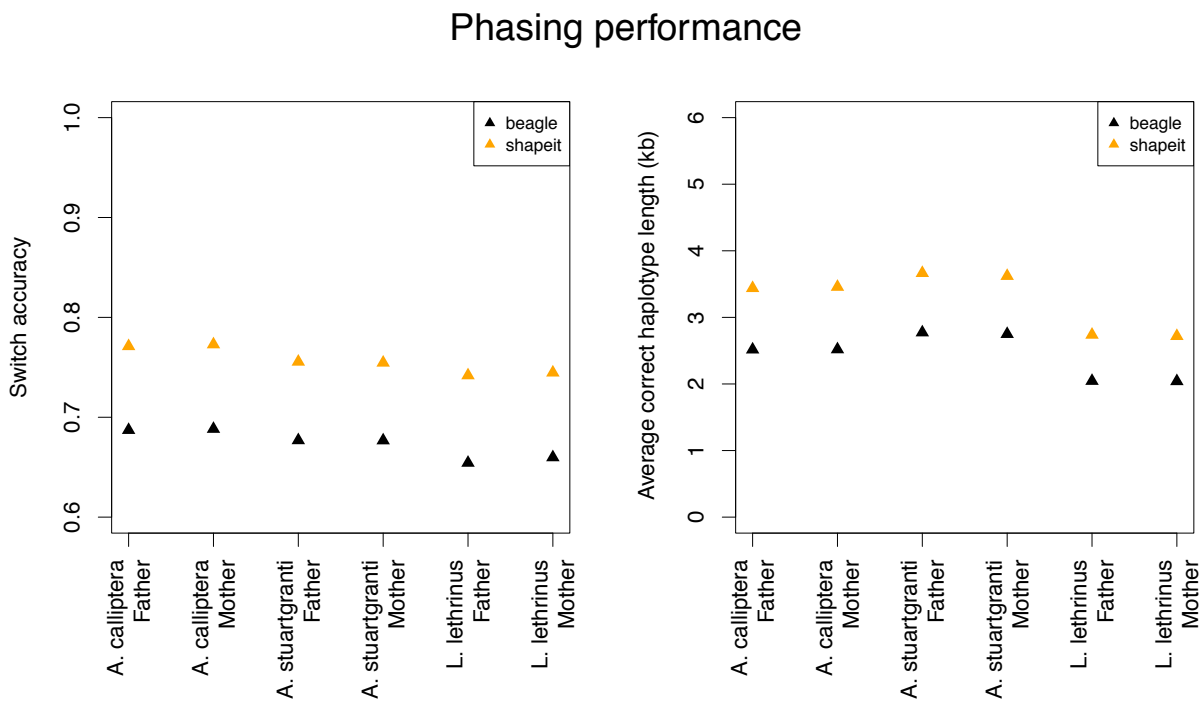
Supplementary Figure 23

Excess allele sharing may be driven by *de novo* mutations in arrestin-C and peripherin-2. a, The f_{dM} measure of excess allele sharing between Diplotaxodon and deep benthic, calculated separately for all synonymous and non-synonymous (missense) variants within each gene. **b, c,** The local patterns of f_{dM} around the two genes.



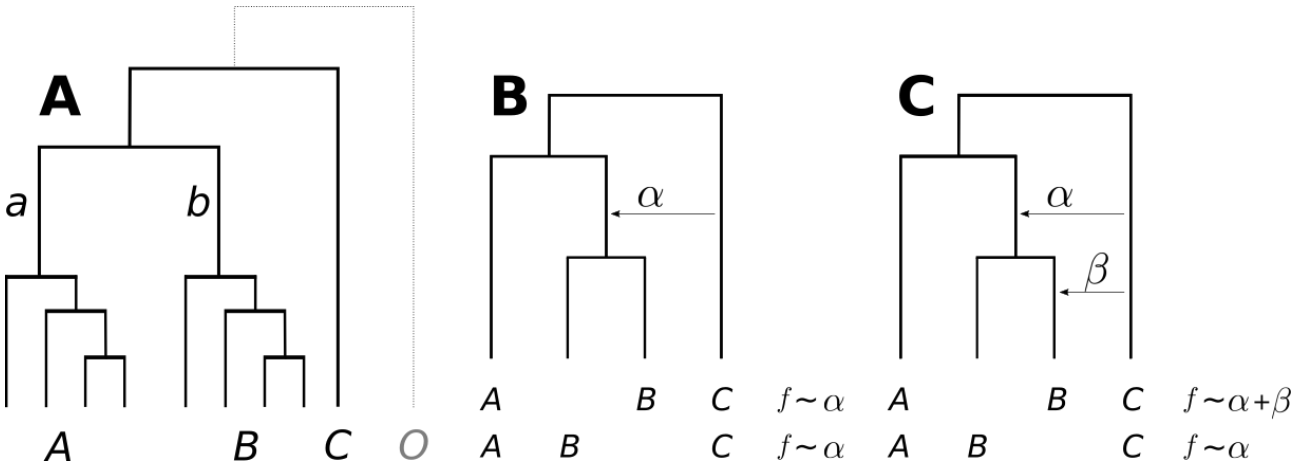
Supplementary Figure 24

Combined MCMC state traces for the tree log-likelihood for the three runs of the SNAPP software. Although all three runs started with the same initial parameters, the first run reached different likelihood values compared with the other two. We interpret this as evidence that the likelihood surface is complicated and includes multiple optima, consistent with there being discordant phylogenetic signals.



