Resubmission to Molecular Ecology BIORXIV/2020/961987

Movement of transposable elements contributes to cichlid diversity

Karen L. Carleton[1+], Matthew A Conte[1], Milan Malinsky[2,3], Sri Pratima Nandamuri[1*], Benjamin A. Sandkam[1&], Joana I Meier[4,5,6,%], Salome Mwaiko[4,5], Ole Seehausen[4,5], Thomas D Kocher[1]

[1]Department of Biology, University of Maryland, College Park MD 20742 USA

[2] Wellcome Sanger Institute, Cambridge, UK.

[3]Zoological Institute, University of Basel, Basel, Switzerland

[4]Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland

[5]Department of Fish Ecology and Evolution, Centre for Ecology, Evolution & Biogeochemistry, Eawag: Swiss Federal Institute of Aquatic Science and Technology, 6047 Kastanienbaum, Switzerland

[6]Computational and Molecular Population Genetics Lab, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland

[+]corresponding author: kcarleto@umd.edu

*Current address: Department of Human Genetics, University of Utah, Salt Lake City UT, USA

[&]Current address: Department of Zoology, University of British Columbia, Vancouver, BC, Canada

[%]Current address: Department of Biology, University of Cambridge, Cambridge UK

This document is the accepted manuscript version of the following article:
Carleton, K. L., Conte, M. A., Malinsky, M., Nandamuri, S. P., Sandkam, B. A., Meier, J. I., … Kocher, T. D. (2020). Movement of transposable elements contributes to cichlid diversity. Molecular Ecology, 29(24), 4956-4969.
https://doi.org/10.1111/mec.15685

Abstract

African cichlid fishes are a prime model for studying speciation mechanisms. Despite the development of extensive genomic resources, it has been difficult to determine which sources of genetic variation are responsible for cichlid phenotypic variation. One of their most variable phenotypes is visual sensitivity, with some of the largest spectral shifts among vertebrates. These shifts arise primarily from differential expression of seven cone opsin genes. By mapping expression quantitative trait loci (eQTL) in intergeneric crosses of Lake Malawi cichlids, we previously identified four causative genetic variants that correspond to indels in the promoters of either key transcription factors or an opsin gene. In this comprehensive study, we show that these indels are the result of the movement of transposable elements (TEs) that correlate with opsin expression variation across the Malawi flock. In tracking the evolutionary history of these particular indels, we found they are endemic to Lake Malawi, suggesting that these TEs are recently active and are segregating within the Malawi cichlid lineage. However, an independent indel has arisen at a similar genomic location in one locus outside of the Malawi flock. The convergence in TE movement suggests these loci are primed for TE insertion and subsequent deletions. Increased TE mobility may be associated with interspecific hybridization, which disrupts mechanisms of TE suppression. This might provide a link between cichlid hybridization and accelerated regulatory variation. Overall, our study suggests that TEs may be an important driver of key regulatory changes, facilitating rapid phenotypic change and possibly speciation in African cichlids.

Introduction

The genomic era promises to unlock the molecular links between genotype and phenotype, including the evolution of new species. Questions of evolutionary predictability have focused on whether new phenotypes result from changes in coding sequence or gene regulation (Carroll, 2008; Hoekstra & Coyne, 2007; Stern & Orgogozo, 2008). In addition, there has been a strong focus on single nucleotide polymorphisms (SNPs) during the search for highly selected regions, magic genes, and islands of speciation (Malinsky et al., 2015; Malinsky et al., 2018; Servedio, Van Doorn, Kopp, Frame, & Nosil, 2011; Svardal et al., 2020; Turner & Hahn, 2010). However, evolution occurs through more than changes to single DNA bases. Structural rearrangements, such as insertions and deletions (indels; Green et al., 2010; Mills et al., 2006), inversions, gene duplications (Cortesi et al., 2015; Ohno, 1970), or whole genome duplications (Crow, Wagner, & Investigators, 2006; Otto & Whitton, 2000) can be critical to generating evolutionary novelty.

The approximately 1500 species of African cichlid fishes are a textbook example of adaptive radiations, characterized by phenotypic change in rapidly speciating lineages (Kocher, 2004; Malinsky & Salzburger, 2016; Seehausen, 2006). Cichlid species differ in many extensively studied and ecologically important phenotypes, including jaw morphology (Albertson, Streelman, & Kocher, 2003), color patterns (Allender, Seehausen, Knight, Turner, & Maclean, 2003; Danley & Kocher, 2001; Konings, 2007; Seehausen, 1996), sex determination (Gammerdinger & Kocher, 2018; R. Roberts, Ser, & Kocher, 2009), and parental care (Barlow, 2000; Sefc, 2011).

To understand the genetic basis of this diversity, the genomes of five species were sequenced by the Cichlid Genome Consortium (Brawand et al., 2014). This included one species from each of the three Great African Lakes, two haplochromine (Malawi, Victoria) and one lamprologine cichlid (Tangaynika), as well as a haplochromine and an oreochromine riverine species. Based on the genomic analyses, at least five mechanisms were suggested to contribute to the unusually rapid and extensive evolutionary diversification in African lacustrine cichlid fish. First, there was an acceleration of gene sequence evolution in African cichlids when compared to other fish. A further acceleration, and higher dN/dS ratios, occurred in the haplochromine and lamprologine lineages, including genes controlling development, pigmentation, and vision. Second, there was a 4.5-6x increase in the rate of gene duplications in the ancestor of the lamprologine and haplochromine cichlids, when compared to *Oreochromis* and to non-cichlid fish such as stickleback and zebrafish. Third, there were 625 regulatory regions which showed accelerated evolution within the haplochromine and lamprologine species. Fourth, there were 40 gains and 9 losses of miRNAs. Finally, there was evidence for transposable element insertions that were associated with changes in gene expression. Although all of these mechanisms are likely to contribute to cichlid diversification, they have yet to be tied to specific phenotypes.

