# Large inversions in Lake Malawi cichlids are associated with habitat preference, lineage, and sex determination

**Nikesh M. Kumar[1], Taylor L. Cooper[1], Thomas D. Kocher[2], J. Todd Streelman[1†], and Patrick T. McGrath[1,3†]**

**[1] School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA**

**[2] Department of Biology, University of Maryland, College Park, MD USA**

**[3] School of Physics, Georgia Institute of Technology, Atlanta, GA, USA**

**[†] Co-correspondence: patrick.mcgrath@biology.gatech.edu, todd.streelman@biology.gatech.edu**

## ABSTRACT

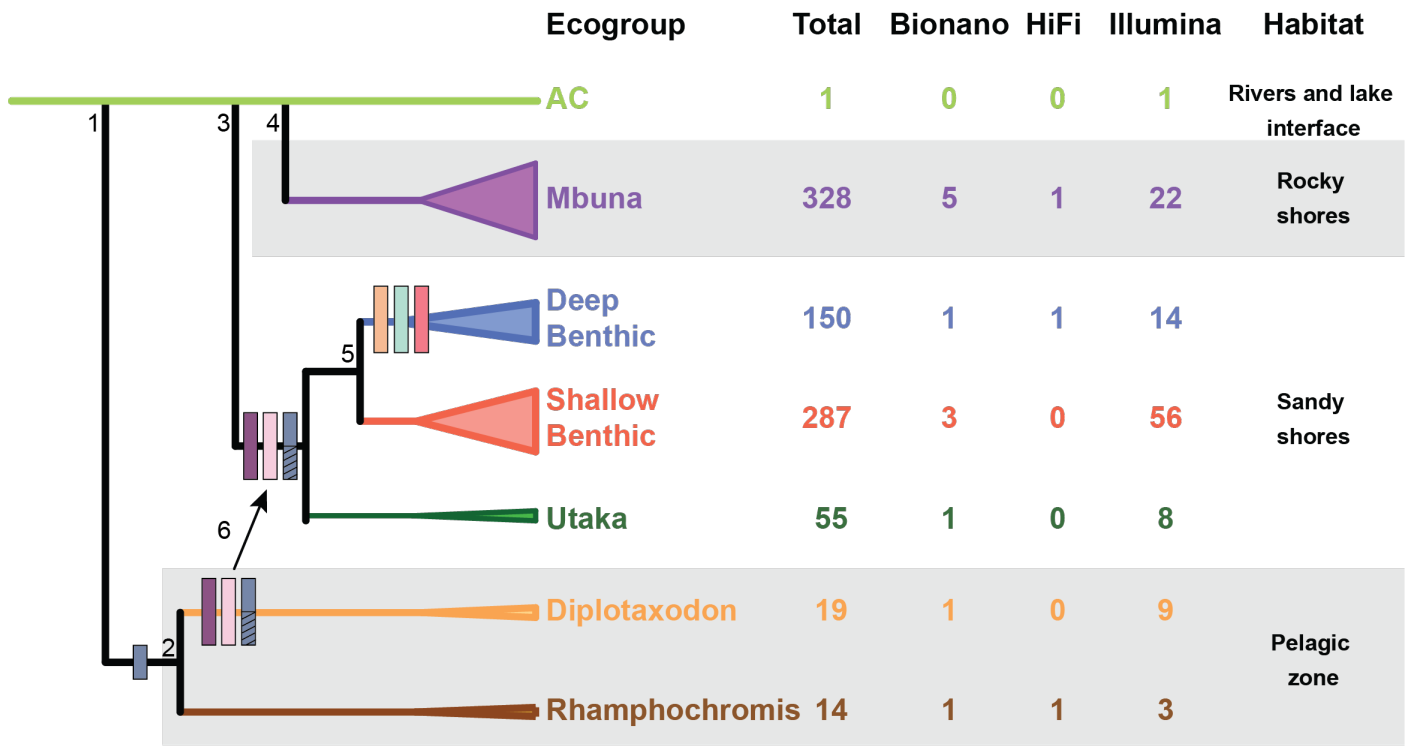
*Chromosomal inversions are an important class of genetic variation that link multiple alleles together into a single inherited block that can have important effects on fitness. To study the role of large inversions in the massive evolutionary radiation of Lake Malawi cichlids, we used long-read technologies to identify four single and two tandem inversions that span half of each respective chromosome, and which together encompass over 10% of the genome. Each inversion is fixed in one of the two states within the seven major ecogroups, suggesting they played a role in the separation of the major lake lineages into specific lake habitats. One exception is within the* benthic *sub-radiation, where both inverted and non-inverted alleles continue to segregate within the group. The evolutionary histories of three of the six inversions suggest they transferred from the pelagic* Diplotaxodon *group into* benthic *ancestors at the time the* benthic *sub-radiation was seeded. The remaining three inversions are found in a subset of* benthic *species living in deep waters. We show that some of these inversions are used as XY sex-determination systems but are also likely limited to a subset of total lake species. Our work suggests that inversions have been under both sexual and natural selection in Lake Malawi cichlids and that they will be important to understanding how this adaptive radiation evolved.*

## INTRODUCTION

Large genomic inversions are a particularly interesting type of genetic variation in the evolutionary process[1–4]. These rearrangements of DNA sequence strongly suppress recombination between inverted and non-inverted alleles, enabling the capture and accumulation of genetic variants of smaller molecular size that are linked together on each inversion haplotype[5]. These alternative haplotypes largely follow independent evolutionary paths within a species, accumulating new mutations that can affect phenotypes under natural and sexual selection. Structural rearrangements can be large, spanning multiple megabases of DNA containing hundreds of genes[2,6]. They have been shown to play a role in adaptive divergence within a variety of species[7–12], in creating alternative reproductive strategies and mating types[13–16], in evolution of sex chromosomes[1,17], and in the formation of pre- and post-zygotic barriers between incipient species[18,19].

When comparing the genomes of extant species, chromosomal inversions are often found as fixed differences, consistent with a role in speciation[20–22]. Inversions have been shown to play a role in adaptation to different environments and the establishment of prezygotic and postzygotic barriers[21,23]. However, the evolutionary forces responsible for establishing and spreading inversions remain obscure. Natural selection and sexual selection are thought to play a role but neutral forces like genetic drift can also be responsible for their spread. Perhaps most likely, all of these different evolutionary forces contribute to their fixation[24,25].

Few vertebrates offer as much potential for investigating evolution and speciation as the *Cichlidae* family of fish, one of the most species-rich and diverse families of vertebrates, with an estimated 2,000-3,000 species across the globe [26–28]. Evolutionary radiations have occurred at least three times in the African Great Lakes (Lake Malawi, Lake Tanganyika, and Lake Victoria). In Lake Malawi, more than 800 species of cichlids evolved in the past 1.2 million years[27,29–31]. Speciation in cichlid fishes involved extremely high levels of phenotypic divergence, including changes in body shape, jaw structures, and feeding behaviors related to prey in their ecological niches[26,32]. Additionally, traits that create prezygotic barriers between species are also extremely diverse. Color patterns are quite dramatic and diverse among species and include both differences in color and species-specific patterns including bars, stripes, and blotches that are thought to reinforce barriers between species in sympatry[33]. In some species, mating is seasonal, with gatherings on 'leks' where males build bowers by scooping and spitting out sand over the course of a week




1. Pelagic lineage splits from ancestral *ACs.*
2. *Diplotaxodon* and *Rhamphochromis* lineages split.
3. *Benthic* lineage splits from ancestral *ACs*.
4. *Mbuna* lineage splits from ancestral *ACs*.
5. *Deep benthics* and *shallow benthics* lineages split.
6. We propose that a hybridization event between ancestral *Diplotaxodons* and *ACs* preceeded the seeding of the benthic radiation, transferring the 9, 11, and 20 inversions between the *Diplotaxodons* and the new lineage. Alternatively, incomplete lineage sorting could be responsible for the violation of the species tree.

**Figure 1. Overview of Lake Malawi cichlids.** A summary of the Lake Malawi cichlids and the results from this paper. On the left, a phylogeny showing the seven major lineages that currently live in the lake which are referred to as ecogroups. AC is *Astatotilapia calliptera*, a riverine species that lives in the margins of the river/lake interface. For each ecogroup, we summarize the 1) estimated number of species in the lake, 2) the number of species we have Bionano optical read data for, 3) the number of species we have PacBio based genome assembies, 4) the number of species with Illumina short-read sequencing data, and 5) the primary habitat these ecogroups live in. Deep and shallow benthics are distinguished by the lake bottom depths where they live while utakas live in the water column but mate over the shore. We also summarize the distribution of 6 large inversions we have identified, colored by the chromosome they are found on (legend). We propose that the inversions on 9, 11, and 20 originated in the pelagic lineage and spread to the benthics by hybridization very early on in the evolution of the benthic lineage. Alternatively, these inversions distribution could be explained by incomplete lineage sorting.

or more[34]. Females choose mates based upon features of these bowers including shape and size[35]. Most of the Lake Malawi species can be induced to interbreed in the lab, providing the opportunity to use genetic approaches to study causality of associated changes. Despite the presence of these prezygotic barriers, substantial gene flow has occurred multiple times within the lake[30,36].

Our current understanding of how the Lake Malawi radiation created >800 species is a set of three serial diversifications, starting from a single riverine-like ancestor lineage, that created seven major ecogroups that currently live in the lake[30] (**Figure 1**). Each ecogroup forms its own clade in a whole-lake phylogeny and occupies distinct habitats within Lake Malawi. A pelagic lineage separated first, further diversifying into a deep-water ecogroup (*Diplotaxodon*) and a mid-water ecogroup (*Rhamphochromis*). A muddy/sandy benthic lineage evolved from the ancestral riverine lineage next, which split into three ecogroups

(collectively referred to as *benthics*) living either in the water column (*utaka*), over shallow sandy/muddy shores (*shallow benthics*), or in deep-water habitats (*deep benthics*). Finally, an ecogroup living over rocky habitats evolved (*mbuna*). The riverine generalist *Astatotilapia calliptera (AC)* living in the border regions of the lake and surrounding rivers, is the seventh ecogroup, and is thought to represent the ancestral lineage that seeded the three primary Lake Malawi lineages. Subsequent radiations within these ecogroups based on trophic specializations and sexual selection further diversified the species flock[30,32].