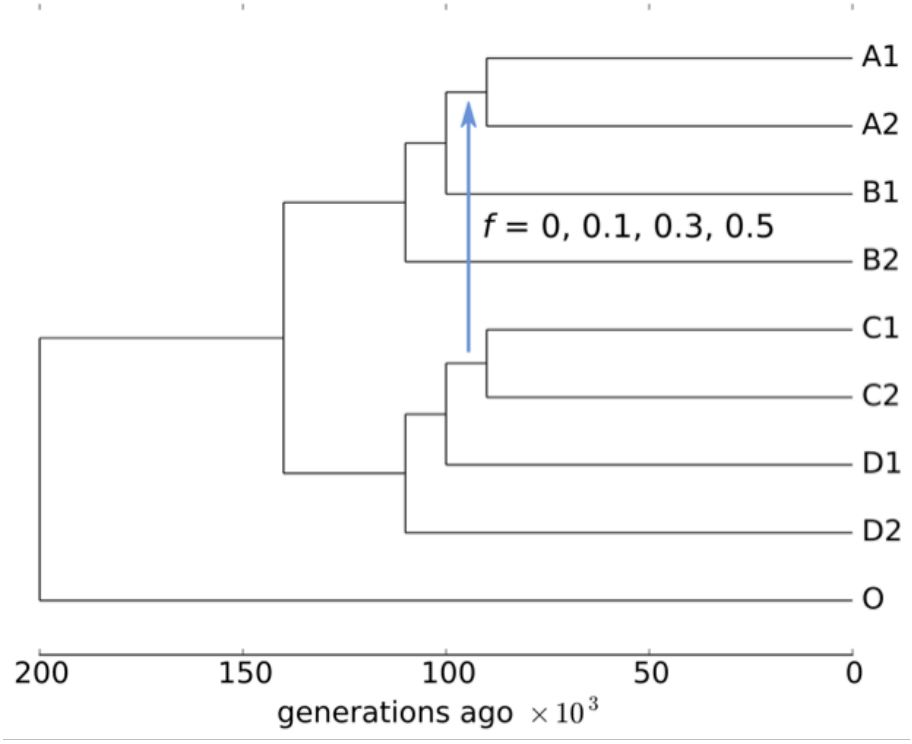
Supplementary Figure 25

Haplotype phasing performance. Improvement in haplotype phasing accuracy achieved by re-phasing the dataset using the `shapeit` software.



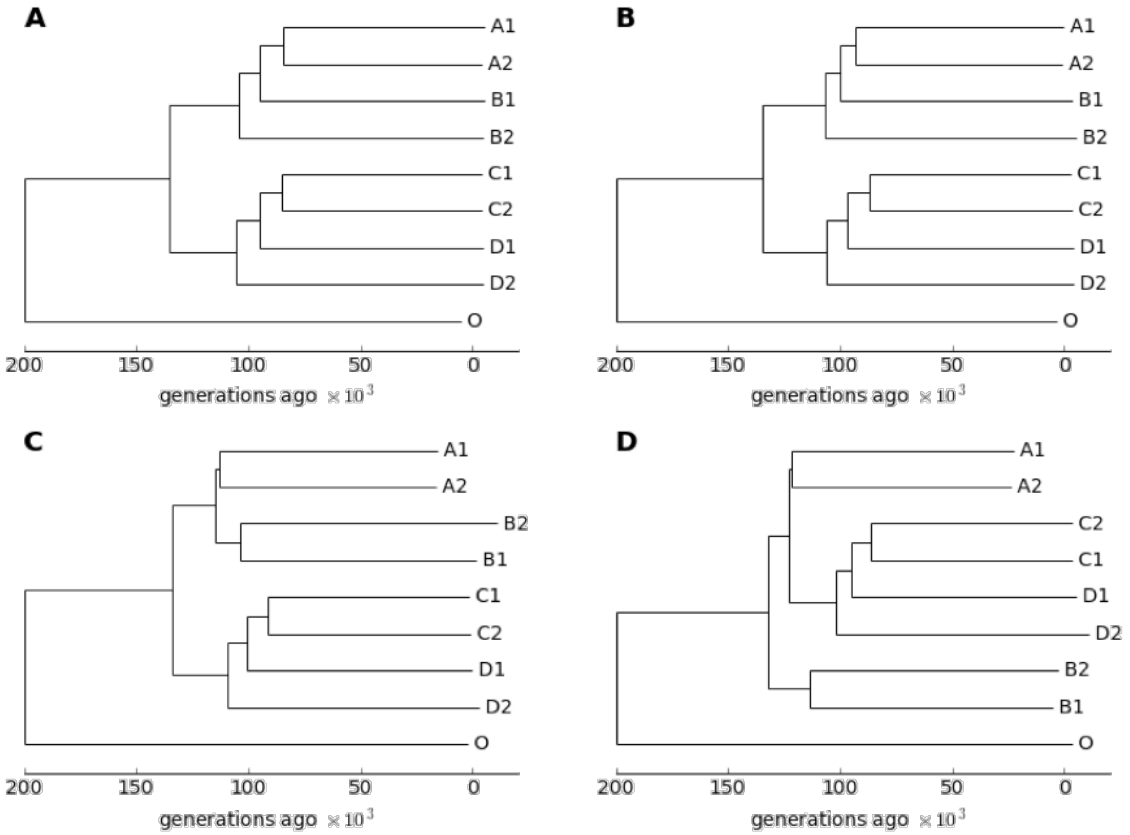
Supplementary Figure 26

The calculation of *f* scores on trees. (**A)** A schematic illustration accompanying the methods section which explains how the reduced branch specific $f_b(C)$ scores are calculated. (**B**, **C)** Illustration of the way in which interdependences between different *f* scores can be informative about the timing of introgression. As shown in (B), gene flow into the common ancestor of two species is expected to be equally detectable in both of them. Additional gene flow into only one species after their split will add to the *f* statistic in that species, but not in its sister species.