The genetic architecture of some ecologically important cichlid traits, such as body shape, is likely to be highly polygenic. For such traits, the individual contributions of numerous loci of small effect are very difficult to determine. The link between genotype and phenotype is easier to make for traits with simpler genetic architectures dominated by large effect loci, including pigmentation patterns (Feller, Haesler, Peichel, & Seehausen, 2020; Kratochwil et al.,

2018; Santos et al., 2014), some aspects of jaw development (e.g. Conith et al., 2018; Parsons, Trent Taylor, Powder, & Albertson, 2014; R. B. Roberts, Hu, Albertson, & Kocher, 2011), or various adaptations of the visual system (e.g. Carleton, Dalton, Escobar-Camacho, & Nandamuri, 2016; Malinsky et al., 2015; Malinsky et al., 2018; Seehausen et al., 2008; Sugawara et al., 2005).

In this study, we focus on the evolution of cichlid visual systems. Visual systems have the advantage that opsin genes are already known to play a key role in shaping visual sensitivity phenotypes. Opsin proteins combine with a chromophore, such as 11-cis retinal, to produce visual pigments that absorb light (Yokoyama, 2008). The sequence of the opsin protein is key to tuning visual pigment sensitivity. Because of the importance of visual systems to survival, they are likely under strong natural selection (Davies, Collin, & Hunt, 2012; Goldsmith, 2013; Maan, Seehausen, & Groothuis, 2017).

Microspectrophotometric measurements of cichlid rods and cones suggest that cichlid visual systems are highly variable among species (Jordan et al., 2006; Levine, MacNichol, Kraft, & Collins, 1979; van der Meer & Bowmaker, 1995). Cichlids have short wavelength single cones and longer wavelength double cones. Species can differ in whether their single cones are ultraviolet, violet or blue sensitive with peak sensitivities varying by up to 90 nm, while green to red sensitive double cones vary by 30 to 50 nm (Carleton, 2009; Jordan et al., 2006). Closely related species show remarkable shifts in sensitivity (Hofmann et al., 2009; Seehausen et al., 2008; Terai et al., 2006).

Previous work has started to identify the genetic variation underlying visual diversity. Cichlids have seven cone opsin genes, which belong to the four vertebrate classes: very short

wavelength sensitive 1 (SWS1), short wavelength sensitive 2 (SWS2A and SWS2B), rhodopsin like (RH2Aa, RH2Ab and RH2B) and long wavelength sensitive (LWS). Protein expression confirmed that these seven genes produce distinct visual pigments with peak sensitivities distributed across the spectrum from ultraviolet to red wavelengths (Parry et al., 2005; Spady et al., 2006). Comparisons across different species suggest that opsin sequences may evolve rapidly and adaptively, but sequence differences contribute relatively small spectral shifts (Malinsky et al., 2015; Nagai et al., 2011; Seehausen et al., 2008; Spady et al., 2005; Sugawara et al., 2005; Terai et al., 2006). Therefore, while opsin protein sequences are important, they are not the primary driver of larger visual sensitivity shifts (Carleton & Kocher, 2001; Hofmann et al., 2009).

The larger visual sensitivity shifts are the result of differential expression of opsin genes. Adults are typically trichromatic, expressing three cone opsins (Carleton, 2009; Carleton et al., 2016). The three common gene combinations are the short (*SWS1, RH2B, RH2A*), medium (*SWS2B, RH2B, RH2A*) and long (*SWS2A, RH2A, LWS*) visual palettes (with *RH2A* representing either of two highly similar genes, *RH2A*α and *RH2A*β). Although many species only express one palette throughout life, some species progress from the short to medium to long combinations through ontogeny (Carleton et al., 2008; O'Quin, Smith, Sharma, & Carleton, 2011).

To identify the loci underlying these differences in opsin expression, we previously made genetic crosses between species expressing different palettes. With one exception, the causative genetic factors are not directly linked to the opsin genes themselves. Instead we found several expression quantitative trait loci (eQTL) acting in trans to the opsin genes (Nandamuri, Conte, & Carleton, 2018; O'Quin et al., 2012). Using fine mapping in crosses, and

association mapping in natural populations, we identified the causative genes as well as putative mutations that might underlie these changes (Table 1). Retinal homeobox 1 (Rx1; Schulte, O'Brien, Conte, O'Quin, & Carleton, 2014) and microphthalmia associated transcription factor a (Mitfa; Nandamuri, 2018) are trans factors associated with changes in the expression of the *SWS2A* opsin gene. A 413 bp deletion that is 2.5 kb upstream of the Rx1 translational start site is correlated with decreases in *SWS2A* expression in Lake Malawi cichlids and explains 62% of its variance across more than 50 species (Schulte et al., 2014). A 1.4kb insertion in intron 1 of *Mitfa* is correlated with an increase in *SWS2A* expression, though it has a smaller effect than *Rx1* (Nandamuri, 2018). T-box 2a (*Tbx2a*) is a trans factor associated with expression of the *LWS* opsin (Sandkam et al., 2020). *Tbx2a* binds to regulatory regions for both *LWS* and *RH2* and acts to switch between these opsins. A 967 bp deletion that is 13.5 kb upstream of its translational start site causes a shift from *LWS* to *RH2* expression. In addition to these three trans-acting regulatory mutations, the one cis regulatory change is a 692 bp deletion in the *SWS1* promoter (Nandamuri et al., 2018). This deletion removes several conserved regulatory elements and acts to shut off *SWS1* expression. Therefore, in each of these cases, we have found a regulatory indel that either removes (*Rx1, Tbx2a, SWS1*) or adds (*Mitfa*) a critical regulatory region. These indels alter either the expression of the critical transcription factor affecting opsin expression, or directly alter the promoter sequence of the opsin itself (*SWS1*).

In the current study, we wanted to characterize the evolutionary origins and history of these loci. In addition, we wanted to explore the genomic variation underlying these four mutations. We find that they involve either insertions or deletions of large size (400-1400 bp). The boundaries of these indels largely correspond to transposable elements. By examining

species within and outside the Malawi flock, including from the sister Lake Victoria lineage, we discovered that these indels are recent and, with one exception, specific to the cichlids of Lake Malawi. Further, they are linked to an increase in the number of copies of particular TE families. Therefore, these indels explain some of the opsin expression variation segregating in the Lake Malawi flock. We suggest that the movement of transposable elements generates sizeable indels that modify important regulatory regions (Feschotte, 2008; Zeng, Pederson, Kortschak, & Adelson, 2018) and acts as an underappreciated but key source of variation helping to drive cichlid diversification.