Here, we investigate the role large inversions played in the Lake Malawi radiation using three technologies. Using the Bionano Saphyr optical mapping system, we identify four single large inversions and two double inversions, ranging from 9.9Mb to 20.6Mb in size. A fifth single inversion, composed of just one of the two tandem inversions on chromosome 20, was also identified, indicating a serial set of rearrangements created this structural variant. We validate these inversions using new genome assemblies created using PacBio Hifi reads. Finally, we use existing and newly generated short-read Illumina sequencing to genotype these inversions in DNA samples from cichlids collected from Lake Malawi. These inversions primarily segregate by ecogroup, with no inversions found in the *mbuna* and *AC* samples, one inversion fixed in *Rhamphochromis*, three inversions fixed in *Diplotaxodon*, and all six segregating within the *benthic* sub-radiation. The evolutionary histories of these inversions are inconsistent with the species phylogeny and suggest that the three *Diplotaxodon* inversions spread into the *benthics* via hybridization around the time the benthic lineage diverged from the ancestral riverine species. The three additional inversions are found primarily in *benthic* species living in deep water habitats. We provide evidence that three of the inversions are involved in sex determination in a subset of *benthic* species. Our work provides a framework for understanding the role of inversions in the adaptive radiation of Lake Malawi cichlids, and supports future investigations into their roles in the establishment of ecogroups, adaptation to deep water habitats, and in controlling traits under sexual selection.

## RESULTS

### *Improvement of the* Metriaclima zebra *reference*

The *mbuna* cichlid, *Metriaclima zebra*, was the first species sequenced as a reference genome for Lake Malawi cichlids[37]. The current version of its genome, M_zebra_UMD2a, has been created by assembling PacBio reads into contigs which were anchored to 22 chromosomes using recombinant maps generated using crosses between different Lake Malawi species[38]. While the overall quality of this reference is high compared to other non-model organisms, a total of 20.5% of the DNA was not anchored to a chromosome. Additionally, the orientation of contigs can be ambiguous given how contigs were anchored to linkage groups using recombinant maps. These imperfections could complicate inversion discovery, so we first sought to improve the reference sequence.

The Bionano Saphyr system generates large (150 kb - multiple Mbs) DNA molecules and labels a specific motif (CTTAAG) with a fluorescent probe[39]. While these molecules do not provide sequencing information, the distance between observed fluorescent loci can be used to assemble multiple molecules into large maps, with lengths spanning up to an entire chromosome. Comparison of these maps with an *in-silico* digest of the reference sequence identifies large structural changes in the observed sample or errors in the reference genome. We processed DNA collected from the blood of three *Metriaclima zebra* individuals using the Bionano pipeline and assembled DNA molecules into optical maps. We identified many discrepancies between the observed optical maps and the M_zebra_UMD2a assembly (**Figure 2**). These differences fall into two primary categories: 1) contigs that were incorrectly ordered or oriented and 2) unplaced contigs that Bionano anchored to specific chromosomes. The first category is consistent with errors in the assembly caused by ambiguity in placing contigs into a linkage group using recombination maps. Our results indicate that 178 of the contigs are incorrectly oriented and 133 of the unanchored contigs could be placed onto specific chromosomes.

To further improve the reference, we used PacBio HiFi reads generated from an additional male *Metriaclima zebra* individual to create a new genome assembly, taking advantage of the low overall error rate (< 0.1%) of HiFi reads compared to previous generation PacBio technologies. Contigs generated from these data were combined with Bionano maps to generate a new hybrid genome assembly, which we refer to as the M_zebra_GT3a assembly. The overall quality of the assembly was excellent, with 933Mb of 962 Mb of DNA assigned to the 22 linkage groups, (**Table S1**) reducing the amount of unplaced contigs from 196 Mb to 28 Mb. The N50 length was 32.542Mb, with many of the scaffolds containing entire chromosomes.

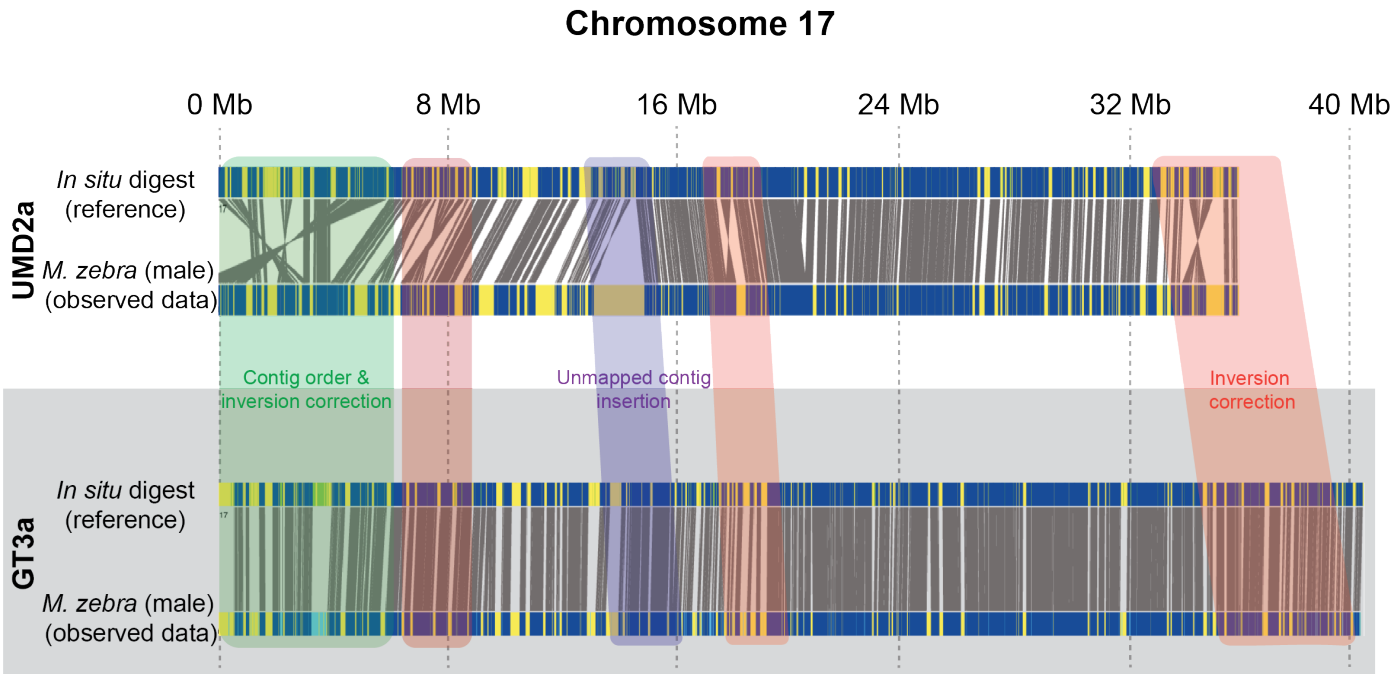



**Figure 2. Generation of a new high quality reference genome for *Metriaclima zebra*.** Alignment of Bionano maps derived from a single *Metriaclima zebra* male to two different reference sequences for Chromosome 17 (white and grey, respectively). M_zebra_UMD2a reference (white) is the existing reference genome and M_zebra_GT3a reference (grey) was created for this manuscript using a combination of PacBio HiFi reads and Bionano molecules. For each reference, the top bar shows the expected location of each CTTAAG motif (blue) created from an *in silico* annotation of the reference sequence. These are aligned (grey lines) to observed locations in maps generated from Bionano Saphyr optical molecules (bottom bar). Deviations from co-linearity indicate reference errors (green, blue, and red highlights). Yellow indicates regions of the genome that lack aligned motifs. A summary of the improvements in the new genome is found in **Table S1**.

*Identification of six large inversions from 11 species of Lake Malawi cichlids*

Armed with the improved reference, we used the Bionano system to test 31 samples from 11 different species found in Lake Malawi. Eight of the species were bred in the Streelman or McGrath laboratories (**Table 1 and Table S2**). These included four *mbuna* species (*Pseudotropheus demasoni*, *Cynotilapia zebroides*, *Labeotropheus fuelleborni, and Labeotropheus trewavasae),* two *shallow benthic* species (*Mchenga conophoros* and *Nyassachromis prostoma* "Orange Cap"), a single *deep benthic* species (*Aulonocara sp. 'chitande type north' Nkhata Bay*), and a single *utaka* species (*Copadichromis virginalis)*. To expand our coverage of the Lake Malawi radiation, we also obtained samples from the pet trade for a single species of the *Diplotaxodon* ecogroup, *Diplotaxodon limnothrissa,* and a single species of the *Rhamphochromis* ecogroup, *Rhamphochromis longiceps*. Finally, we also tested a single male *Protomelas taeniolatus* *shallow benthic*