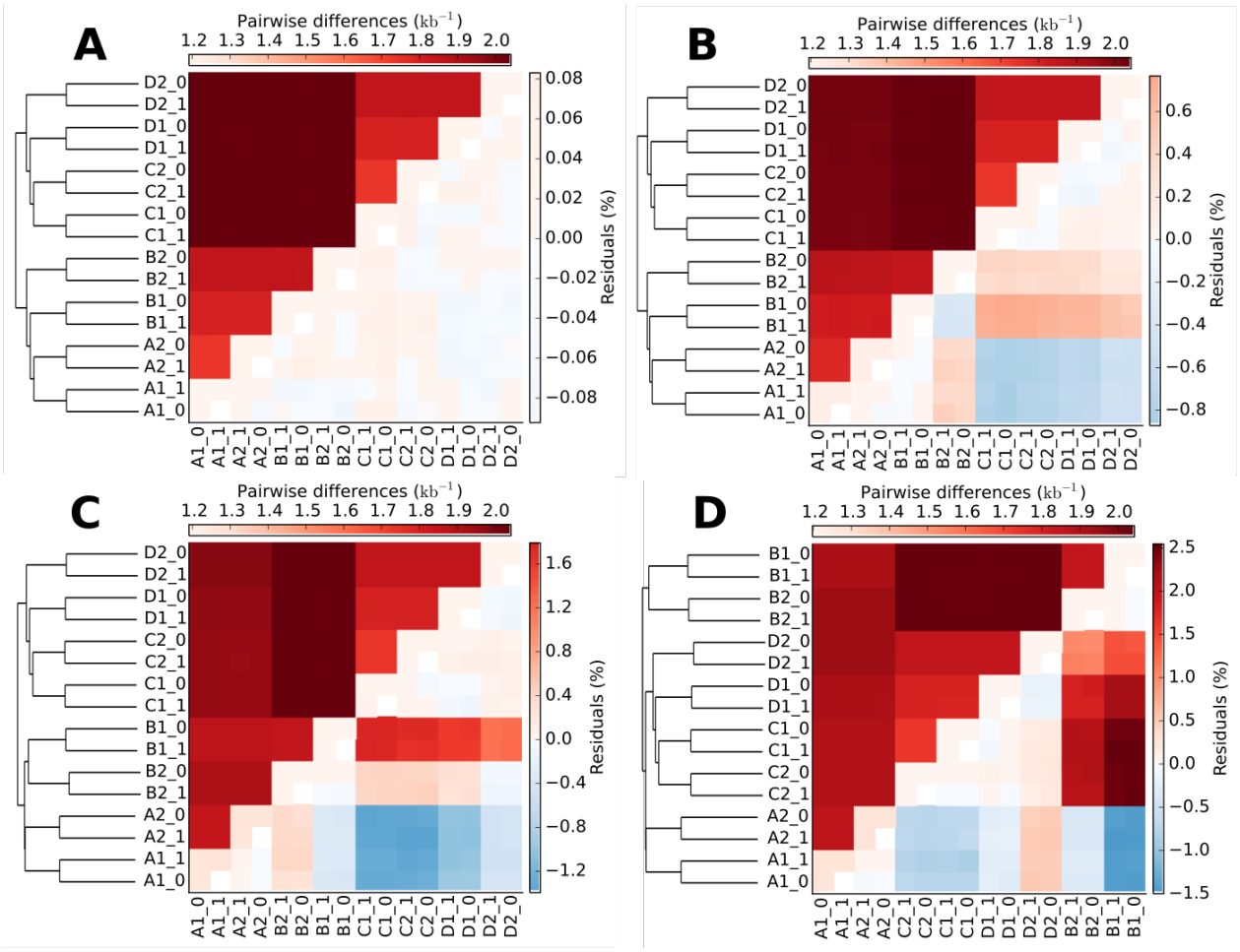
Supplementary Figure 27

Phylogeny used for the coalescent simulations. Two diploid individuals were sampled from species A1, A2, …, D2 and a single diploid individual from the outgroup O. Effective population size was kept constant at 10^5, recombination rate was set to 2×10^{-8} and mutation rate to 3×10^{-9}. For each sample we simulated 120 independent stretches of 5Mb (600Gb in total per individual, corresponding approximately to the size of the accessible genome). We simulated a single pulsed gene flow event from the common ancestor of C1, C2 into the common ancestor of A1, A2 (blue arrow). Four independent runs where performed with $f = 0$, 0.1, 0.3, and 0.5, respectively, where f is the fraction of lineages in the common ancestor of A1, A2 tracing their ancestry though the gene flow event.



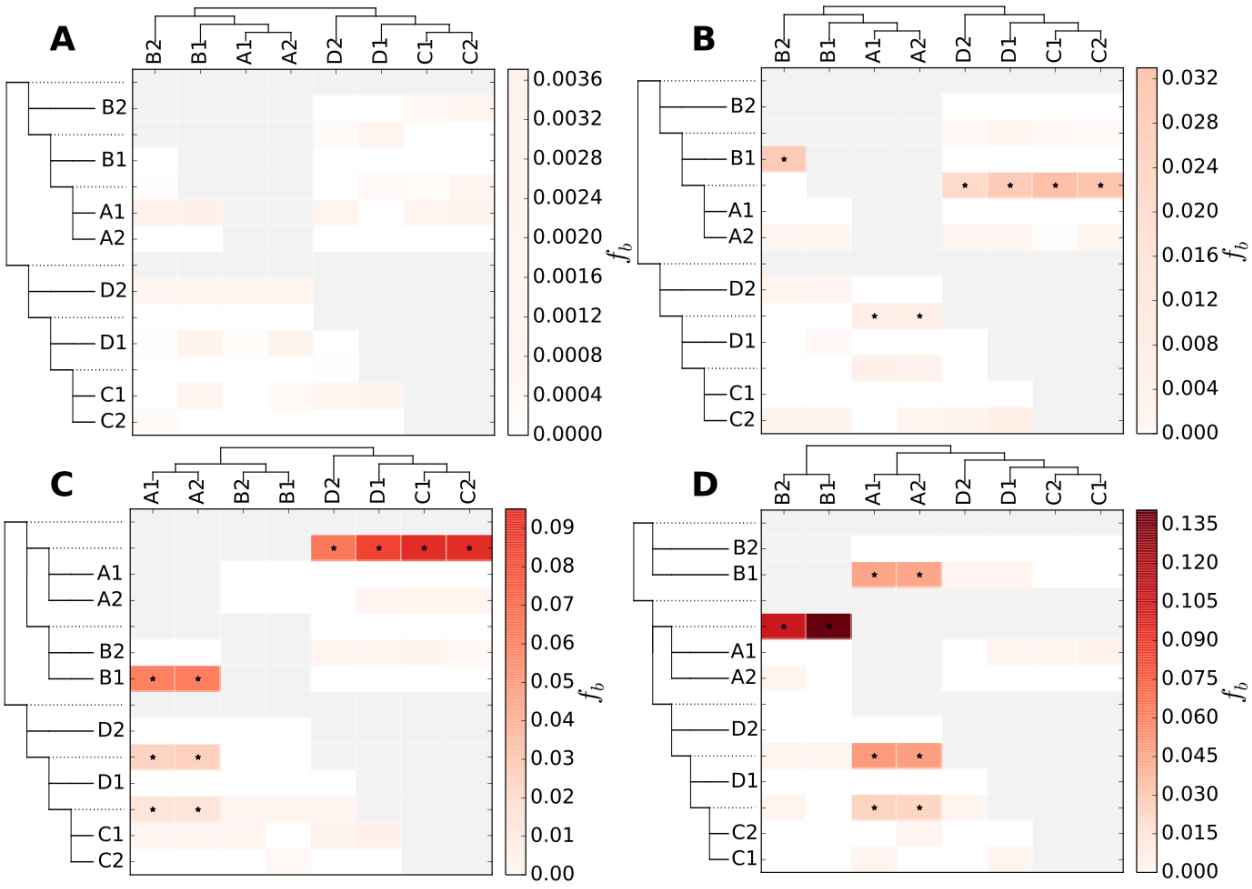
Supplementary Figure 28

Inferred species trees for simulated data of the model in Supplementary Fig. 27. (A) f=0, **(B)** f=0.1, **(C)** f=0.3, **(D)** f=0.5. Trees were constructed using the neighbour-joining method. Coalescent times were inferred from the pairwise genetic differences. Species split times were then calculated by subtracting within population differences from between population differences.