METHODS

Indel analysis

We focus on the four mutations identified in our previous QTL studies. The loci for *Rx1, Tbx2a* and *Mitfa* were identified in a cross between *Tramitichromis intermedius* and *Aulonocara baenschi*. The *SWS1* locus was characterized in a cross between *Metriaclima* 'mbenji' and *A. baenschi.* These loci were then compared to the genome of *Metriaclima zebra* as well as genomes of outgroups to Lake Malawi including *Astatotilapia burtoni*, *Pundamilia nyererei* and *Oreochromis niloticus* (Brawand et al., 2014; Conte, Gammerdinger, Bartie, Penman, & Kocher, 2017; Conte et al., 2019; Feulner, Schwarzer, Haesler, Meier, & Seehausen, 2018). This determined the relative indel size and whether a locus was an insertion or a deletion relative to the outgroup species. Inserted sequences were analyzed using Repbase (Kohany, Gentles, Hankus, & Jurka, 2006) using the CENSOR website (https://girinst.org/censor/index.php) to determine if they corresponded to known transposable elements.

Origin and phylogenetic diversity of indel sequences

To identify the phylogenetic origin and to estimate the age of these indels, we searched for them across species within and outside of Lake Malawi using a combination of targeted PCR and bioinformatic queries of whole genome sequence data. In total, 209 species and 239 individuals were surveyed (individuals and species listed in Supp. Table S1). The regions were first examined in the five cichlid genome project species (Brawand et al., 2014; Conte et al., 2017; Conte et al., 2019; Feulner et al., 2018). Next, single individuals of 53 Lake Malawi species with known retinal opsin gene expression (Hofmann et al., 2009) were screened by PCR (primers in Supp. Table S2). To determine the phylogenetic origins of the indels, we searched additional species within and outside of the Lake Malawi flock. We used PCR to screen for the *Rx1* and *Mitfa* indels in additional species including three (n=1) from Lake Malawi, one from Lake Chilingali (a lake very close to Lake Malawi), and several samples of riverine *Astatotilapia* including two populations (n=1-2) of *A. calliptera* from Lake Malawi, and two *A. calliptera* (n=1-2) and one sample of *A.* sp. “tweddlei/kilossana” from the Rovuma River catchment. The *A. calliptera* are a sister group to the Lake Malawi rock dwelling clade, the ‘Mbuna’, (Malinsky et al., 2018). We also searched newly sequenced genomes of 103 Lake Victoria species (86 from Lake Victoria proper, 2 from Lake Nabugabo and 15 from Lake Kyoga) as well as 14 riverine outgroup species. Lake Victoria species were Illumina sequenced and the reads mapped onto the *Pundamilia nyererei* genome v. 2.0 (Feulner et al., 2018) with Bowtie v. 2 (Langmead & Salzberg, 2012). Variants were called with Haplotype Caller (GATK v. 3.5, McKenna et al. 2010) for each individual separately and subsequently combined with GenotypeGVCFs, GATK v. 3.5,

(McKenna et al., 2010) to make VCF tracks for viewing in the Integrative Genome Viewer (Robinson et al., 2011; Thorvaldsdottir, Robinson, & Mesirov, 2013). Finally, we searched 53 additional genomes including 15 species from Lake Malawi, 4 from LVRS, 4 from Lake Tanganyika and 30 outgroups (Malinsky et al., 2018; McGee et al., 2015; Svardal et al., 2020). For these taxa, we searched raw reads (10-15x coverage) from the Illumina platform using k-mer based analysis with overlapping 27-mers identified from the indels and several kb of surrounding sequences from the consensus of the reference genomes. The counts of occurrences of these k-mers were smoothed using a rolling average in windows of 20 k-mers. The deletions had consistent k-mer counts of zero throughout their sequences, whereas sequences present in the genomes had positive k-mer counts.

In total we examined 64 species (75 individuals) from the Lake Malawi clade including 6 individuals of *Astatotilapia* (3 *A. calliptera* from Lake Malawi, 1 from Lake Massoko and 2 *A. calliptera*), 106 species (108 individuals) from the Lake Victoria region superflock (LVRS), 4 species (4 individuals) from Lake Tanganyika, and 35 species (52 individuals) from smaller African lakes and rivers (Supp. Table S1). This amounts to 239 individuals from 209 species of haplochromine cichlids.

To place these taxa in a phylogenetic context, we used the tree of Meier et al. (2017), which is based on RAD sequences from haplochromines of all African catchments and all major clades (Meier et al. 2017, Supp. Fig. S1). We replaced the Malawi clade in that tree to include the more extensive set of Malawi taxa included here, using a taxonomic tree divided into four Lake Malawi clades: rock, sand, pelagic, and deep (Hofmann et al., 2009) and which includes *A. calliptera* as sister to the rock dwelling clade (Malinsky et al., 2018). The additional 53 samples

genotyped by k-mer analysis were grouped by neighbor joining based on distances calculated from a set of SNPs identified across the genome (tree includes 58 species of which 53 were analyzed for indels; Supp. Fig S2; Malinsky et al., 2018). This tree had 12 clades that shared at least one species in common with the RAD tree from Meier et al. (2017). We therefore noted how many of the species examined in this study fall within the clades identified by Meier et al. 2017.

Comparison of TE family sizes

We wanted to determine which TE families might be expanding in these cichlids. TE families generally consist of many highly similar copies spread throughout the genome. Therefore, confidently quantifying TE family loci can only be done with high quality genomes built from long sequencing reads that span TE boundaries. Therefore, we focused our analysis of TE families on the long-read based genome assemblies of the Malawi zebra cichlid *Metriaclima zebra* (UMD2a; Conte et al., 2019) and the Nile tilapia *Oreochromis niloticus* (UMD1; Conte et al., 2017). TEs were identified and assigned to families using a combination of RepeatModeler and RepeatMasker. First, RepeatModeler *version open-1.0.8* (Smit & Hubley, 2010) was used to identify and classify *de novo* repeat families separately for each assembly. These *de novo* repeats were then combined with the RepBase-derived RepeatMasker libraries (Bao, Kojima, & Kohany, 2015). RepeatMasker *version open-4.0.5* (Smit, Hubley, & Green, 2010) was run on the final anchored assembly using NCBI BLAST+ (*version 2.3.0+*) as the engine (‘*-e ncbi*’) and specifying the combined repeat library (‘*-lib*’). The more sensitive slow search mode (‘*-s*’) was used.