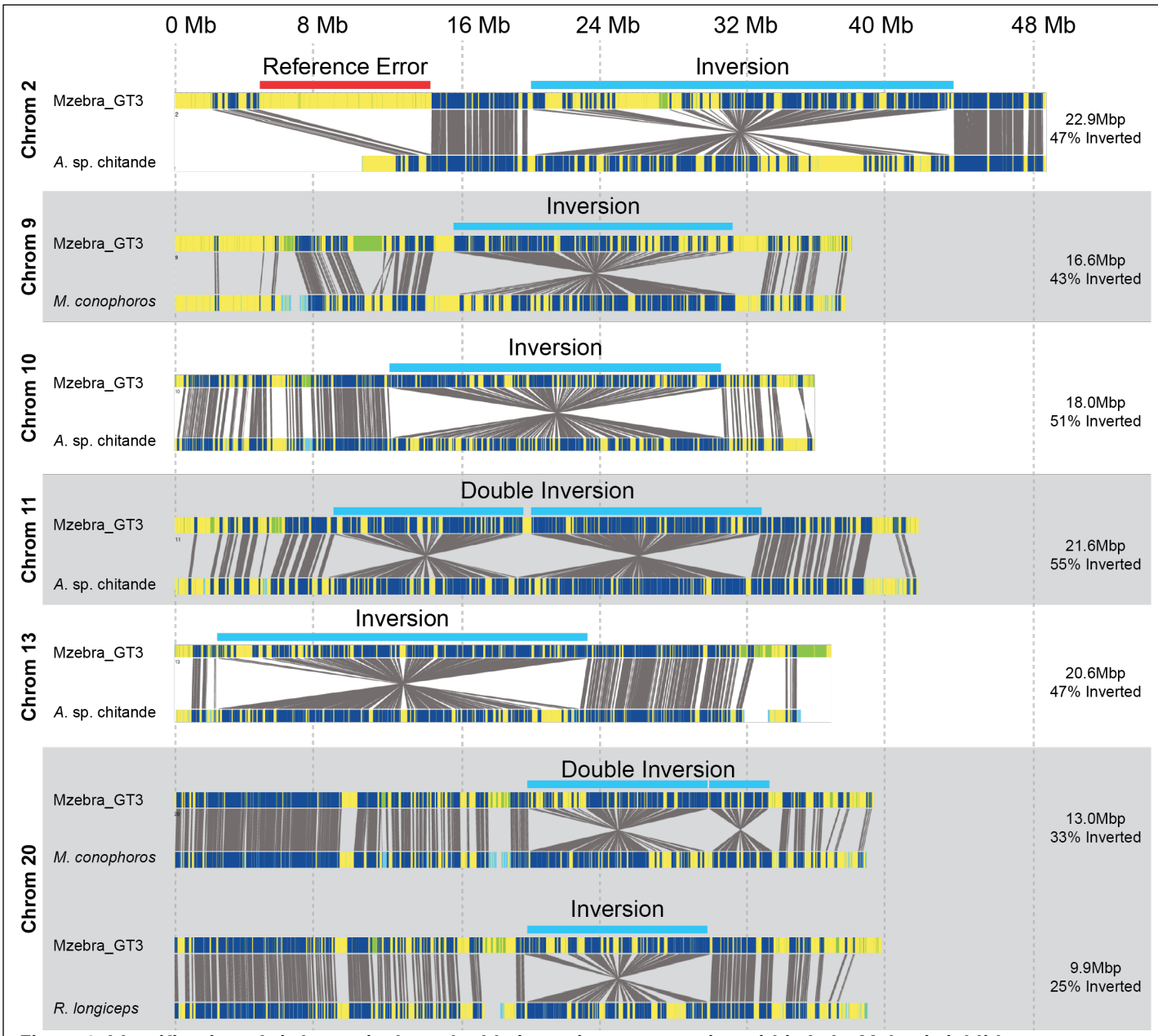



**Figure 3. Identification of six large single or double inversions segregating within Lake Malawi cichlids.** Representative alignments of inversions or double inversions identified from blood samples obtained from 30 individuals from eleven species. For each inversion, the top shows predicted CTTAAG motifs from an *in-silico annotation* of the reference genome, the bottom shows motifs identified using Bionano molecules obtained from an individual of the species indicated on the left, and the grey lines indicate matching motifs based upon predicted and observed distances. Yellow indicates regions containing no matches between motifs in the reference genome and the mapped sample. Single inversions were identified on chromosomes 2, 9, 10, 13, and 20. Tandem double inversions were identified on chromosomes 11 and 20. The estimated length and percentage of the chromosome spanned by the inversion is shown on the right. The single inversion on 20 identified in *Rhamphochromis longiceps* has the same position as the first inversion of the double inversion on 20 identified from *Mchenga conophoros*. A list of all samples and their inversion genotype is in Table 1. Note that an error or polymorphism in the reference genome in chromosome 2 is indicated in the first panel.

individual collected from Lake Malawi based on its position in principal component analysis of SNVs described below. In total, we were able to obtain samples from 6 of the 7 ecogroups, excluding *Astatotilapia calliptera*.

From these 31 individuals, we identified six new inversions ranging from 9.9 - 22.9 Mb in size (**Figure 3**): single inversions on chromosomes 2, 9, 10, and 13, and double inversions on chromosomes 11 and 20.

153 Interestingly, regarding the double inversion on chromosome 20, we also identified four *Rhamphochromis*
154 individuals that carried only the first inversion, lacking the 4.1 Mb long second part of this rearrangement
155 (**Figure 3**). This indicates that the tandem inversions on 20 were formed by at least a two-part process,
156 most parsimoniously with the left arm of the inversion (20a) forming first followed by the right inversion
157 (20b) occurring in that genetic background.

158 The distribution within these species is summarized in **Table 1**. For simplicity, we will refer to all six
159 rearrangements as inversions, regardless of whether they are single or double inversions. The most
160 common inversions were on chromosome 9, 11, and 20 which were found in a combination of *benthic* and
161 *Diplotaxodon* individuals. The position of these three inversions were also consistent with regions of low
162 recombination rate previously identified in an intercross between a *mbuna* and *benthic* individual[38], as
163 expected, since inversions suppress recombination between inverted and non-inverted haplotypes. The
164 remaining three inversions (2, 10, and 13) were only found in the deep benthic individuals (*Aulonocara sp.*
165 *'chitande type north' Nkhata Bay*). While for most samples the inversion genotype was homozygous, for
166 five individuals the optical maps supported both inverted and non-inverted haplotypes, indicating the
167 inversions were heterozygous in those samples (**Figure S1**).

168 de novo *reference genomes support the presence of these six inversions*

169 To provide further support for these structural rearrangements, we generated *de novo* hybrid genome

***Table 1***. *Inversion genotypes of 31 samples based on Bionano data.*

| | | | Inversion | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Species*** | **Ecogroup** | **#** | **2** | **9** | **10** | **11** | **13** | **20** | **Notes** |
| *Cynotilapia zebroides* 'Cobue' | *Mbuna* | 1m/1f | - | - | - | - | - | - | |
| *Labeotropheus fuelleborni* | *Mbuna* | 1m/1f | - | - | - | - | - | - | |
| *Labeotropheus trewavasae* | *Mbuna* | 1m/1f | - | - | - | - | - | - | |
| *Metriaclima zebra* | *Mbuna* | 3m/1f | - | - | - | - | - | - | Reference species |
| *Pseudotropheus demasoni* | *Mbuna* | 1m/1f | - | - | - | - | - | - | |
| *Copidachromis virginalis* | *Utaka* | 1m/2f | - | **X** | - | -* | - | **X** | One female was het on 11 |
| *Protomelas taeniolatus* | *Shallow benthic* | 1m | - | **X*** | - | **-** | - | **X** | The male was het on 9 |
| *Mchenga conophoros* | *Shallow benthic* | 1m/2f | - | **X** | - | **X** | - | **X** | |
| *Nyassochromis prostoma* 'Orange Cap' | *Shallow benthic* | 1m/1f | - | **X** | - | **X*** | - | **X** | The male was het on 11 |
| *Aulonocara* sp. 'chitande type north' *Nkhata* Bay | *Deep benthic* | 3m/1f | **X** | **X** | **X*** | **X** | **X** | **X** | All three males were het on 10 |
| *Diplotaxodon limnothrissa* | *Diplotaxodon* | 2m/0f | - | **X** | - | **X** | - | **X** | |
| *Rhamphochromis longiceps* | *Rhamphochromis* | 3m/1f | - | - | - | - | - | **X*** | All samples had the single left inversion on 20 |

assemblies of three *Aulonocara sp. 'chitande type north' Nkhata Bay* individuals, which carry all six of the identified inversions, and two additional *Metriaclima zebra* individuals using a combination of PacBio HiFi reads and Bionano maps. The overall quality of these assemblies is high, with large N50 values and a small number of individual contigs (**Table S3**). We performed whole genome alignments between the five new assemblies and the GT3a reference genome (**Figure S2-S7**). These alignments show that the new assemblies support the presence of all six inversions in the *Aulonocara sp. 'chitande type north' Nkhata Bay* individuals. For most inversions, a single scaffold covered the entire inversion with breakpoints in locations consistent with the Bionano maps. The two new control assemblies were consistent with the M_zebra_GT3a reference and did not show any evidence of inversions in those regions. In other regions of the genome, however, there were differences in the alignment of the M_zebra_GT3a assembly and the *Metriaclima zebra* genomes, most prominently in the left arm of chromosome 2 (**Figure 3**) and in the highly repetitive chromosome 3, which likely indicates some errors in the M_zebra_GT3a assembly or that a large polymorphism segregates within the species.

We refined the location of the end points for each inversion using the genomic alignments. However, the presence of a large amount of repetitive DNA in regions near these regions prevented us from defining inversion boundaries at single base pair resolution. The mechanisms that create large inversions often involve repetitive sequence near the breakpoints, making exact localization difficult if not impossible[40].

We also took advantage of a genome assembly for a *Rhamphochromis sp 'chilingali'* individual (fRhaChi2.1) created by the Darwin Tree of Life Project produced using PacBio data and Arima2 Hi-C data (Accession PRJEB72870). The genome alignment between the fRhaChi2.1 assembly and the GT3a assembly confirmed a single inversion on chromosome 20, consistent with the Bionano data (**Figure S2-S7)**.

For each inversion, we determined the ancestral haplotype using two species as outgroups, *Oreochromis niloticus* and *Pundamilia nyererei*. *Oreochromis niloticus,* or Nile tilapia, is an important food fish that's estimated to have diverged from Lake Malawi cichlids 14.1 to 30 million years ago while *Pundamilia nyererei* is a more closely related haplochromine found in Lake Victoria[37,41]. For this analysis we used a published whole genome assembly for *Oreochromis niloticus* (UMD_NMBU)[38] and generated a hybrid genome assembly for *Pundamilia nyererei* using a previously published whole genome assembly[37] combined with Bionano maps we produced from a single individual (**Figure S8**). For all six inversions, the reference haplotype found in the GT3a assembly represents the ancestral state.

*Inference of the segregation of these inversions in Lake Malawi using short-read sequencing*

The distribution of these inversions within the lake can provide important information to their function. Genotyping inversions using short read sequencing is possible because each haplotype is on an independent evolutionary trajectory due to the suppression of recombination. Population genetics approaches take advantage of the fact that inversions create patterns of inheritance between SNVs that are readily detected by approaches such as Principal Component Analysis (PCA)[42].

We analyzed short read sequencing data from the 31 samples with Bionano data along with 297 wild individuals published by the Durbin lab[30], 15 wild individuals published by the Streelman lab [31,43], and 22 wild individuals sequenced for this paper, aligning reads and calling small variants against our new M_zebra_GT3a reference genome. PCA analysis was used to analyze the major axes of variation across the whole genome and within the six inversions. We separately analyzed the left and right inversions on 11 and 20. For the whole genome, the species within an ecogroup clustered together within the first two principal components, reflecting their evolutionary history. This matched the previously reported results of Malinsky et al[30]. Individuals from four of the ecogroups - *Rhamphochromis*, *Diplotaxodon*, *Astatotilapia calliptera*,