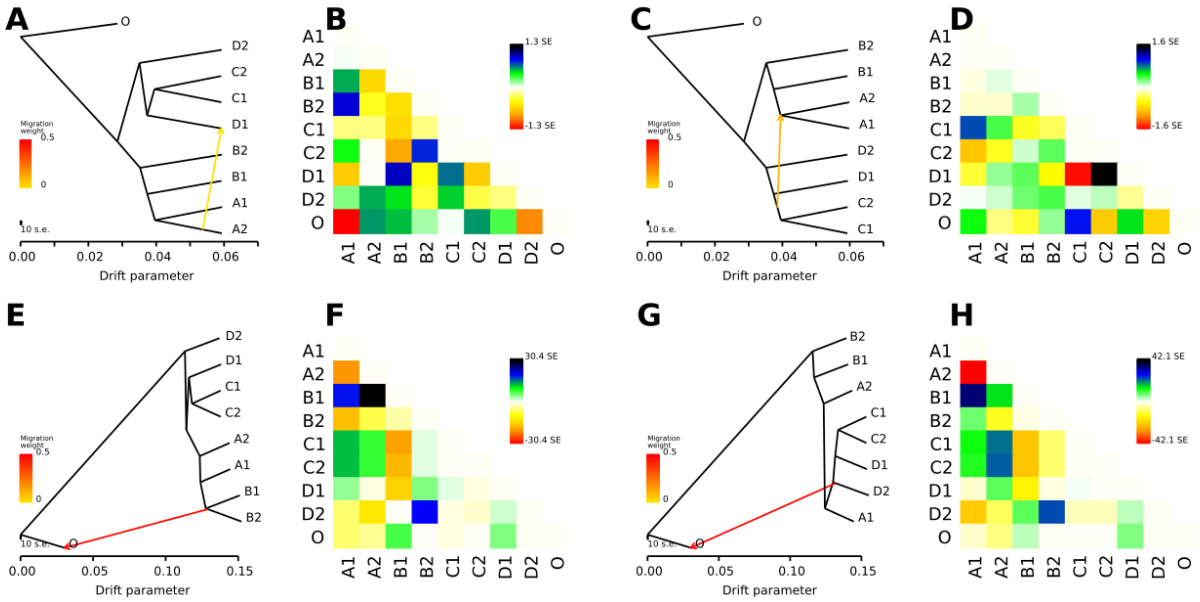
Supplementary Figure 29

Residuals of neighbour-joining trees for simulated data of the model in Supplementary Fig. 27. Pairwise genetic differences (above diagonal) and residuals of pairwise difference and tree distance (below diagonal) trees for (A) f=0, (B) f=0.1, (C) f=0.3, (D) f=0.5. The analysis shown here is equivalent to Supplementary Fig. 12.



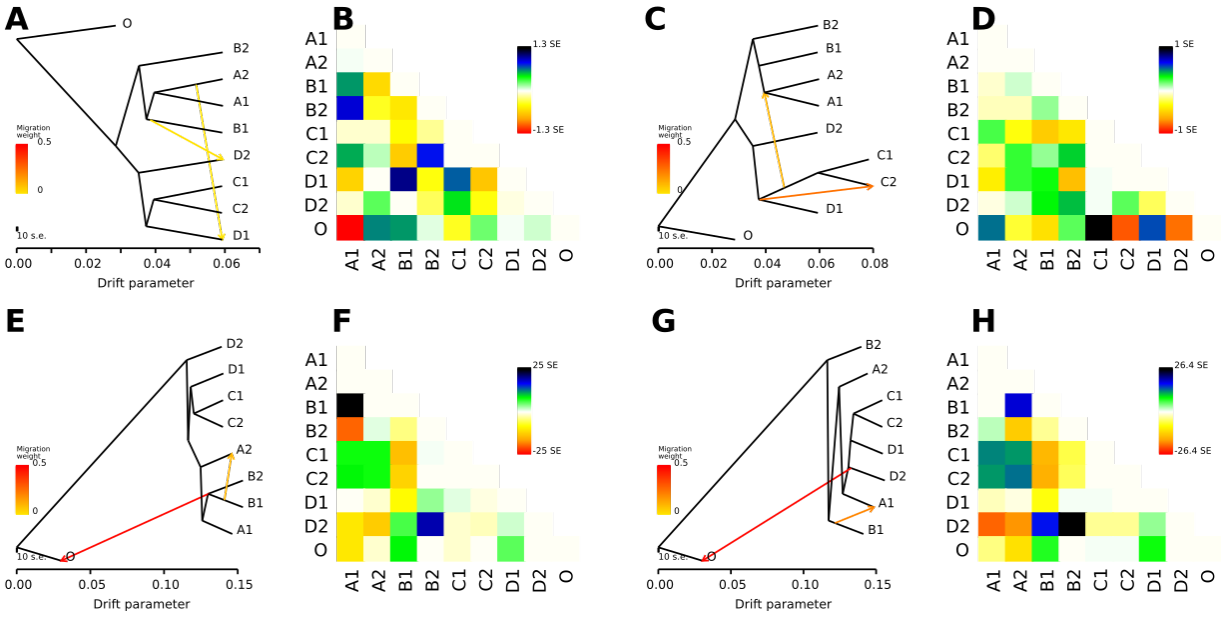
Supplementary Figure 30

The f_b statistic calculated on simulated data of the model in Supplementary Fig. 27. (A) f=0, **(B)** f=0.1, **(C)** f=0.3, **(D)** f=0.5. The analysis shown here is equivalent to Fig. 3 from the main manuscript.