TEs were then ascribed to genome locations including promoters, exons, introns and intergenic regions (UTRs were included in the exon category). GFF3 annotations for the O_niloticus_UMDNMBU and M_zebra_UMD2a assemblies were downloaded from the NCBI RefSeq FTP site (NCBI; O'Leary et al., 2016). These GFF3 files were first converted to the UCSC GenePred format using the 'gff3ToGenePred' utility with the '-refseqHacks' option enabled. The GenePred formatted files were then converted to BED12 using the genePredToBed utility. These BED files were then used to define the gene parts within the genomation R package (Akalin, Franke, Vlahovicek, Mason, & Schubeler, 2015) to determine the overlap with the various TE families.

Age of TE insertions

To obtain additional information regarding the origin of the alleles at these loci, we estimated the relative divergence of the alleles with and without the indels using sequence divergence in flanking regions (i.e. not including sequence for either the deleted or inserted regions themselves). Homologous sequences for several regions were also extracted from the *P. nyererei* (Brawand et al., 2014; Feulner et al., 2018) *M. zebra* (Conte et al., 2019) and *A. calliptera* (Malinsky et al., 2018) genomes. For *Rx1*, and *Mitfa*, key regions were PCR amplified from up to 18 Lake Malawi cichlids and then Sanger sequenced using primers listed in Supp. Table S2. For *Rx1*, we compared 1.6kb of sequence surrounding the indel for six species having the insertion with 11 species having the deletion, along with *A. calliptera* and *P. nyererei* which had the ancestral state lacking either insertion or deletion. For the *Mitfa* insertion in intron 1, we compared 1.88kb of sequence covering exons 5 to 7. This included 6 species with the

insertion, and 6 species plus *P. nyererei* without it. Sequences for the *SWS1* promoter region were previously sequenced (Nandamuri et al., 2018) and we compared 1kb surrounding the deletion in *A. baenschi* with 16 species plus *P. nyererei* that had an intact promoter. For *Tbx2a*, we used genomic sequences and compared 1.8 kb of sequence surrounding the deletion for two *Aulonocara baenschi* individuals, which had the deletion, and two individuals of *Tramitichromis intermedius,* (Sandkam et al., 2020) as well as *M. zebra,* and *P. nyererei* that did not. Sequences for each locus were aligned using MUSCLE (Edgar, 2004) in Geneious 10.1.3 (https://www.geneious.com). Sections of the alignment containing any inserted or deleted sequence were removed to retain just the surrounding sequence. Tamura Nei corrected distance trees were generated using Geneious's neighbor joining algorithm (Tamura & Nei, 1993). Sequence distances between alleles containing the indel and alleles that lack the indel were averaged across all possible comparisons. These average allelic divergences were put into context by calculating their ratio either to the average sequence divergence between the deepest lineages within Lake Malawi (AlleleDiv/LM) or the sequence divergence for Lake Malawi and Victoria species (AlleleDiv/MV). When focusing on SNPs only, the genome wide sequence divergence between deep Malawi lineages is 0.001924 (Malawi Mbuna, *A. calliptera*, shallow benthic, deep benthic and sand dwelling utaka (N=78) compared to (*Diplotaxodon, Rhamphochromis* (N=10)) (Malinsky et al., 2018). The genome sequence divergence for Malawi lineages compared to *P. nyererei* in Lake Victoria is ~0.0076 (Svardal et al., 2020). By comparing the allelic divergence at each locus to these two calibrations, we can determine if the alleles arose relatively recently (AlleleDiv/LM <1), around the emergence of the Malawi lineages (AlleleDiv/LM ~ 1) or before the Malawi flock (AlleleDiv/LM >1 and AlleleDiv/MV ~ 0.5-1).

RESULTS

Comparative indel analysis

The indel variants underlying differences in opsin expression are illustrated in Figure 1, where the genotypes of several Malawi cichlid species with known opsin expression are compared with three outgroup species (*A. burtoni, P. nyererei* and *O. niloticus*). The indel locations in the *M. zebra* UMD2a genome are listed in Supp. Table S3, with the corresponding indel sequences given in Supp. Table S4. TE insertions and deletions are common and there are other indels in the broader regions, as shown in Supp. Fig S3A-D.

With the exception of *Astatotilapia calliptera* and some *Rhamphochromis*, the Lake Malawi taxa at the *Rx1* locus have either a fixed length deletion (413 bp relative to *P. nyererei*) or an insertion of varying length (268-421 bp relative to *P. nyererei*). The deletion occurs in short and medium palette Lake Malawi species at exactly the same location. The insertion occurs in long palette Lake Malawi species (and is shared with *A. calliptera* from the Rovuma River (Lucheringo)) but varies in length across species because of its repetitive microsatellites. The longest version occurs in *Dimidiochromis compressiceps.* Interestingly, the boundaries of the deletion are outside the boundaries of this insertion. We hypothesize that when the deletion occurred, it removed both the insertion as well as some surrounding ancestral sequence. However, it is possible that the deletion occurred on the ancestral haplotype at the same location as the insertion.

The *Mitfa* locus involves a 1408 bp insertion in intron 1 of the gene and occurs in numerous species. No species with deletions were identified. The *Tbx2a* locus includes a 1081

bp insertion in *M. zebra* and a 967 bp deletion in *A. baenschi*, relative to the outgroup *P. nyererei.* None of the other species show variation at this location. Of the species where we have sequenced the adjacent region, the *SWS1* promoter involves a 692 bp deletion that occurs only in *A. baenschi,* with no instances of insertions.