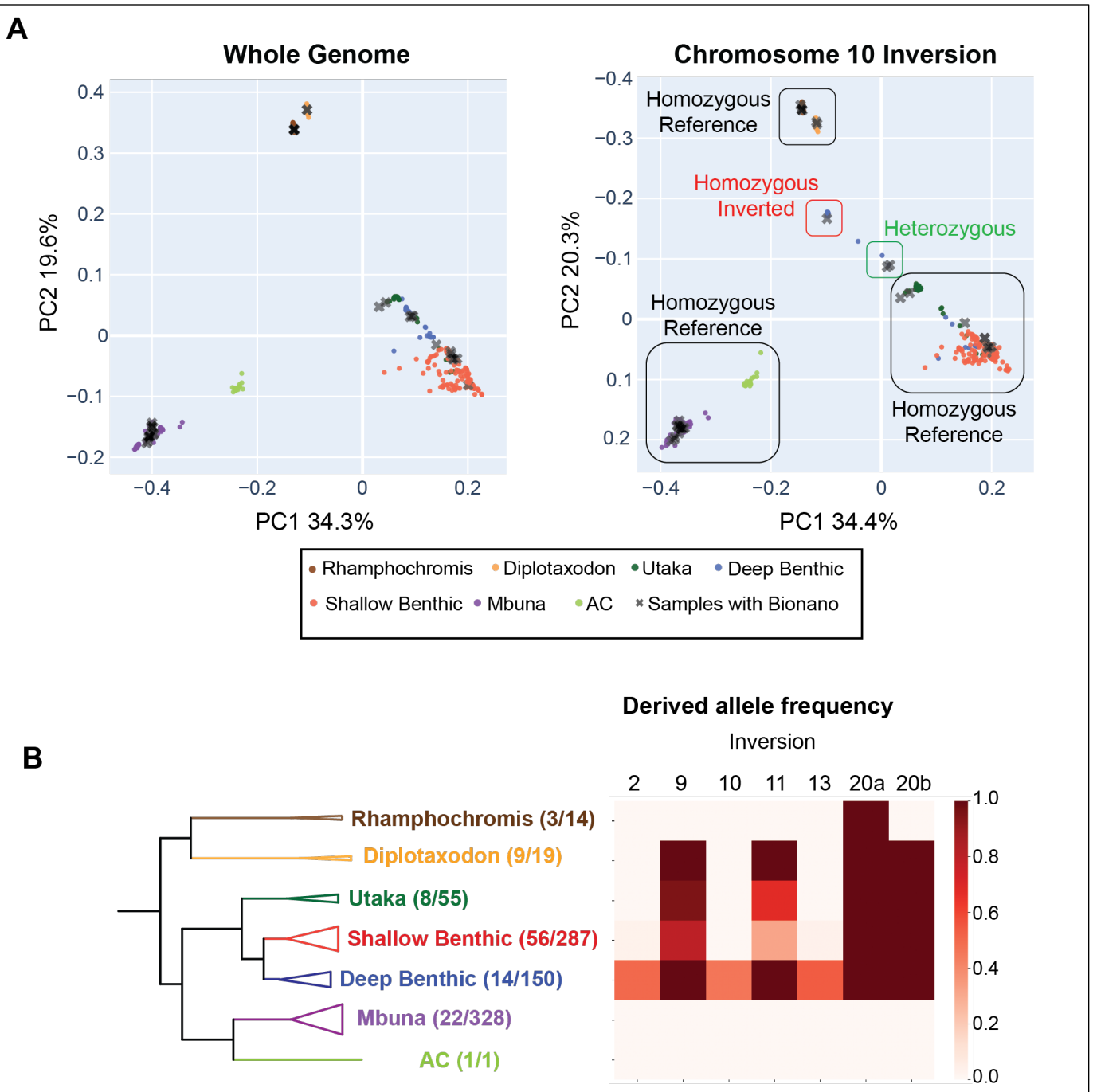



**Figure 4. Distribution of inversions within the Lake Malawi ecogroups.** Principal component analysis was used to analyze SNVs distributions in 365 samples within each inversion to genotype the inversion haplotype for all six inversions. (**A**) We illustrate the approach for genotyping the chromosome 10 inversions. The PCA plot for the entire genome is shown on the left, with each sample colored by its ecogroup. Samples from each ecogroup cluster together, reflecting their evolutionary history. Individuals with Bionano data are shown as grey Xs. On the right, for the SNVs that fall within the chromosome 10 inversion, the deep benthic/shallow benthic and utaka individuals split into multiple clusters. Using the samples genotyped with Bionano data, we assign each cluster to an inversion genotype, which is shown by the black, green, and red boxes. To make comparisons of the whole genome PCA and chromosome 10 PCA plots easier, we reversed the y-axis on the right panel. The assigned genotypes of each sample can be found in Supplemental Table 2. Interactive PCA plots for each inversion are included in **Figures S9-S17**. (**B**) The derived allele frequency was calculated for each inversion within the seven ecogroups. The number of species that were genotyped and the estimated number of species that live in Lake Malawi are listed in the whole-genome phylogeny (left).

and *mbuna* - formed four distinct clusters (**Figure 4A and Figure S9**). A fifth, more diffuse cluster, composed of *shallow benthics*, *deep benthics*, and *utakas*, was also present. The broader distribution of

the *benthic* ecogroups in PC space is consistent with both their recent separation as well as large amounts of gene flow thought to have occurred between these different species[30,36].

The PCA plots using the SNVs within the six inversions often violated the pattern of the whole genome PCA (**Figure 4A** and **Figure S10-S17**). Using samples that were sequenced with both Illumina and Bionano technologies, we assigned each cluster to specific inversion genotypes (**Table S2**). We did not have Bionano data for two clusters (one on chromosome 9 and one on 11 – see **Figures S11, S13, and S14**), so we supplemented our dataset by adding Bionano data from an additional *shallow benthic* individual (*Protomelas taeniolatus*) that fell into both clusters. The PCA clusters varied in their complexities for each inversion depending largely on their distribution within the *benthic* ecogroup (more on this below) and included both homozygous genotypes and heterozygous genotypes. We summarize the overall distribution of the inversions within each of the ecogroups in **Figure 4B**.

All specimens of the *mbuna* and *AC* ecogroup were fixed for the homozygous non-inverted genotype for all six inversions. In *Rhamphochromis*, all the inversions were homozygous non-inverted except for the left arm of 20 which was fixed for the single inversion state. The *Diplotaxodon* individuals all carried the non-inverted alleles for the inversions on 2, 10 and 13 and were fixed for the inversions on 9, 11, and 20.

The distribution of these inversions was the most complicated within the *benthics*. The inversions on 2, 10, and 13, which were identified in *Aulonocara sp. 'chitande type north' Nkhata Bay*, were mostly found together in other *deep benthics*, including the *Alticorpus* (*geoffreyi*, *macrocleithrum*, *peterdaviesi*), *Aulonocara* (*blue chilumba*, *gold* and *minutus)*, and *Lethrinops* (*gossei, longimanus 'redhead', and sp. olivera*) genera. The inversions on 2 and 13 were also found in the shallow benthic species *Placidochromis longimanus*. However, not all the *deep benthics* carried the three inversions, including five *Aulonocara* species (*baenschi*, *gertrudae*, *steveni*, *stuartgranti*, and *stuartgranti* Maisoni*)* and a single *Lethrinops* species (*longipinnis*). The distribution of these three inversions in *benthic* species known to inhabit the deepest habitats are suggestive for a role in depth adaptation.

Both inversions on 20 were homozygous inverted in all benthic individuals. For the single inversion on 9, many *benthics* were homozygous for the inverted haplotype. However, an additional PC cluster was present, composed of 64 *shallow benthic* and *utaka* individuals heterozygous for the non-inverted haplotype (**Figure S1**). Interestingly, no *benthic* individuals were identified as homozygous for the non-inverted state.

Finally, the distribution of the chromosome 11 inversion was also determined. Individuals of all three genotypes (homozygous non-inverted, homozygous inverted, and heterozygous) were observed (**Figure 4a**).

*Evolutionary history of 9,11, and 20 suggest introgression from* Diplotaxodon *into benthic ancestors*

The distribution of these inversions conflicts with the species tree, suggesting that these inversions could have spread to the *benthics* via hybridization or carried forward by incomplete lineage sorting. To determine their evolutionary history, we built phylogenies for the whole genome as well as for each inversion using SNVs that fell within each of the six inversions using representative species from each ecogroup (73 total – **Table S2**) as well as two previously sequenced *Pundamilia nyererei* individuals as an outgroup (**Figure 5A** and **Figures S18 – S25**)[44,45]. To avoid issues with haplotype phasing, we focused on individuals that were homozygous for one of the two inverted genotypes.

While the whole genome phylogeny matched the expected topology based upon our understanding of the Lake Malawi species tree, the topologies of each of the inversions showed key differences between the separation of ecogroups. The *benthic* group often split into separate clades, correlated with their inversion genotype. Prominently, we found the three inversions on 9, 11, and 20 showed key differences in the

relationship between the *Diplotaxodon* and *benthic* ecogroups. While *Diplotaxodon* is a sister group to *Rhamphochromis* at the species level, for these three inversions, the *Diplotaxodon* individuals formed a clade with the *benthics* that also carried the inversion. Additionally, genetic distance between *Diplotaxodon* and *benthics* carrying the inversion was also much smaller than the rest of the genome (**Figure 5B**). Both the phylogenies and genetic distances between ecogroups are consistent with the