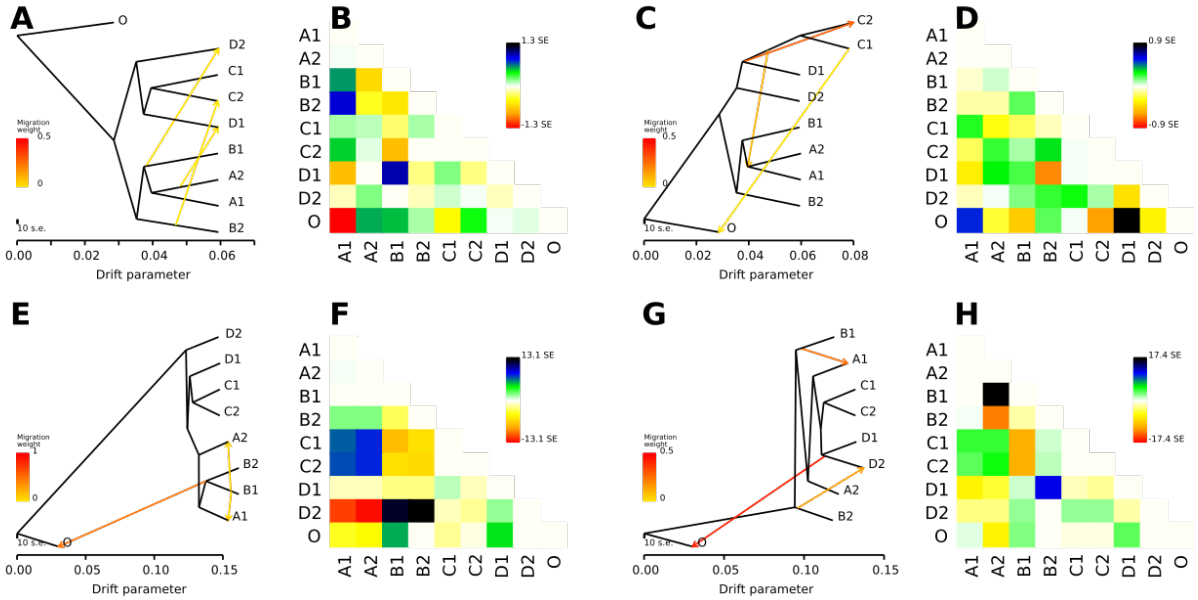
Supplementary Figure 31

Inference using the `treemix` software with `m=1` for simulated data of the model in Supplementary Fig. 27. (A-B) *f*=0, (C-D) *f*=0.1, (E-F) *f*=0.3, (G-H) *f*=0.5. The `m=1` parameter tells `treemix` to expect exactly one migration event.



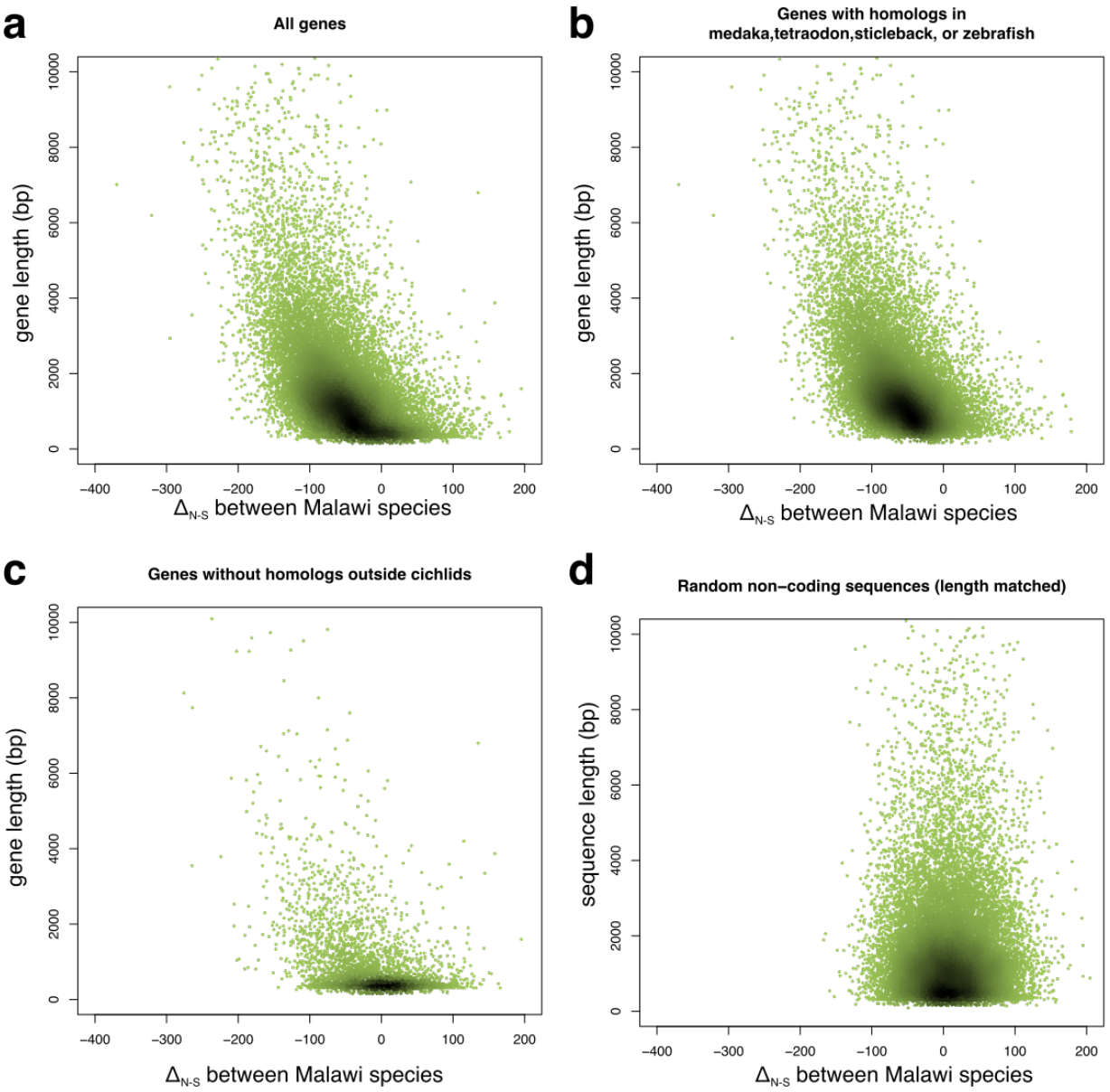
Supplementary Figure 32

Inference using the `treemix` software with `m=2` for simulated data of the model in Supplementary Fig. 27. (A-B) f=0, **(C-D)** f=0.1, **(E-F)** f=0.3, **(G-H)** f=0.5. The `m=2` parameter tells `treemix` to expect exactly two migration events.