A comparison of the inserted sequences against known repeats (see Methods) revealed matches to TEs. From the longest 421 bp (*D. compressiceps*) version of the insertion in the Rx1 promoter, 384 bp match Rex1-5 AFC, a known nonLTR retrotransposon (Supp. Fig. S4A). For the 1081 bp insertion for the *Tbx2a* regulatory region in *M. zebra*, all but the first 9 bp match hAT-8 AFC, a DNA transposon from the hAT family (Supp. Fig. S4B). The AFC in the names of these repetitive elements indicates they were first described in African cichlids (Brawand et al., 2014). The last insertion, in intron 1 of *Mitfa,* is a bit more complex (Supp. Fig. S4C). It includes matches to four different transposable elements. The longest match is 522 bp of a Rex1 TE from *Petromyzon marinus*. There are several other fragments, which are 50-70 bp long, matching L1 LINES, DNA/Mariner and Copia LTR elements from diverse species. We also checked the deleted sequences against known repeats, and while they contained a number of small TE fragments, they do not appear to be entire TE excisions (Supp. Fig. S4A-D).

Phylogenetic origin

We found that all the major indels that affect visual tuning occur only in Lake Malawi species, or in *A. calliptera* of the Rovuma River (N=75 individuals). The Rovuma River is not connected to the Malawi catchment but the species clusters phylogenetically very close to or as part of the Lake Malawi radiation. None of the indels are found in any of the species from Lake

Victoria proper (N=90), other species in the Lake Victoria Region Superflock (N=18) or rivers and smaller lakes across eastern Africa (N=56). This provides strong evidence that these indels are specific to the Malawi flock and have arisen either in its immediate ancestor or even within the timescales of the Malawi radiation (complete dataset for all 209 species / 239 individuals given in Supp. Table S1).

The indel for the *Rx1* locus occurred in two samples from outside of Lake Malawi proper. The *Rx1* deletion was found in *Rhamphochromis longicep* from Lake Chilangali which is 10 km from Lake Malawi. However, the lake is within the Lake Malawi catchment. This species occurs also in Lake Malawi and is clearly part of the monophyletic genus *Rhamphochromis* within the Lake Malawi flock (Figure 2). The *Rx1* insertion was also found in *A. calliptera* of the Lucheringo River in the Rovuma/Indian Ocean catchment. *A. calliptera* from this part of Africa are very closely related to *A. calliptera* from the Lake Malawi basin, and to the Lake Malawi rock dwelling (Mbuna) clade with which the Rovuma clade of *A. calliptera* (but not the Malawi clade) also shares the mitochondrial haplotype (Joyce et al., 2011; Malinsky et al., 2018).

There is one additional taxon that has a similar, though distinct, deletion in the *Rx1* locus. *Ctenochromis pectoralis* was sampled from Chemka Hot Springs in northern Tanzania, close to the Kenyan border. This individual has a deletion that overlaps with the Lake Malawi *Rx1* deletion. On aligning the *C. pectoralis* and *A. baenschi* reads to the *Pundamilia* genome, the *C. pectoralis* deletion length is only 323 bp, instead of the 413 bp found in *A. baenschi* and the other Malawi species (Supp. Fig. S5). Both edges of the deletion are different, with the *C. pectoralis* genome having 94 bp of sequence on the left side that are missing in *A. baenschi* and *A. baenschi* having 4 bp of sequence on the right side that are missing in *C. pectoralis.* Because

of the physical (~750km) and phylogenetic distance of *C. pectoralis* from the Lake Malawi flock, and the differences in indel boundaries, this is likely an independent deletion. This suggests that the *Rx1* regulatory region may be predisposed to deletions.

There is variation in the prevalence of the indels within Lake Malawi. Of the 75 Lake Malawi individuals examined at the *Rx1* locus, 12 taxa had the insertion, 55 had the deletion, and 4 had the ancestral sequence matching *P. pundamilia* and *A. burtoni.* The latter include two *A. calliptera* individuals from Lake Malawi, one *Rhamphochromis* sp from Lake Malawi, and one *A. calliptera* from the Rovuma River. In addition, four of the individuals are heterozygous, with three having both an insertion and a deletion allele (*Protomelas simulans, Tramitichromis brevis,* and *Tyrranochromis maculiceps*) and one having an insertion plus an ancestral allele (*Tramitichromis intermedius*). This is consistent with the fact that the insertion and deletion are quite common across species (except in *Astatotilapia calliptera* which lacks the deletion) and even co-occur within individuals of several species to produce heterozygotes. The *Mitfa* insertion is also quite common. Of 74 individuals in 65 species, 20 were homozygous for the insertion, 12 were heterozygous and 42 had the ancestral sequence. Again, since we are examining single individuals for most of these species, this means that there are at least twelve different species that are polymorphic for the insertion and the ancestral sequence. Interestingly though, the insertion seems absent from *A. calliptera* and also nearly absent from the Mbuna clade.

In contrast to *Rx1* and *Mitfa*, the indels at *Tbx2a* and *SWS1* loci appear to be much less prevalent. For the *Tbx2a* locus, only *M. zebra* had the insertion and only *A. baenschi* had the deletion. Of the four *M. zebra* individuals examined, two from Mazinzi Reef had the insertion

and two from elsewhere in the lake did not. For the deletion, all *A baenschi* individuals that we have previously examined (N=4 lab reared and N=7 wild caught) show the deletion. For the *SWS1* locus, we have found no evidence of any TE insertions, and among the taxa included here only two species (*A. baenschi* and *P. milomo*) had the deletion. However, our previous studies which included more taxa found two additional species with the *SWS1* deletion (*A. stuartgranti* and *Trematocranus placodon*; Nandamuri et al., 2018).