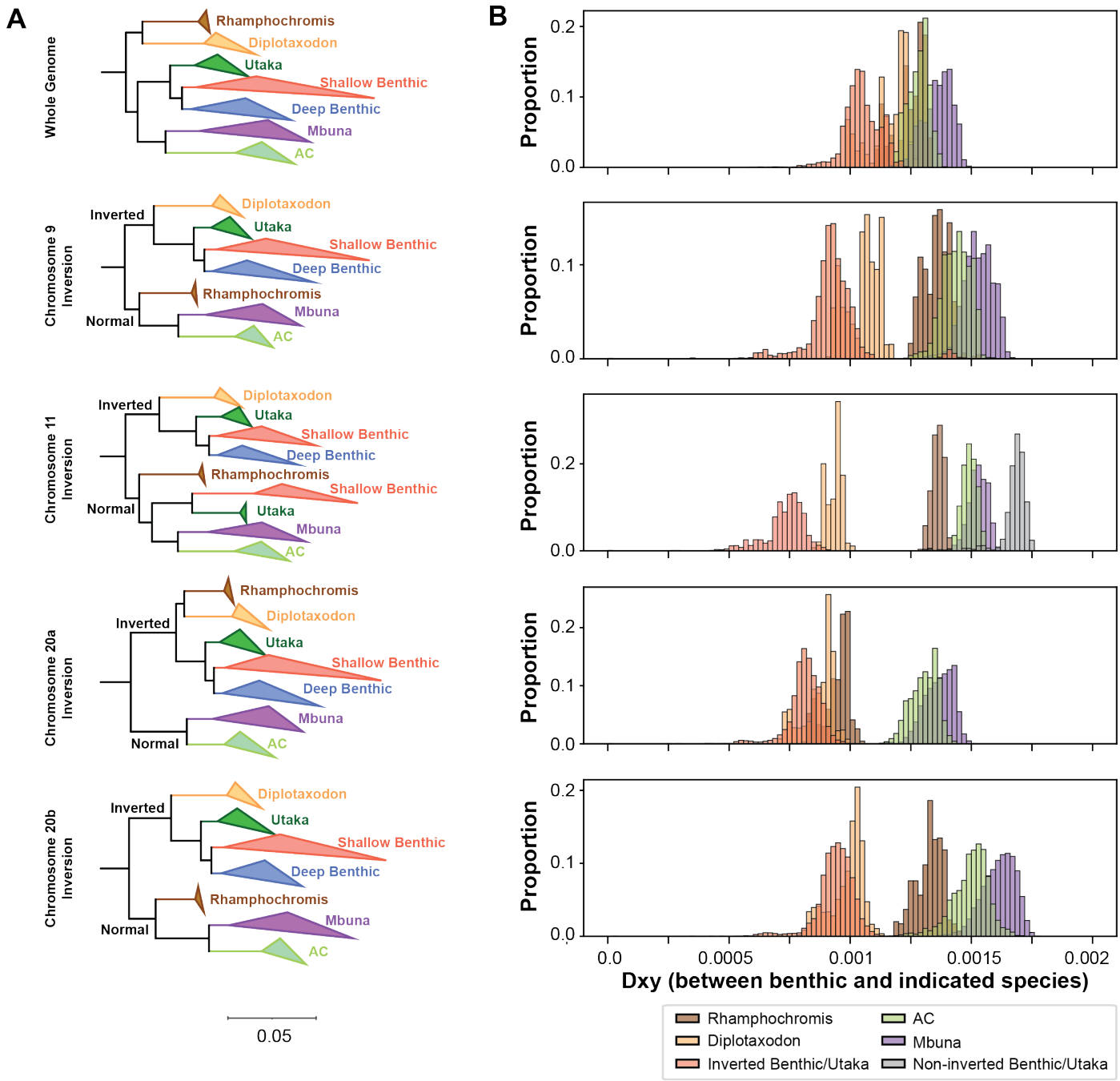



**Figure 5. Evolutionary history of large inversions violates species tree.** (**A**) Maximum likelihood trees built from either whole genome SNVs (top) or within the 9, 11, and 20 inversions. Between 2-25 representative samples were selected for each ecogroup (see **Table S2**) and individuals that were heterozygous were excluded from this analysis. Clades were collapsed by ecogroup for visualization purposes. Full phylogenies are presented in **Figures S18** to **S25**. The whole genome tree indicates *Diplotaxodon* and *Rhamphochromis* species split first and formed their own clade. For each of the displayed inversions, the first split was between samples carrying the inverted vs. non-inverted genotypes. Additionally, the *Diplotaxodon* individuals now formed a clade with the *benthics* that carry the inversion. (**B**) Density plots of $d_{XY}$ values comparing *benthic* species and species of the indicated ecogroup. For the displayed inversions, the *Diplotaxodon* species were much closer in evolutionary distance when compared to the whole genome plot.

introgression of the 9, 11, and 20 inversions from the *Diplotaxodon* to *benthic* ancestors at the time of the benthic radiation.

*A role for large inversions in sex determination within the* benthic *ecogroup*

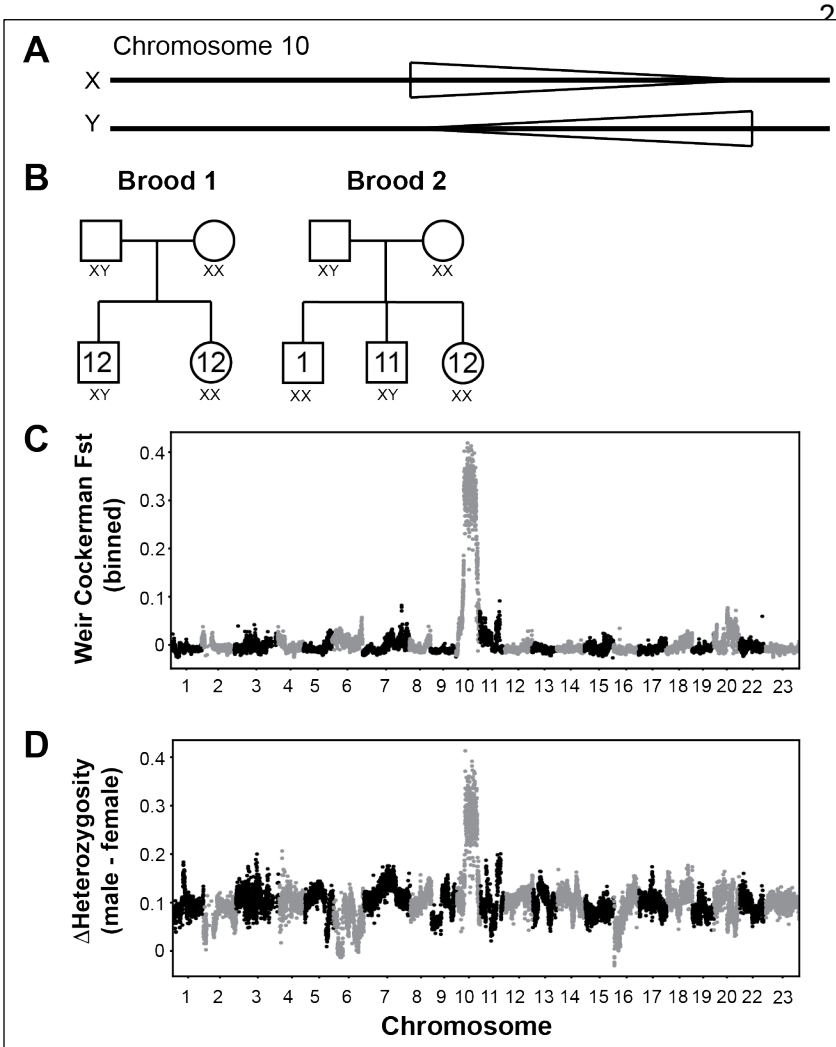



**Figure 6. A role for the chromosome 10 inversion in sex determination** (**A**) Schematic of the two inversion states (normal and inverted) and their assignment as X or Y chromosomes. (**B**) 24 offspring (12 male and 12 female) from two separate male/female pairs were genotyped for the inversion. For 47 of 48 individuals, the XY animals were male and the XX animals were female. One male was XX. (**C**) $F_{ST}$ analysis comparing 24 males and 24 females identified chromosome 10 as the primary genetic region that segregated with sex. (**D**) Comparison of heterozygosity levels of 24 male and 24 female animals also identified chromosome 10 as an outlier.

The evolution of sex determination is often associated with the presence of inversions to repress recombination between the sex chromosomes[46,47]. We tested whether these inversions could play a role in sex determination in *benthic* species, as we identified individuals that were heterozygous for some of these inversions (**Table 1**). We started with the inversion on 10, as our Bionano data identified three male *Aulonocara sp. 'chitande type north' Nkhata Bay* individuals that were heterozygous and a single female that was homozygous for the inverted haplotype. To test whether this inversion is linked to a genetic locus (or loci) that act as an XY system in this species, we tested two separate broods (Brood 1; 2 parents, 24 offspring and Brood 2: 2 parents, 24 offspring) currently growing in the laboratory, sequencing 12 males and 12 females from each brood with Illumina paired-end sequencing. We used PCA to genotype each of the 52 animals (4 parents and 48 offspring) for the chromosome 10 inversion genotype. The association in the offspring between the inversion 10 genotype and sex was almost perfect: 24 females and 1 male were homozygous for the inversion, while 23 males were heterozygous for the inversion (**Figure 6B**). These individuals were sexed by venting so the imperfect segregation could either indicate an error in sexing or that other modifiers influence sex.

We also used binned Weir Cockerham $F_{st}$ analysis, grouping male and female individuals to scan the genome for regions that were associated with sex (**Figure 6C**). The inversion on 10 was the region of the genome that was in strongest linkage disequilibrium with sex. Males also had a much higher heterozygosity in this region than females (**Figure 6D**). Altogether, these results indicate that the inversion on 10 acts as an XY sex determiner in the *Aulonocara sp. 'chitande type north' Nkhata Bay* species.

We also tested eight laboratory *Nyassachromis prostoma* "Orange Cap" individuals, as our Bionano data identified a male that was heterozygous for the inversion on 11 and a female that was homozygous for the inversion. We performed short-read sequencing on three males and five females growing in the lab, genotyping the inversion using PCA analysis. All three males were heterozygous for the inversion while the five females were homozygous, a significant association between genotype and sex (Fisher exact test p-value = 0.0179). This suggests a separate XY system segregates in *Nyassachromis prostoma* "Orange Cap", linked to the inversion on chromosome 11.

## DISCUSSION

Here, we characterize the presence of large inversions found in the Lake Malawi cichlid flock, identifying six inversions segregating within the lake, providing a framework to address the role structural variants played in this lake's adaptive radiation (**Figure 6**). What are the evolutionary forces responsible for their spread and what role do they currently play in the lake? Inversions have multiple, well-characterized roles in adaptive evolution: 1) they can capture beneficial alleles responsible for adaptation to local habitats, 2) they can capture alleles that create prezygotic and postzygotic barriers that prevent gene flow during speciation, and 3) they can capture alleles that influence sex-specific phenotypes that are linked to sex-determination systems[1,3,4]. These inversions could have also spread through neutral processes like genetic drift or through other adaptive processes not listed here.

This question is much more complicated in the context of the Lake Malawi cichlid radiation with ~800 different species. The role of these inversions must be considered in a species- and time-specific way. The recent history of the lake was quite dynamic, with large cyclical changes in water depth and shoreline in the past 800,000 years that are predicted to create alternating periods of extinctions and radiations[48–50]. Similarly, the role the inversions play in each species (such as sex determination) will not be the same. Determining whether these inversions carry alleles that contribute to the phenotypic diversity found in Lake Malawi will be an important first step towards resolving the evolutionary forces responsible for their spread through the lake. This is possible as most species in the lake can interbreed, allowing for quantitative genetics approaches to link genotype variation to phenotypic variation and determine what traits are influenced by the inversions.

The local adaptation mechanism is one well-characterized method to explain how inversions can spread through adaptive evolution[6,8]. Under this hypothesis, two populations undergoing adaptive divergence in the presence of migration and gene flow create conditions that can select for inversions, provided they capture and link adaptive alleles together that increase fitness to their local environments[51]. Apart from the *benthics*, the inversions are fixed in each ecogroup in the inverted or non-inverted state, each of which live in unique habitats in the lake that place very different evolutionary pressures on the animals that live within them. In this scenario, the inversions would capture adaptive alleles that increase fitness of ancestral species to pelagic and deep-water environments. To illustrate this scenario, we consider the time the pelagic groups split from the ancestral AC fish that is thought to live in rivers and the borders of the lake (**Figure 1**). If the first inversion on 20 captured alleles that allowed these pelagic ancestors to inhabit the lake, it could be selected on provided that these pelagic ancestors continue to interbreed with the AC ancestors. Similarly, the diversification of the *Rhamphochromis* and *Diplotaxodon* groups could have involved the inversions on 9, 11, and right arm of 20, and the adaptation of *benthics* to deeper water habitats could have involved the inversions on 2, 10, and 13. We do note that the existing theory for recruiting inversions to local adaptation does not account for multiple unlinked inversions (i.e. 9, 11, and right arm of 20). Future experimental and comparative genomics approaches can determine if local adaptation was at play at the time these lineages separated.

We also provide evidence that two inversions play a role in sex determination in the *benthics*, with the strongest evidence showing the inversion on 10 acts as an XY system in a deep benthic species.