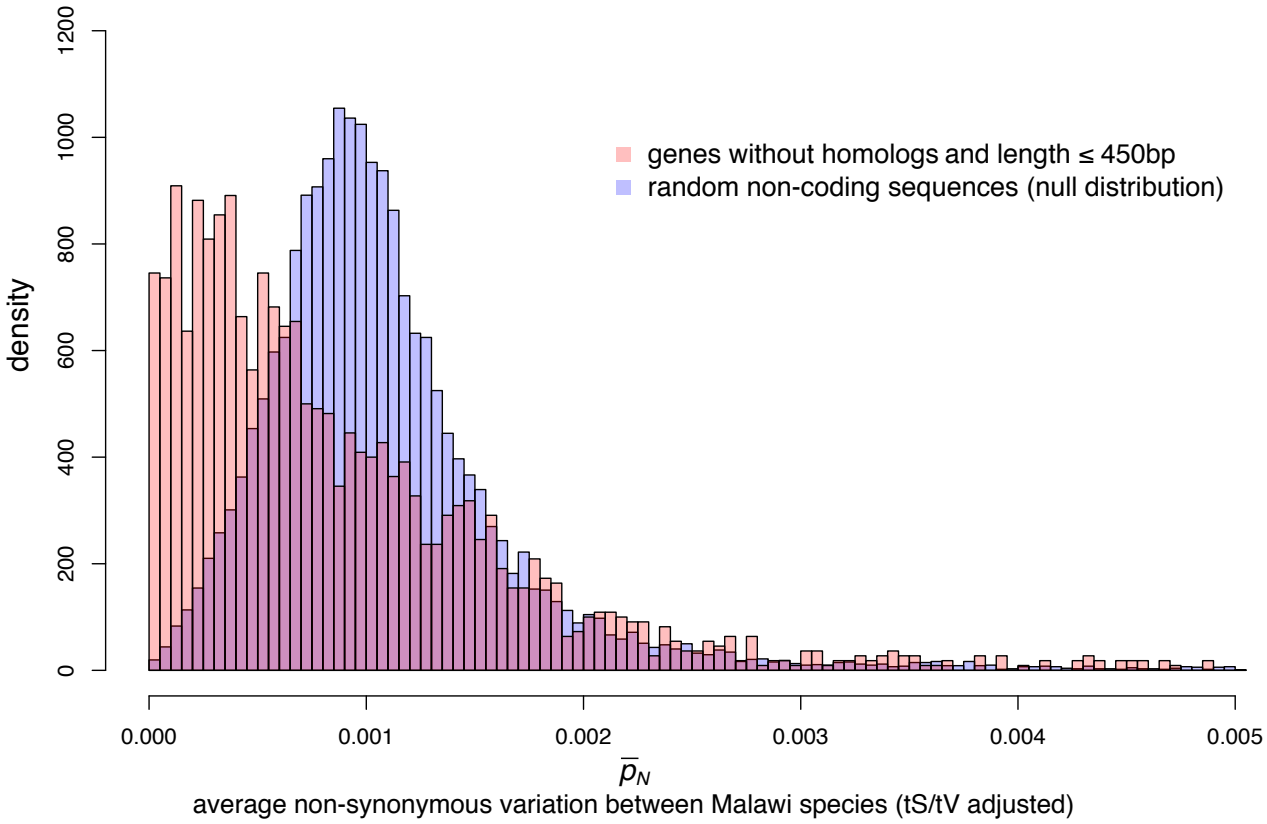
Supplementary Figure 33

Inference using the `treemix` software with `m=3` for simulated data of the model in Supplementary Fig. 27. (A-B) *f*=0, **(C-D)** *f*=0.1, **(E-F)** *f*=0.3, **(G-H)** *f*=0.5. The `m=3` parameter tells `treemix` to expect exactly three migration events.



Supplementary Figure 34

Non-synonymous variation excess scores (Δ_{N-S}) and gene-length. Darker colour indicates greater density of data points. **a,** Overall, there is a disproportionally large fraction of short genes with high Δ_{N-S} values. **b,** This effect is greatly reduced when considering genes for which there are homologs in other non-cichlid teleost genome assemblies separately. **c,** Genes without homologs in other teleosts tend to be short and have higher Δ_{N-S} values; especially the shorter ones have a Δ_{N-S} similar to that observed in non-coding sequences shown in panel **d**.



Supplementary Figure 35

Overlaid density histograms of non-synonymous diversity ($\bar{p}_N$). The distributions for short genes without homologs on other teleosts and for matched randomly selected non-coding sequences. The comparison of these distributions shows that short genes without homologs have a component with low $\bar{p}_N$, suggesting purifying selection, as well as a component with high $\bar{p}_N$, suggesting diversifying selection.

Supplementary Table 1

An overview of Lake Malawi and species and samples sequenced in this study. Approximate coverage and Illumina sequencing chemistry used (v3 or v4) are indicated.