Genomic distribution of TEs

The TE families that we have identified in the insertions include the hAT DNA transposons and the LINE/Rex1/Babar nonLTR transposons. Recent analyses of the highly contiguous PacBio genomes of *M. zebra* and *O. niloticus* indicate that these transposon families are more prevalent in the *M. zebra* genome than in *O. niloticus*. To more broadly examine the prevalence and location of these TEs, we analyzed the top three TE families, which include these two along with the Tc1 Mariner family. Genomic locations were divided between 15 kb promoters, exons, introns and intergenic regions (Fig. 3; see Conte et al., 2019, Table S5). Very few of the TE insertions in either *M. zebra* or *O. niloticus* occurred in exons. For the promoter, intron and intergenic regions, all three TE families had more insertions in *M. zebra* than in *O. niloticus* , except for hAT transposon promoter insertions which occur more frequently in *O. niloticus* . Even for this class, there are still over 2000 instances of promoter insertions in *M. zebra*. Overall, the high prevalence of TE insertions in the *M. zebra* genome, and especially in gene promoters, suggests that they could have an impact on gene expression of a broad range of genes and the opsin related examples described here might be just the tip of an iceberg.

352

353 **Indel origins by divergence estimates**

354 In order to estimate when the TE insertions arose, we compared 1-2 kb flanking
355 sequences to determine when the indel alleles diverged from the ancestral sequence. These
356 regions also have some repetitive sequence, though they vary from only one tiny TE (SWS1) to
357 several TE fragments (Supp. Fig. S6A-D). The Lake Malawi species examined, Genbank accession
358 numbers, and divergences are given in Supp. Table S6 and distance trees are shown in Supp.
359 Fig. S7. The sequences typically do not form monophyletic clades for either of the indel
360 variants. Therefore, we calculate the average divergences for all pairwise allele comparisons.
361 These divergences are then normalized to genome wide divergences for deep lineages within
362 Lake Malawi (AlleleDiv/LM) or the average Malawi-Victoria divergence (AlleleDiv/MV). The
363 values normalized to deep Malawi lineages, AlleleDiv/LM, varied for the different loci. In
364 increasing order they were: 0.52 ± 0.38 (*SWS1*), 0.90 ± 0.24 (*Rx1* deletion), 0.95 ± 0.39 (*Tbx2a*),
365 1.39 ± 0.27 (*Rx1* insertion), and 1.47 ± 0.46 (*Mitfa*) with an average of 1.18 ± 0.29. The values
366 normalized to Malawi-Victoria comparisons, AlleleDiv/MV, were smaller: 0.13 ± 0.10 (*SWS1*),
367 0.23 ± 0.06 (*Rx1* deletion), 0.24 ± 0.10 (*Tbx2a*), 0.35 ± 0.07 (*Rx1* insertion), and 0.37 ± 0.12
368 (*Mitfa*) with an average of 0.30 ± 0.7. Interestingly, the *Rx1* deletion alleles are less diverged
369 and so potentially younger than the *Rx1* insertion alleles, suggesting the insertion may have
370 occurred prior to the deletion. Overall, the combination of flanking allele divergence being less
371 than the Malawi-Victoria split, and of similar magnitude to the deep Malawi split, suggest that
372 all the mutations arose either in the ancestors of the Malawi radiation or early within the
373 Malawi species flock. Therefore, we cannot distinguish with confidence whether the

recruitment of these functional indel alleles represents utilization of pre-existing genetic variation or new mutations in the course of the adaptive radiation. These estimates are consistent with the exclusive distribution of the indels within species of the Malawi radiation.

DISCUSSION

This study explored the genetic variation causing differences in a key cichlid phenotype, visual sensitivity. Although the cichlid genome project suggested several possible sources of regulatory mutation, most of these do not seem to play a role in visual system evolution. Whereas opsin coding sequence differences can be associated with fine-tuning to local light environment (Carleton, Parry, Bowmaker, Hunt, & Seehausen, 2005; Terai et al., 2006), they cannot explain the large shifts in retinal cone cell peak sensitivities observed across the radiations and within the Lake Malawi radiation. These are instead caused by regulatory changes (Hofmann et al., 2009; O'Quin, Hofmann, Hofmann, & Carleton, 2010). There is no evidence for cichlid-specific duplications of opsin genes. The regulatory changes do not result from evolution of miRNA target sites in opsin 3'UTRs (O'Quin, Smith, Naseer, et al., 2011). Instead, eQTL studies suggest that three of the four identified eQTL result from indel mutations in the promoters of transcription factors acting in trans, while one eQTL is a cis regulatory indel mutation in the *SWS1* opsin promoter (Nandamuri, 2018; Nandamuri et al., 2018; O'Quin et al., 2012; Sandkam et al., 2020; Schulte et al., 2014). Our results suggest the movement of TEs explains much of this indel-associated regulatory diversity.

These regulatory indels are correlated with altered gene expression of either *SWS1* opsin or one of the three transcription factors that affect opsin gene expression. The *SWS1* opsin regulatory deletion removes both a conserved noncoding element (CNE) and a miRNA (Nandamuri et al., 2018). These two regulatory elements are conserved across 230 MY of fish evolution (zebrafish, *Danio rerio*, to cichlids). Work in medaka, *Oryzias latipes*, has shown that the *SWS1* opsin and the miRNA (miR-729) are expressed in the same photoreceptor and coregulated by the CNE (Daido, Hamanishi, & Kusakabe, 2014). Therefore, the deletion removes the CNE and miR-729, which affects *SWS1* expression (Nandamuri et al., 2018). The *Rx1* regulatory deletion removes several potential transcription factor binding sites (TFBS) including sites for *Tbx2a* and *Mitfa*. The *Tbx2a* regulatory deletion removes a CNE conserved in sticklebacks and medaka (130 MY divergence; timetree.org) and also contains a TFBS for *Rx1.* Finally, the *Mitfa* regulatory insertion contains a potential *Rx1* TFBS. The presence of these shared TFBS across the different TF loci suggest that these different regulatory regions may work together to co-regulate opsins, switching them on or off at the same time. Several previous studies have demonstrated expression of particular opsins being highly correlated (summarized in Supp. Table S7) across developmental series (Carleton et al., 2008), within F_2 crosses (Nandamuri, Dalton, & Carleton, 2017; O'Quin et al., 2012) and among adult species from Lake Malawi (Hofmann et al., 2009). For example, short palette species express *SWS1* and *RH2B*. However, both these genes are downregulated in the long palette species, as *SWS2A* and *LWS* opsin are upregulated. Some feedback between the loci through key regulatory elements likely work to achieve this co-regulation (Sandkam et al., 2020; Schulte et al., 2014).