Additionally we show that the inversion on 11 is significantly associated with sex determination in a shallow benthic species as an XY system. Despite the importance of the maintenance of distinct, reproductively compatible sexes, sex determination systems can be evolutionarily labile, and Lake Malawi cichlids show exceptional variability, with at least 5 different genetic sex determination systems described to date[52,53]. While we implicated two inversions in sex determination, our work also underscores how dynamic sex determination is within the *benthics*. While many benthic *species* carry the different inversions, their genotypic distribution in most species is inconsistent with a role for sex determination. Effectively, the genetic determination of sex determination largely must be determined in a species by species way, and individual species may be segregating multiple sex determination systems. The selective forces on sex chromosomes are strong, including balancing selection for sex ratio and sexually antagonistic selection on sexually dimorphic phenotypes. Lake Malawi cichlids show amazing sex differences, characterized by diversification in numerous phenotypes, including pigmentation, behavior, and morphology thought to contribute to species barriers. The use of inversions in sex-determination allows for the accumulation of alleles in linkage with the sex-determiner gene which will have sex-specific effects in the male individuals that carry the Y chromosomes. The presence of these inversions in various ecogroups could simply be remnants of their role in sex determination in the ancestors of pelagic ecogroups, and have nothing to do with local adaptation, which could have contributed to their spread before sex chromosome turnover ended balancing selection on the inverted/non-inverted haplotypes and subsequent fixation of these inversions. Identification of the causal alleles for sex determination on these inversions will help us to resolve this question.

The presence of non-inverted haplotypes in the *benthics* could have been retained from the time the lineage initially formed, with both haplotypes under balancing selection if they were used for sex determination. A recent preprint, however, which also identified these 6 inversions presented here using multiple technologies, also utilized short-read sequencing of 1375 individuals to create an alternative model[54]. They inferred the non-inverted haplotypes of the 9 and 11 inversions were introgressed back into the *benthics* from riverine species within and outside of the lake. Similarly, the non-inverted haplotypes on chromosomes 2, 10, and 13 in the *deep benthics* were introgressed from *shallow benthics*. In this model, it is unclear what evolutionary forces drove these back introgressions of the non-inverted alleles, but would imply these inversions have not continuously provided the genetic sex determination mechanism since their inception.

The distribution of the inversions in the benthic sub-radiation is uniquely complicated in Lake Malawi. Inverted and non-inverted haplotypes for five of the six inversions are segregating within this group. Additionally, the presence of the 9, 11, and 20 inversions within *benthics* violates the whole genome phylogeny. We propose these three inversions originated within the *Diplotaxodon* group and spread to a very early *benthic* ancestor via introgression. We favor this over incomplete lineage sorting, which would also lead to violations of the phylogeny, for a few reasons. First, the dynamic history of the lake levels is consistent with times of extinction and hybridization[48–50]. Second, genomic analysis of extant cichlids provided evidence for a large amount of hybridization events[36,55]. Finally, the close genetic distance between *Diplotaxodons* and *benthics* that carry the inverted haplotype (**Figure 5**) is consistent with hybridization. Hybridization has been proposed to fuel adaptive radiations of cichlids in all three great African lakes (Malawi, Tanganyika, and Victoria) [27,28,56] and our work suggests that an ancient hybridization event between early *benthic* and *Diplotaxodon* ancestors occurred prior to the benthic radiation. It will be interesting to study the role of these three inversions in *benthic* species. *Diplotaxodon* are pelagic, deep-water dwellers that feed on plankton or small fish while *benthics* live near the shore over sandy and muddy habitats[30]. It is not obvious how the traits of the *Diplotaxodon* would benefit *benthic* animals that live in such dissimilar environments. However, the initial habitats that the *benthic* ancestors utilized is not known and potentially their initial habitat was in deeper waters than the *AC* ancestors.

404 This report provides an initial framework for understanding the role of inversions in the adaptive radiation
405 of Lake Malawi cichlids. Future work identifying and characterizing causal alleles captured within these
406 inversions will enable us to discern the specific mechanisms by which they contribute to speciation using
407 this remarkable evolutionary system as a model for the evolution of all animals.

408 **ACKNOWLEDGEMENTS**

409 We thank Krish Roy for the acquisition of the Bionano Saphyr system, Manasi Pimpley for assistance in
410 adapting existing Bionano kits to cichlid tissue, and the Molecular Evolution Core Laboratory at the Parker
411 H. Petit Institute for Bioengineering and Bioscience at the Georgia Institute of Technology for the use of
412 their shared equipment, services, and expertise. This work was supported in part by NIH R35 GM139594
413 and the Nelson and Bennie Abell Professorship to P.T.M., R01GM144560 to J.T.S., and U.S. National Science
414 Foundation (DEB-1830753) to T.D.K..

415 **METHODS**

416 **Samples.** Wild-caught cichlids were acquired from Old World Exotic Fish Inc in 2022 and Cichlidenstadel
417 in 2024. Lab-reared species were purchased from Southeast Cichlids, Old World Exotic Fish Inc., and
418 Cichlidenstadel at various points in the past two decades (**Table S2**). Fish were delivered live to the Georgia
419 Institute of Technology cichlid aquaculture facilities. All samples were collected following anesthetization
420 with tricaine according to procedures approved by the Institutional Animal Care and Use Committee
421 (IACUC protocol numbers A100029 and A100569).

422 Caudal fins were collected from subjects and flash frozen on powdered dry ice before storage at -80°C.
423 Care was taken to minimize freeze-thaw cycles prior to DNA extraction via Bionano or Qiagen protocols. If
424 necessary, fins were sectioned on a sterile aluminum block set in dry ice prior to digestion.

425 Blood was collected following rapid decapitation of anesthetized subjects. Wide-bore, low-retention
426 pipette tips, 0.5M EDTA (pH 8.0, Invitrogen Cat# 15575-038), and Bionano Cell Buffer (Part Number 20374)
427 prevented blood from clotting at room temperature. Samples were used immediately for Bionano DNA
428 extraction (see below).

429 **Bionano.** Bionano DNA extractions were predominantly performed on fresh whole blood samples (see
430 above). Blood concentration was determined via a manual logical count performed on a hemocytometer
431 and 2 million cells were carried forward for DNA extraction with the Bionano SP-G2 Blood & Cell Culture
432 DNA Isolation Kit (Part Number 80060) as per the Bionano Prep SP-G2 Fresh Human Blood DNA Isolation
433 Protocol (Document Number CG-00005, Rev C). Because nucleated cichlid blood had high
434 concentrations, steps 3 and 4 were skipped and samples were resuspended in Bionano Cell Buffer to a
435 total volume of 200µL. Following extraction, samples were allowed to homogenize undisturbed at room
436 temperature for at minimum 72 hours prior to quantification using the Qubit dsDNA Quantitation, Broad
437 Range kit (ThermoFisher, Cat # Q32853).

438 Bionano data for a single *Protomelas taeniolatus* sample (PT_2003_m, see **Table_S2**) was generated from
439 fin tissue using a combination of the Bionano Prep SP Tissue and Tumor DNA Isolation Protocol
440 (Document Number 30339, Rev A) and the Bionano Prep SP-G2 Fresh Human Blood DNA Isolation
441 Protocol.

442 750ng of homogenized DNA was fluorescently labeled using the Bionano Direct Label and Stain-G2 (DLS-
443 G2) Kit (Part Number 80046, Protocol - CG-30553-1, Rev E). Labeled DNA was quantified using the Qubit
444 dsDNA Quantitation, High Sensitivity kit (ThermoFisher, Cat # Q32854) and loaded onto a flow cell within
445 a Bionano Saphyr Chip G3.3 (Part Number 20440). Samples were imaged on the Bionano Saphyr with at
446 least 500Gbp of data collected per sample.

447 Molecules for each sample were assembled using the Bionano *de novo* assembly pipeline on Bionano
448 Access Version 1.7. The preassembly setting was turned on, and the variant annotation parameters were
449 deselected for every assembly. *De novo* maps were aligned to the Mzebra_GT3a genome.

450 **PacBio HiFi sequencing and genome assembly.** PacBio *de novo* assemblies were generated from HiFi
451 reads across three sequencing runs. The MZ_GT3.3 and YH_GT1.3 were sequenced following DNA
452 extraction from frozen caudal fin tissue using the Qiagen MagAttract HMW DNA Kit (Cat. No. 67563). In the
453 first sequencing run a *Metriaclima zebra* female was sequenced sequenced on a PacBio Sequel II system
454 by the Georgia Genomics and Bioinformatics core. In the second run, the same the *Metriaclima zebra*
455 female and an *Aulonocara sp. 'chitande type north' Nkhata Bay* female was sequenced on a PacBio Revio
456 instrument by the HudsonAlpha Institute for Biotechnology. Reads from both runs were used to assemble
457 the MZ_GT3.3 genome and reads from the second run alone were used to assemble the YH_GT1.3 genome
458 (see details below).

459 Mzebra_GT3a, MZ_GT3.2, YH_GT1.1, and YH_GT1.2 were assembled using HiFi reads generated with DNA
460 extracted from fresh whole blood and heart tissue. DNA was extracted via the PacBio Nanobind Tissue Kit
461 RT (Part Number 102-208-000) following a modified DNA from animal tissue using the Nanobind®
462 PanDNA kit protocol (Part Number 102-574-600). Blood and heart tissue were combined and
463 homogenized together using a Qiagen TissueRuptor (Qiagen 9002755) and centrifugation g-force was
464 halved during the DNA extraction. DNA fragments <25kb were removed using the PacBio Short Read
465 Elimination kit (Part Number 102-208-300). Library preparation and sequencing was performed by the
466 University of Maryland Institute for Genome Sciences on a PacBio Revio instrument.

467 HiFi reads from all samples were assembled using the Mabs de novo assembler (v2.28)[57] using the `mabs-`
468 `hifiasm` algorithm and default parameters. Each assembly from mabs-hifiasm was evaluated using
469 Inspector (`inspector.py` v1.0.1)[58] run on default parameters. The assemblies were error corrected with
470 `inspector-correct.py` (v1.0) to resolve large structural errors. These error-corrected contigs from
471 were uploaded to Bionano Access.

472 The MZ_GT3.2, MZ_GT3.3, YH_GT1.1, YH_GT1.2 and YH_GT1.3 scaffold-level assemblies were generated
473 by bridging gaps in their contigs using Bionano maps via the single-enzyme Bionano Hybrid Scaffold
474 pipeline using default parameters. A unique set of Bionano maps was used to scaffold the contigs from
475 each *de novo* assembly. The resulting hybrid scaffold `NCBI.fasta` file (which denotes gaps filled by
476 Bionano maps with stretches of N nucleotides) was concatenated with any unscaffolded contigs from
477 the error-corrected mabs assembly. These five scaffold-level assemblies have been deposited at
478 DDBJ/ENA/GenBank (accessions TBD).

479 A scaffold-level genome was identically generated for Mzebra_GT3a. These scaffolds were further aligned
480 to the corrected UMD2a genome (see below) and anchored to linkage groups using D-GENIES[59] with the
481 Minimap2 v2.28 aligner and "Many repeats" options. An anchored genome was output using the Query
482 assembled as reference export option.