Group	Species	Number of individuals: ~coverage/chemistry				Total individuals
		~15x/v3	~15x/v4	~6x/v3	~6x/v4	
mbuna	*Cynotilapia afra*		1			**1**
	Cynotilapia axelrodi		1			**1**
	Genyochromis mento		1			**1**
	Iodotropheus sprengerae		1			**1**
	Labeotropheus trewavasae				1	**1**
	Metriaclima zebra		1			**1**
	Petrotilapia genalutea		1			**1**
	Tropheops tropheops		1			**1**
deep benthic	*Alticorpus geoffreyi*	1				**1**
	Alticorpus macrocleithrum	1				**1**
	Aulonocara 'minutus'		1			**1**
	Aulonocara steveni	1				**1**
	Aulonocara stuartgranti	3				**3 (trio)**
	Aulonocara 'yellow'		1			**1**
	Lethrinops gossei	1				**1**
	Lethrinops 'longimanus redhead'	1				**1**
	Lethrinops 'oliveri'	1				**1**
shallow benthic	*Buccochromis nototaenia*		1			**1**
	Buccochromis rhoadesii	1				**1**
	Champsochromis caeruleus		2			**2**
	Chilotilapia rhoadesii		1		3	**4**
	Copadichromis cf. trewavasae	1				**1**
	Ctenopharynx intermedius		1		2	**3**
	Ctenopharynx nitidus		2			**2**
	Dimidiochromis compressiceps		1			**1**
	Dimidiochromis dimidiatus		1			**1**
	Dimidiochromis kiwinge		1			**1**
	Dimidiochromis strigatus		2			**2**
	Fossorochromis rostratus				2	**2**
	Hemitaeniochromis spilopterus		2			**2**
	Hemitilapia oxyrhynchus		2			**2**
	Lethrinops albus	1				**1**
	Lethrinops auritus	1				**1**
	Lethrinops lethrinus	4				**4 (trio+1)**
	Mylochromis anaphyrmus	1		4		**5**
	Mylochromis ericotaenia		1			**1**
	Mylochromis melanotaenia		1			**1**
	Nimbochromis linni	1				**1**
	Nimbochromis livingstoni	1				**1**
	Nimbochromis polystigma	1				**1**
	Otopharynx 'brooksi nkhata'	1				**1**
	Otopharynx lithobates		1			**1**
	Otopharynx speciosus		2			**2**
	Otopharynx tetrastigma (Lake Ilamba)	1				**1**
	Placidochromis electra	1				**1**
	Placidochromis johnstoni	1				**1**
	Placidochromis cf. longimanus		1		4	**5**
	Placidochromis milomo	1				**1**
	Placidochromis subocularis				8	**8**
	Protomelas ornatus		2			**2**
	Stigmatochromis guttatus		1			**1**
	Stigmatochromis modestus		1			**1**
	Taeniochromis holotaenia		1			**1**
	Taeniolethrinops furcicauda		1			**1**
	Taeniolethrinops macrorhynchus		1			**1**
	Taeniolethrinops praeorbitalis		1			**1**
	Trematocranus placodon	1		4		**5**
	Tyrannochromis nigriventer		1			**1**
utaka	*Copadichromis likomae*		1			**1**
	Copadichromis quadrimaculatus	1				**1**
	Copadichromis trimaculatus		1			**1**
	Copadichromis virginalis	1			4	**5**
Rhampochromis	*Rhamphochromis esox*		1			**1**
	Rhamphochromis longiceps		1			**1**
	Rhamphochromis woodi		1			**1**

Diplotaxodon	*Diplotaxodon greenwoodi*		1			**1**
	Diplotaxodon limnothrissa		1			**1**
	Diplotaxodon macrops		1			**1**
	Diplotaxodon 'macrops black dorsal'		1			**1**
	Diplotaxodon 'macrops ngulube'		1			**1**
	Diplotaxodon 'similis white back'		1			**1**
	Pallidochromis tokolosh		1			**1**
A. calliptera	*Astatotilapia calliptera*	8	13			**21 (trio+18)**
	Number of species: 73	**36**	**67**	**8**	**23**	**134**

Supplementary Table 2

Outgroup *Astatotilapia* species and specimens sequenced for this study. All individuals were sequenced to approximately 15x coverage using Illumina v4 sequencing chemistry (125bp paired end reads).

Species	Sampling location(s)	Number of individuals
Astatotilapia bloyeti	Kilosa, Wami System Lake Burungi Lake Kumba, Korogwe District, Tanzania	3
Astatotilapia 'rukwa'	Lake Rukwa	3
Astatotilapia 'rufiji blue'	Oxbow Lake (Rufiji river delta) Lake Mansi (Rufiji river system)	2
Astatotilapia 'ruaha 1'	Kidatu, Ruaha river	2
Astatotilapia tweddlei	Ruhuhu river Ruaha river Mindu Dam Kitele Lake (Ruvuma catchment)	5
Astatotilapia 'ruaha 2'	Kidatu, Ruaha river	2
Astatotilapia burtoni	Rusizi, Bujumbura Lab strain from ref. 11	2
Total		**19**

Supplementary Table 3

Gene ontology terms with significant (p≤0.01) enrichment when using the `weight` algorithm [60] implemented in the `topGO` package [111].

Molecular function:

	GO.ID	Term	Annotated	Significant	Expected	Rank in classic	classic	weight
1	GO:0009881	photoreceptor activity	20	8	0.51	1	1.6e-08	1.6e-08
2	GO:0020037	heme binding	60	11	1.54	2	2.9e-07	2.9e-07
3	GO:0005344	oxygen transporter activity	7	4	0.18	5	1.4e-05	1.4e-05
4	GO:0005506	iron ion binding	89	11	2.28	6	1.6e-05	1.6e-05
5	GO:0019825	oxygen binding	8	4	0.21	7	2.7e-05	2.7e-05
6	GO:0005126	cytokine receptor binding	58	8	1.49	8	0.00011	0.00029
7	GO:0005125	cytokine activity	51	6	1.31	9	0.00184	0.00184
8	GO:0030414	peptidase inhibitor activity	61	6	1.57	10	0.00458	0.00458
9	GO:0051537	2 iron, 2 sulfur cluster binding	14	3	0.36	12	0.00493	0.00493
10	GO:0005509	calcium ion binding	366	18	9.39	13	0.00616	0.00616

Cellular component:

	GO.ID	Term	Annotated	Significant	Expected	Rank in classic	classic	weight
1	GO:0001750	photoreceptor outer segment	11	6	0.29	1	1.4e-07	1.4e-07
2	GO:0005833	hemoglobin complex	5	4	0.13	2	2.5e-06	2.5e-06
3	GO:0031224	intrinsic component of membrane	3002	110	80.51	4	4.3e-05	9.8e-05
4	GO:0005576	extracellular region	542	28	14.54	8	0.00061	0.0033

Biological Process:

	GO.ID	Term	Annotated	Significant	Expected	Rank in classic	classic	weight
1	GO:0007602	phototransduction	24	8	0.58	2	5.5e-08	5.5e-08
2	GO:0018298	protein-chromophore linkage	19	7	0.46	4	1.8e-07	1.8e-07
3	GO:0007601	visual perception	64	10	1.55	7	2.7e-06	2.7e-06
4	GO:0016226	iron-sulfur cluster assembly	12	5	0.29	11	5.5e-06	5.5e-06
5	GO:0015671	oxygen transport	7	4	0.17	13	1.1e-05	1.1e-05
6	GO:0071482	cellular response to light stimulus	29	6	0.70	16	5.6e-05	5.6e-05
7	GO:0009615	response to virus	21	4	0.51	30	0.00145	0.0015
8	GO:0071466	cellular response to xenobiotic stimulus	23	4	0.56	32	0.00207	0.0021
9	GO:0006954	inflammatory response	77	7	1.87	33	0.00255	0.0025
10	GO:0050730	regulation of peptidyl-tyrosine phosphor...	11	3	0.27	31	0.00200	0.0034
11	GO:0055113	epiboly involved in gastrulation with mo...	28	4	0.68	42	0.00435	0.0043
12	GO:0009204	deoxyribonucleoside triphosphate catabol...	5	2	0.12	47	0.00557	0.0056
13	GO:0009264	deoxyribonucleotide catabolic process	5	2	0.12	48	0.00557	0.0056
14	GO:0034587	piRNA metabolic process	5	2	0.12	49	0.00557	0.0056
15	GO:0070098	chemokine-mediated signaling pathway	5	2	0.12	50	0.00557	0.0056
16	GO:0009394	2'-deoxyribonucleotide metabolic process	6	2	0.15	56	0.00822	0.0082

Supplementary Table 4

Individual BioSample accessions for whole genome sequencing data used in this study.