Our hypothesis for the ultimate source of the genetic variation generating this regulatory diversity is the movement of TEs, which alters gene expression by insertion (*Mitfa, Rx1*) or excision (*Rx1, Tbx2a, SWS1*) of regulatory regions. While the deleted sequences are not themselves TEs, for two of the deletions we identified lineages that had TE insertions in the exact same location as the deletion. Since the deletions have start and end points surrounding the insertion, this suggests the TE insertions were removed when the deletion occurred. These TE excisions likely removed significant regulatory sequence to generate the large (*Rx1:* 413 and *Tbx2a:* 967 bp) deletions that we observed. Although we did not find a TE insertion associated with the *SWS1* locus, we hypothesize that the large 692 bp deletion may be the result of an insertion and subsequent excision of a mobile TE. These results suggest that TEs are an important factor in changing phenotypes, in this case shifting expression of both transcription factors and their downstream opsin targets.

The origin of these indels, both through phylogenetic distributions and divergence estimates suggests they are unique to the Lake Malawi radiation and a very closely related riverine lineage. In addition, the *Rx1* locus may show convergent evolution, as a similar though distinct deletion arose separately outside of the Malawi flock in a phylogenetically distant lineage (*Ctenochromis pectoralis* of North Eastern Tanzania). This suggests that this regulatory region may be susceptible to TE movement, enabling convergent regulatory mutations, although this does not seem to have happened often.

The idea that TE movement might be important in the evolution of cichlid phenotypes is supported by another study. In cichlid pigmentation patterns, egg spots are a key innovation that occurs on the anal fin of male haplochromine cichlids, and is implicated in mating behavior.

438 In a study of the genetic basis of cichlid egg spots, Santos et al (2014) identified the gene four
439 and a half LIM domain protein 2 (*fhl2b*) as important in egg spot formation. The
440 haplochromines with egg spots had a SINE element in the *fhl2b* promoter that was missing in
441 non-haplochromines, which lacked egg spots. This supports the idea that mobilization of a SINE
442 introduced new regulatory sequence that induced the egg spot phenotype.
443 Previous genome-wide analyses also support the idea of a link between repetitive
444 elements and structural variation in cichlid fishes. The original cichlid genome project found the
445 composition of cichlid genomes to be 16-19% TEs (Brawand et al., 2014). Subsequent long read
446 sequencing has increased those estimates to 35-37% (Conte et al., 2017; Conte et al., 2019). TE
447 insertions near 5'UTRs were associated with increased gene expression in all tissue types. Other
448 studies have examined structural variants in these five genomes and identified both deletions
449 and inversions associated with SINE elements and DNA transposons with some of the structural
450 variation specific to terminal lineages (Fan & Meyer, 2014; Penso-Dolfin, Man, Haerty, & di
451 Palma, 2018). A new study has found a correlation between speciation rates and the
452 occurrence of large genomic indels across several African cichlid flocks (McGee et al., 2020).
453 Transposable elements have been noted to play a key role in evolution (Oliver &
454 Greene, 2009). TEs have been proposed to contribute to introgression (Choudhury & Parisod,
455 2017) as well as reproductive isolation and speciation (Naciri & Linder, 2020; Serrato-Capuchina
456 & Matute, 2018). TEs may also play a possible role in bird diversification (Suh, Smeds, &
457 Ellegren, 2018) and in speciation in amniotes (Zeng et al., 2018), mammals (Ricci, Peona,
458 Guichard, Taccioli, & Boattini, 2018), and fishes (Volff, 2005). TE movement is sometimes
459 ascribed to result from stress in fishes (Auvinet et al., 2018; Symonova et al., 2013). Studies in

very young hybrids show that TE number can increase as a result of tandem duplication in repetitive regions of the genome (Dennenmoser et al., 2019).

One additional consideration is that TE mobility may be enhanced in hybrids. Since TE movement is repressed by PIWI-interacting RNAs (piRNA), hybridization could lead to an incompatibility of parental piRNAs and the corresponding TEs that allow TEs to increase their mobility (Dion-Cote & Barbash, 2017; Dion-Cote, Renaut, Normandeau, & Bernatchez, 2014; Luo & Lu, 2017). Introgression and hybridization have been demonstrated for cichlids in several African lakes (Meier, Sousa, et al., 2017; Salzburger, Baric, & Sturmbauer, 2002; Smith, Konings, & Kornfield, 2003) with some introgression between more divergent lineages (Malinsky et al., 2018; Meier, Marques, et al., 2017; Meier et al., 2019; Svardal et al., 2020). Therefore, past hybridization events between the lineages that seeded the Lake Malawi radiation (Svardal et al., 2020), between major lineages of the radiation, or between some of these and secondary riverine colonists, might contribute to TE increases within this flock. Such increases could contribute to the genetic and phenotypic diversity we find here. Induced TE movement might be one mechanism for how hybridization contributes to adaptive radiation in cichlids (Seehausen, 2004). Further work is needed to compare PIWI interacting RNAs between Lake Malawi cichlid species with varying degrees of divergence or between the ancestral lineages to see whether they have evolved sufficiently to cause mismatches with their target sites.

One important question concerns the degree to which the retention of the TE induced indels found here results from positive selection, relaxed selection, or genetic drift. While estimates of long term effective population sizes (N_e) in Lake Malawi cichlids are relatively high at 50,000 to 130,000 (Malinsky et al., 2018) or >120,000 (Won, Sivasundar, Wang, & Hey, 2005),

many rocky shore species live in highly structured small populations where current N_e may be much smaller and drift much more important. (Husemann, Nguyen, Ding, & Danley, 2015; Won et al., 2005). However, the roles of these indels in underlying large shifts in visual spectral sensitivity suggest a strong selective component, and the Rx1 and Mitf indels show less variation in the rock dwelling Mbuna compared to the sand dwellers, which have larger population sizes. These considerations argue against drift being the main driver. Perhaps relaxed selection followed by an increase in frequency, due to drift, enables these alleles to be subsequently ecologically selected. Finally, the evidence for numerous heterozygotes at these loci also raises the question as to whether balancing selection could play a role. Distinguishing these possibilities will require larger population sampling and population genetic approaches to test for the role of selection on these regulatory regions.