483 Note that the mitochondrial DNA sequence in Mzebra_GT3a is the same as the mitochondria assembled
484 in UMD2a. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the
485 accession JBEVYI000000000. The version described in this paper is version JBEVYI010000000.

486 **Correcting errors in M_zebra_UMD2a.** We generated Bionano molecules for 3 male and 1 female
487 *Metriaclima zebra*, assembled Bionano maps, and aligned them to the UMD2a reference genome. We
488 reordered or reoriented contigs that were revealed as misassemblies or inversions, respectively, in all four
489 Bionano maps and corresponded to breakpoints between contigs in UMD2a. By filtering interchromosomal
490 translocation calls in the Bionano Access software, we were able to insert unmapped contigs into the
491 established linkage groups.

**Illumina extractions and sequencing.** DNA was extracted from fresh or frozen fin tissues for short-read sequencing using the Qiagen MagAttract HMW DNA Kit (Cat. No. 67563) or the Qiagen DNeasy Blood & Tissue Kit (Cat No. 69504) using the Manual Purification of High-Molecular Weight Genomic DNA from Fresh or Frozen Tissue and Purification of Total DNA from Animal Tissues (Spin-Column) protocols respectively. DNA was delivered to the Georgia Tech Molecular Evolution Core where libraries were prepared with the NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina (NEB #E6177) using the Protocol for FS DNA Library Prep Kit (E7805, E6177) with Inputs ≥100 ng. A small subset of samples was sent to an external collaborator by the Molecular Evolution Core where library preparation was performed using the KAPA HyperPrep Kit (Roche, Material Number: 07962363001). Samples across all runs were sequenced on a NovaSeq 6000 instrument using v1.5 chemistry.

**Variant calling and PCA analysis.** We provide a link to a github repository (https://github.com/ptmcgrat/CichlidSRSequencing/tree/Kumar_eLife) containing the scripts used for the major analysis in the paper. Because our data is behind a secure Dropbox account, readers will be limited to seeing the exact programs, filters, and parameters used for this manuscript embedded within each script.

Fastq files from Illumina sequencing were converted to UBAM format using the `gatk`[60] `FastqToSam` algorithm. UBAM files were used for alignment to Mzebra_GT3 using `bwa`[61] `mem`; the `-M` and `-p` flags were used. To carry forward metadata from the unaligned BAMs, `gatk MergeBamAlignment` was called on the alignments. BAM files were converted to GVCF format using `gatk HaplotypeCaller`. Variant calling was performed using the GVCF files for the analysis cohort. The `gatk GenomicsDBImport` and `gatk GenotypeGVCFs` algorithms were used to generate a master vcf file which was subsequently filtered with `gatk VariantFiltration` (parameters below). The variants that passed filtering were used for PCA analysis.

PCA was performed with `plink2`[62]. To avoid overrepresenting species from the cohort, a core subset of the samples was used for PC 1 and 2 calculations (**Table S2**, column S). Eigenvectors from the whole cohort were then plotted on this PC space and visualized using `plotly`.

Note that GATK ≥ v4.3.0.0 and python ≥ 3.7 were used for these analyses. Custom python scripts were written to automate and parallelize processing of the samples in the cohorts.

The commands used for variant calling and PCA are the following:

*UBAM Generation:*

```
> gatk FastqToSam --FASTQ <fq1> --FASTQ2 <fq2> --READ_GROUP_NAME <RUNID> --
TMP_DIR <Temp_dir> --OUTPUT <temp_bam_file> --SAMPLE_NAME <sample_name> --
LIBRARY_NAME <library_name> --PLATFORM <platform>
```

*Alignment to GT3:*

```
> bwa mem -t <threads> -M -p <GT3_FASTA> <UBAM_file>
```

*Merge metadata from UBAM to BAM:*

```
> gatkMergeBamAlignment -R <GT3_FASTA> --UNMAPPED_BAM <UBAM_file> --
ALIGNED_BAM <BAM_file>
```

*Generate GVCFS:*

```
> gatk HaplotypeCaller -R <GT3_FASTA> -I <BAM_file> -ERC GVCF -O <GVCF_file>
```

`-ERC GVCF` is used to generate GVCF files that can be used for subsequent joint genotyping

*Read Depth Information:*

```
> gatk CountReads -I <BAM_file>
```

*Generate GenomicsDB:*

```
> gatk --java-options -Xmx <memory> GenomicsDBImport --genomicsdb-workspace-path <database_output_path> --intervals <interval> --sample-name-map <cohort_samples.txt>
```

*Joint Genotyping:*

```
> gatk --java-options -Xmx <memory> GenotypeGVCFs -R <GT3_FASTA> -V <path_to_database> -O <VCF_file> --heterozygosity 0.00175 -A DepthPerAlleleBySample -A Coverage -A GenotypeSummaries -A TandemRepeat -A StrandBiasBySample -A ReadPosRankSumTest -A AS_ReadPosRankSumTest -A AS_QualByDepth -A AS_StrandOddsRatio -A AS_MappingQualityRankSumTest -A FisherStrand -A QualByDepth RMSMappingQuality -A DepthPerSampleHC -G StandardAnnotation AS_StandardAnnotation -G StandardHCAnnotation
```

Note that an average measure of heterozygosity, 0.00175 was used from what was reported in Malinksy *et al*.[30].

*Filtering Variants:*

```
> gatk VariantFiltration -R <GT3_FASTA> <VCF_file> -O <Filtered_VCF> --filter-name ‘allele_freq’ --filter-expression 'AF < 0.05' --filter-name 'inbreeding_test' --filter-expression 'InbreedingCoeff < -0.6' --filter-name 'depth_Qual' --filter-expression 'QD < 2.0’ --filter-name 'max_DP’ --filter-expression 'DP > 11000' --filter-name 'min_DP' --filter-expression 'DP < 7600' --filter-name 'strand_bias' --filter-expression 'FS > 40.0' --filter-name 'mapping_quality' --filter-expression 'MQ < 50.0' --filter-name 'no_calls’ --filter-expression 'NCC > 119' --verbosity ERROR"))
```

Variants were filtered by allele frequency, inbreeding coefficient, quality by depth, depth, fisher’s exact test for strand bias, mapping quality, and by excess missingness according to methods published in Malinksky *et al*.[30]. We filtered by depth to include variants with a mean depth per sample between 10% and 95% of the distribution of all variant depths.

```
> gatk SelectVariants -V <Filtered_VCF> --exclude-filtered -O <pass_VCF_file>
```

*PCA:*

For PCA, first a subset of the pass_VCF_file containing the samples we used to create the principal components was generated using `bcftools`[63]. Next, the relevant pfiles were generated for both the pass_VCF_file and the subset_samples_VCF. Linkage pruning was only performed for the whole genome and whole chromosome PCAs using the –indep-pairwise 50 5 0.1 flag and parameters. Since inversions inherently link together variants, linkage pruning was not used when restricting the analysis to within inverted regions. A linear scoring system is generated using the subset_samples_VCF pfiles. This scoring system is applied to the genotype matrix for the whole cohort’s samples to scale each sample consistently. The resulting first two eigenvectors per sample from the .sscore file are plotted with `plotly`.

```
> plink2 --vcf <pass_VCF_file>  --out <whole_pfiles>
> plink2 --vcf <subset_samples_VCF_file> --out <subset_pfiles>
> plink2 --pfile <whole_pfiles> --set-missing-var-ids @:# --make-pgen --out <whole_corrected>
```

LD Pruning for Whole genome and whole chromosome PCAs:

```
> plink2 -pfile <subset_pfiles> --out <subset_samples_ld_pruning_intermediate> --allow-extra-chr --set-missing-var-ids @:#' --indep-pairwise 50 5 0.1
```

```
> plink2 -pfile <subset_pfiles> --freq counts -pca allele-wts -out <sample_subset_yes_ld_pruning_pca> --allow-extra-chr --set-missing-var-ids @:# --max-alleles 2 --extract <subset_samples_ld_pruning_intermediate.prune.in
```

```
> plink2 -pfile <whole_corrected> --read-freq <sample_subset_yes_ld_pruning_pca.acount> --score <sample_subset_yes_ld_pruning_pca.eigenval.allele> 2 5 header-read no-mean-imputation variance-standardize --score-col-nums 6-15 -out <projected_pca_yes_ld --allow-extra-chr
```

No LD Pruning for inverted region PCAs:

```
> plink2 --pfile <subset_pfiles> --freq counts -pca allele-wts -out <subset_samples_no_ld_pruning_pca> --allow-extra-chr --set-missing-var-ids @:# --max-alleles 2
```

```
> plink2 -pfile <whole_corrected> --read-freq <subset_samples_no_ld_pruning_pca.acount> --score' <subset_samples_no_ld_pruning_pca.eigenvec.allele> 2 5 header-read no-mean-imputation variance-standardize --score-col-nums 6-15 --out <projected_PCA_no_ld> --allow-extra-chr
```

**Genome alignment.** Genomes were aligned using `minimap2`[64] (v2.22) to align the query genome to the M_zebra_GT3 reference using default settings. Custom scripts were used to convert the output paf files to dataframe objects, filtering removed alignments that were 1) secondary matches, 2) had less than 30% percent identify, and 3) were less than 10,000 base pairs in length. Further, we excluded any alignments from a contig containing less than 2,000,000 total bp of alignments, to account for repetitive DNA. Alignments were plotted using `seaborn`. We estimated the intervals containing the breakpoints for the inversions using genome alignments of Aulonocara sp. 'chitande type north' Nkhata Bay and M_zebra_GT3 (**Table S4**).

**Phylogenies.** To create phylogenies, `bcftools` was used to filter out non SNV variants from the master vcf file. `Bcftools` was also used to create individual vcf files for each inversion, filtering out SNVs that fell outside of the inverted region. `Vcf2philip`[65] was used to convert the to the phylip format, using the -m 50 option to filter out SNVs with less than 50 genotyped samples. `Iqtree2`[66] was used to create trees and estimate confidence values for each node using the following options: `-nt','24','-mem','96G','-v','--seqtype', 'DNA','-m','GTR+I+G','-B','1000'`. Trees were visualized with `iTol`[67] or `ete3`[68].

**Nucleotide diversity analysis.** To analyze the Pixy between individuals of different species, we used `scikit-allel` (v1.3.8)[69] to calculate the genetic difference between each sample using the `allel.pairwise_distance` function and the `citblock` for the metric. We restricted the analysis to benthic animals that carried the inversion and compared these to animals of various ecogroups. We used the `seaborn` function `hist plot` to create density histograms for each of the inversions.

**Fst & pedigree.** A vcf file was created for the *Aulonocara* samples listed in **Table S2** (Column I). The PCA approach (described above) was used to genotype each of the samples for the inversion on 10. For the $F_{ST}$ analysis, we used `scikit-allel` (v1.3.8) to perform `allel.average_weir_cockerham_fst` with a window size of 100. For the heterozygosity analysis, we used the `allel.heterozygosity_observed` function on both the male and female population, subtracting the average heterozygosity from the male offspring from the female offspring.