Samples	BioSample Accessions
Lake Malawi populations	SAMEA1877409, SAMEA1877411, SAMEA1877414, SAMEA1877417, SAMEA1877421, SAMEA1877429, SAMEA1877440, SAMEA1877451, SAMEA1877455, SAMEA1877459, SAMEA1877464, SAMEA1877472, SAMEA1877476, SAMEA1877480, SAMEA1877484, SAMEA1877499, SAMEA1877503, SAMEA1904322, SAMEA1904324, SAMEA1904329-SAMEA1904331, SAMEA2661216-SAMEA2661246, SAMEA2661250-SAMEA2661253, SAMEA2661255-SAMEA2661258, SAMEA2661260, SAMEA2661262, SAMEA2661264-SAMEA2661270, SAMEA2661272, SAMEA2661275-SAMEA2661282, SAMEA2661287-SAMEA2661290, SAMEA3388853-SAMEA3388860, SAMEA3388862-SAMEA3388864, SAMEA3388868, SAMEA3388870-SAMEA3388874
Lake Malawi trios	SAMEA1920096-SAMEA1920098 *A. stuartgranti* SAMEA1920093-SAMEA1920095 *L. lethrinus* SAMEA1920090-SAMEA1920092 *A. calliptera* Salima, Lake Malawi
***A. calliptera* Malawi catchment**	SAMEA1877400, SAMEA1904326, SAMEA1904327, SAMEA2661273, SAMEA2661381-SAMEA2661385, SAMEA2661389-SAMEA2661391
***A. calliptera* Indian Ocean catchments**	SAMEA1904323, SAMEA1904328, SAMEA2661386-SAMEA2661388
***Astatotilapia* outgroups**	Under PRJEB1254: SAMEA2661249, SAMEA3388867 Under PRJEB15289: SAMEA4033331, SAMEA4033332, SAMEA4033339, SAMEA4033340 SAMEA4033333, SAMEA4033334, SAMEA4033341

Supplementary Table 5

Versions of cichlid genome assemblies used in this study. All described in ref. 11.

Species	Broad Institute Assembly ID	URLs used to download
M. zebra	MetZeb1.1_prescreen	http://www.broadinstitute.org/ftp/pub/assemblies/fish/M_zebra/MetZeb1.1_prescreen/M_zebra_v0.assembly.fasta
P. nyererei	PunNye1.0	http://www.broadinstitute.org/ftp/pub/assemblies/fish/P_nyererei/PunNye1.0/P_nyererei_v1.assembly.fasta.gz
A. burtoni	HapBur1.0	http://www.broadinstitute.org/ftp/pub/assemblies/fish/H_burtoni/HapBur1.0/H_burtoni_v1.assembly.fasta.gz
N. brichardi	NeoBri1.0	http://www.broadinstitute.org/ftp/pub/assemblies/fish/N_brichardi/NeoBri1.0/N_brichardi_v1.assembly.fasta.gz
O. niloticus	Oreni1.1	http://www.broadinstitute.org/ftp/pub/assemblies/fish/tilapia/Oreni1.1/20120125_MapAssembly.anchored.assembly.fasta

Supplementary Table 6

Versions of non-cichlid teleost assemblies used for multiple whole-genome alignments.

Species	UCSC version of assembly	Notes	URLs used to download
medaka	oryLat2	NIG v1.0 assembly	ftp://hgdownload.soe.ucsc.edu/goldenPath/oryLat2/bigZips/oryLat2.fa.gz
stickleback	gasAcu1	Broad Institute v1.0	ftp://hgdownload.soe.ucsc.edu/gbdb/gasAcu1/gasAcu1.2bit
zebrafish	danRer7	Sanger Zv9 assembly	http://hgdownload.cse.ucsc.edu/gbdb/danRer7/danRer7.2bit

Supplementary Table 7

An overview of the specimens whose photographs were used in the morphometric analyses. We used a total of 47 specimens from seven *Astatotilapia* outgroup species, 12 *A. calliptera* specimens, and 109 specimens from 55 Lake Malawi endemic species (a subset of the 72 sequenced Lake Malawi species for which we had available high quality photographs).

Group	N species	N populations	N specimens
***Astatotilapia* outgroups**	**7**	**10**	**47**
A. 'rukwa', 'rufiji blue', 'ruaha 1'	3	3	19
A. tweddlei ,'ruaha 2'	2	2	10
A. bloyeti	1	2	9
A. burtoni	1	3	9
A. calliptera	**1**	**3**	**12**
Malawi radiation	**55**		**109**
mbuna	6		12
shallow benthic	30		55
deep benthic	8		15
utaka	2		6
Diplotaxodon	6		12
Rhamphochromis	3		9
Total	**62**		**168**

Additional References:

114. 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).
115. Harris, R. S. Improved pairwise alignment of genomic DNA. (PhD Thesis: The Pennsylvania State University, 2007).
116. 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).
117. Bertram, C. K. R., Borley, H. J. H. & Trewavas, E. *Report on the fish and fisheries of lake Nyasa*. (1942).
118. Snoeks, J. *The cichlid diversity of Lake Malawi/Nyasa/Niassa*. (Cichlid Press, 2004).
119. 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).
120. Joyce, D. A. *et al.* An extant cichlid fish radiation emerged in an extinct Pleistocene lake. *Nature* **435,** 90–95 (2005).
121. 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).
122. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* **12,** e1004842 (2016).
123. Seehausen, O. *et al.* Speciation through sensory drive in cichlid fish. *Nature* **455,** 620–626 (2008).
124. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45,** D200–D203 (2017).