## DATA AVAILABILITY

All Illumina and PacBio sequencing reads have been deposited to the NCBI Short Read archive at BioProject PRJNA1112855. The VCF files used in these analyses are available through the Dryad Digital Repository (TBD).

## REFERENCES


1. Kirkpatrick, M. How and Why Chromosome Inversions Evolve. *PLOS Biol.* **8**, e1000501 (2010).
2. Wellenreuther, M. & Bernatchez, L. Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends Ecol. Evol.* **33**, 427–440 (2018).
3. Berdan, E. L. *et al.* How chromosomal inversions reorient the evolutionary process. *J. Evol. Biol.* **36**, 1761–1782 (2023).
4. Hoffmann, A. A. & Rieseberg, L. H. Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annu. Rev. Ecol. Evol. Syst.* **39**, 21–42 (2008).
5. Andolfatto, P., Depaulis, F. & Navarro, A. Inversion polymorphisms and nucleotide variability in Drosophila. *Genet. Res.* **77**, 1–8 (2001).
6. Harringmeyer, O. S. & Hoekstra, H. E. Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nat. Ecol. Evol.* **6**, 1965–1979 (2022).
7. Huang, K., Andrew, R. L., Owens, G. L., Ostevik, K. L. & Rieseberg, L. H. Multiple chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype. *Mol. Ecol.* **29**, 2535–2549 (2020).

8. Hager, E. R. *et al.* A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. *Science* **377**, 399–405 (2022).

9. Joron, M. *et al.* Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203–206 (2011).

10. Joron, M. *et al.* A Conserved Supergene Locus Controls Colour Pattern Diversity in Heliconius Butterflies. *PLOS Biol.* **4**, e303 (2006).

11. Palmer, D. H. & Kronforst, M. R. A shared genetic basis of mimicry across swallowtail butterflies points to ancestral co-option of doublesex. *Nat. Commun.* **11**, 6 (2020).

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

13. Lamichhaney, S. *et al.* Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nat. Genet.* **48**, 84–88 (2016).

14. Küpper, C. *et al.* A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* **48**, 79–83 (2016).

15. Wang, J. *et al.* A Y-like social chromosome causes alternative colony organization in fire ants. *Nature* **493**, 664–668 (2013).

16. Tuttle, E. M. *et al.* Divergence and Functional Degradation of a Sex Chromosome-like Supergene. *Curr. Biol. CB* **26**, 344–350 (2016).

17. Hughes, J. F. *et al.* Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).

18. Fishman, L., Stathos, A., Beardsley, P. M., Williams, C. F. & Hill, J. P. CHROMOSOMAL REARRANGEMENTS AND THE GENETICS OF REPRODUCTIVE BARRIERS IN MIMULUS (MONKEY FLOWERS). *Evolution* **67**, 2547–2560 (2013).

673 19. Noor, M. A., Grams, K. L., Bertucci, L. A. & Reiland, J. Chromosomal inversions and the
674 reproductive isolation of species. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 12084–12088 (2001).
675 20. Fuller, Z. L., Koury, S. A., Phadnis, N. & Schaeffer, S. W. How chromosomal rearrangements shape
676 adaptation and speciation: Case studies in Drosophila pseudoobscura and its sibling species
677 Drosophila persimilis. *Mol. Ecol.* **28**, 1283–1301 (2019).
678 21. Fuller, Z. L., Leonard, C. J., Young, R. E., Schaeffer, S. W. & Phadnis, N. Ancestral polymorphisms
679 explain the role of chromosomal inversions in speciation. *PLOS Genet.* **14**, e1007526 (2018).
680 22. Kirkpatrick, M. & Barton, N. Chromosome Inversions, Local Adaptation and Speciation. *Genetics*
681 **173**, 419–434 (2006).
682 23. Thompson, M. J. & Jiggins, C. D. Supergenes and their role in evolution. *Heredity* **113**, 1–8 (2014).
683 24. Lowry, D. B. & Willis, J. H. A Widespread Chromosomal Inversion Polymorphism Contributes to a
684 Major Life-History Transition, Local Adaptation, and Reproductive Isolation. *PLoS Biol.* **8**, e1000500
685 (2010).
686 25. Trickett, A. J. & Butlin, R. K. Recombination suppressors and the evolution of new species.
687 *Heredity* **73**, 339–345 (1994).
688 26. Kocher, T. D. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.*
689 **5**, 288–298 (2004).
690 27. Santos, M. E., Lopes, J. F. & Kratochwil, C. F. East African cichlid fishes. *EvoDevo* **14**, 1 (2023).
691 28. Svardal, H., Salzburger, W. & Malinsky, M. Genetic Variation and Hybridization in Evolutionary
692 Radiations of Cichlid Fishes. *Annu. Rev. Anim. Biosci.* **9**, 55–79 (2021).
693 29. Johnson, Z. V. *et al.* Cellular profiling of a recently-evolved social behavior in cichlid fishes. *Nat.*
694 *Commun.* **14**, 4891 (2023).
695 30. Malinsky, M. *et al.* Whole-genome sequences of Malawi cichlids reveal multiple radiations
696 interconnected by gene flow. *Nat. Ecol. Evol.* **2**, 1940–1955 (2018).

31. Patil, C. *et al*. Genome-enabled discovery of evolutionary divergence in brains and behavior. *Sci. Rep*. **11**, 13016 (2021).

32. Todd Streelman, J. & Danley, P. D. The stages of vertebrate evolutionary radiation. *Trends Ecol. Evol*. **18**, 126–131 (2003).

33. Maan, M. E. & Sefc, K. M. Colour variation in cichlid fish: Developmental mechanisms, selective pressures and evolutionary consequences. *Semin. Cell Dev. Biol*. **24**, 516–528 (2013).

34. Konings, A. Fishes, as well as birds, build bowers.

35. Martin, C. & Genner, M. A role for male bower size as an intrasexual signal in a Lake Malawi cichlid fish. (2009) doi:10.1163/156853908X396836.

36. Loh, Y.-H. E. *et al*. Origins of Shared Genetic Variation in African Cichlids. *Mol. Biol. Evol*. **30**, 906–917 (2013).

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

38. Conte, M. A. *et al*. Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes. *GigaScience* **8**, giz030 (2019).

39. Bocklandt, S., Hastie, A. & Cao, H. Bionano Genome Mapping: High-Throughput, Ultra-Long Molecule Genome Analysis System for Precision Genome Assembly and Haploid-Resolved Structural Variation Discovery. in *Single Molecule and Single Cell Sequencing* (ed. Suzuki, Y.) 97–118 (Springer, Singapore, 2019). doi:10.1007/978-981-13-6037-4_7.

40. Villoutreix, R. *et al*. Inversion breakpoints and the evolution of supergenes. *Mol. Ecol*. **30**, 2738–2755 (2021).

41. Ciezarek, A. G. *et al*. Ancient and Recent Hybridization in the Oreochromis Cichlid Fishes. *Mol. Biol. Evol*. **41**, msae116 (2024).

720 42. Nowling, R. J., Manke, K. R. & Emrich, S. J. Detecting inversions with PCA in the presence of
721 population structure. *PLoS ONE* **15**, e0240429 (2020).
722 43. York, R. A. *et al*. Behavior-dependent cis regulation reveals genes and pathways associated with
723 bower building in cichlid fishes. *Proc. Natl. Acad. Sci.* **115**, E11081–E11090 (2018).
724 44. Kratochwil, C. F., Liang, Y., Urban, S., Torres-Dowdall, J. & Meyer, A. Evolutionary Dynamics of
725 Structural Variation at a Key Locus for Color Pattern Diversification in Cichlid Fishes. *Genome Biol.*
726 *Evol.* **11**, 3452–3465 (2019).
727 45. Feulner, P. G. D., Schwarzer, J., Haesler, M. P., Meier, J. I. & Seehausen, O. A Dense Linkage Map of
728 Lake Victoria Cichlids Improved the Pundamilia Genome Assembly and Revealed a Major QTL for Sex-
729 Determination. *G3 GenesGenomesGenetics* **8**, 2411–2420 (2018).
730 46. Johnson, N. A. & Lachance, J. The genetics of sex chromosomes: evolution and implications for
731 hybrid incompatibility. *Ann. N. Y. Acad. Sci.* **1256**, E1-22 (2012).
732 47. Natri, H. M., Merilä, J. & Shikano, T. The evolution of sex determination associated with a
733 chromosomal inversion. *Nat. Commun.* **10**, 145 (2019).
734 48. Malinsky, M. & Salzburger, W. Environmental context for understanding the iconic adaptive
735 radiation of cichlid fishes in Lake Malawi. *Proc. Natl. Acad. Sci.* **113**, 11654–11656 (2016).
736 49. Lyons, R. P. *et al*. Continuous 1.3-million-year record of East African hydroclimate, and
737 implications for patterns of evolution and biodiversity. *Proc. Natl. Acad. Sci.* **112**, 15568–15573 (2015).
738 50. Ivory, S. J. *et al*. Environmental change explains cichlid adaptive radiation at Lake Malawi over the
739 past 1.2 million years. *Proc. Natl. Acad. Sci.* **113**, 11895–11900 (2016).
740 51. Kondrashov, A. S. & Mina, M. V. Sympatric speciation: when is it possible? *Biol. J. Linn. Soc.* **27**,
741 201–223 (1986).
742 52. Parnell, N. F. & Streelman, J. T. Genetic interactions controlling sex and color establish the
743 potential for sexual conflict in Lake Malawi cichlid fishes. *Heredity* **110**, 239–246 (2013).

53. Behrens, K. A., Koblmueller, S. & Kocher, T. D. Diversity of Sex Chromosomes in Vertebrates: Six Novel Sex Chromosomes in Basal Haplochromines (Teleostei: Cichlidae). *Genome Biol. Evol.* **16**, evae152 (2024).

54. Blumer, L. M. *et al.* Introgression dynamics of sex-linked chromosomal inversions shape the Malawi cichlid adaptive radiation. Preprint at https://doi.org/10.1101/2024.07.28.605452 (2024).

55. Malinsky, M. *et al.* Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat. Ecol. Evol.* **2**, 1940–1955 (2018).

56. Meier, J. I. *et al.* Cycles of fusion and fission enabled rapid parallel adaptive radiations in African cichlids. *Science* **381**, eade2833 (2023).

57. Schelkunov, M. I. Mabs, a suite of tools for gene-informed genome assembly. *BMC Bioinformatics* **24**, 377 (2023).

58. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* **22**, 312 (2021).

59. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).

60. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. 201178 Preprint at https://doi.org/10.1101/201178 (2018).

61. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

62. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).

63. 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).

64. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

65. Ortiz, E. M. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. Zenodo https://doi.org/10.5281/zenodo.2540861 (2019).

66. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

67. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).

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

69. Miles, A. *et al.* cggh/scikit-allel: v1.3.13. Zenodo https://doi.org/10.5281/zenodo.13772087 (2024).