This work has examined many species of haplochromine cichlids from both Lakes Malawi and Victoria as well as a broad swath of species from other lakes and rivers across Eastern Africa. We found substantial evidence for TE induced indels in cichlids from Lake Malawi that correlates with differential opsin expression. In contrast, opsin expression from ten ecologically varied Lake Victoria species suggests they all rely on the long visual palette (Hofmann et al., 2009). This lack of expression variation is consistent with the absence of the target indels from the genomes of over 100 different Lake Victoria species. Although they lack these specific indels, there is some evidence for other indels in these regulatory regions in the *P. nyererei* genome (Supp. Fig. S3). Therefore, further research is still needed before contributions of TE induced indels to opsin expression in Lake Victoria can be completely excluded. One further test for the role of TEs could utilize Lake Tanganyika cichlids to determine

if they have regulatory indels near some of the key genes (*Rx1, Tbx2a, Mitf,* or the opsins themselves). Although we did not detect any of the Malawi indels in the handful of Lake Tanganyika cichlids examined here, Lake Tanganyika cichlids show just as much opsin expression variation as Lake Malawi species (O'Quin et al., 2010). If they do not carry these specific indels, they may have their own indels which shape the diversity of opsin expression in that flock.

Conclusions

Cichlid visual diversity is shaped by mobile transposable elements. We have shown that among the four loci underlying opsin expression differences, three carry insertions which appear to be generated by mobile transposable elements, while one has a large deletion that may be TE induced. In two cases, we find species having either the TE insertions or deletions at the same locus. The boundaries of deleted sequences extend beyond the TE elements. The coincidence of insertions and deletions suggests that these may be linked by the movement of TEs. All indels are recent, in the sense of being either specific to the Malawi radiation or to the radiation and its immediate ancestors, and may make contributions to visual and ecological diversity among Lake Malawi cichlids. Such increased TE movement could be caused by ancestral hybridization. Additionally, relaxed selection allowing mutations to rise in frequency, followed by ecological sorting, could play a role. Therefore, TEs and their associated large indels emerge as an important source of genetic variation for generating visual system diversity. Their impact on other diverse phenotypes and possibly speciation needs further study in this amazingly diverse group.

Acknowledgements: We thank Richard Durbin and Hannes Svardal for helpful discussions. This study was supported by funding from the NIH (1R01EY024639 to KC), the Swiss National Science Foundation (31003A,_163338 to OS and 176039 to Walter Salzburger), an EMBO grant (ALTF 456-2016 to MM), and the US National Science Foundation (DEB-1830753 to TK).

Table 1. Effects of loci controlling opsin expression. This includes whether the variant is an insertion or a deletion, its location, the opsin gene and how it is affected, and the opsin palette.

Locus	Indel	Location	Opsin effect	Palette	Ref
Rx1	Insertion	2.5 kb 5’ of CD	High SWS2A	Long	1
	Deletion	2.5 kb 5’ of CDS	Low SWS2A	Short&Medium	
Mitf	Insertion	Intron 1	High SWS2A	Long	2
Tbx2a	Deletion	13.4 kb 5’ of CDS	Low LWS	Medium	3
SWS1	Deletion	0.75 kb 5’ of CDS	Low SWS1	Medium	4

[1] (Schulte et al., 2014)

[2] (Nandamuri, 2018)

[3] (Sandkam et al., 2020)

[4] (Nandamuri et al., 2018)

References

Akalin, A., Franke, V., Vlahovicek, K., et al. (2015). Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics, 31*(7), 1127-1129. doi:10.1093/bioinformatics/btu775

Albertson, R. C., Streelman, J. T., & Kocher, T. D. (2003). Directional selection has shaped the oral jaws of Lake Malawi cichlid fishes. *Proc Natl Acad Sci U S A, 100*(9), 5252-5257.

Allender, C. J., Seehausen, O., Knight, M. E., et al. (2003). Divergent selection during speciation of Lake Malawi cichlid fishes inferred from parallel radiations in nuptial coloration. *Proc Natl Acad Sci U S A, 100*(24), 14074-14079.

Auvinet, J., Graca, P., Belkadi, L., et al. (2018). Mobilization of retrotransposons as a cause of chromosomal diversification and rapid speciation: the case for the Antarctic teleost genus Trematomus. *BMC Genomics, 19*(1), 339. doi:10.1186/s12864-018-4714-x

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA, 6*, 11. doi:10.1186/s13100-015-0041-9

Barlow, G. W. (2000). *The cichlid fishes: Nature's grand experiment in evolution*. Cambridge, MA: Perseus Publishing.

Brawand, D., Wagner, C. E., Li, Y. I., et al. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature, 513*(7518), 375-381. doi:10.1038/nature13726

Carleton, K. L. (2009). Cichlid fish visual systems: mechanisms of spectral tuning. *Integrative Zoology, 4*, 75-86.

Carleton, K. L., Dalton, B. E., Escobar-Camacho, D., et al. (2016). Proximate and ultimate causes of variable visual sensitivities: Insights from cichlid fish radiations. *Genesis, 54*(6), 299-325. doi:10.1002/dvg.22940

Carleton, K. L., & Kocher, T. D. (2001). Cone opsin genes of african cichlid fishes: tuning spectral sensitivity by differential gene expression. *Mol Biol Evol, 18*(8), 1540-1550.

Carleton, K. L., Parry, J. W., Bowmaker, J. K., et al. (2005). Colour vision and speciation in Lake Victoria cichlids of the genus Pundamilia. *Mol Ecol, 14*(14), 4341-4353.

Carleton, K. L., Spady, T. C., Streelman, J. T., et al. (2008). Visual sensitivities tuned by heterochronic shifts in opsin gene expression. *BMC Biol, 6*(1), 22.

Carroll, S. B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell, 134*(1), 25-36.

Choudhury, R. R., & Parisod, C. (2017). Jumping genes: Genomic ballast or powerhouse of biological diversification. *Mol Ecol, 26*(18), 4587-4590. doi:10.1111/mec.14247

Conith, M. R., Hu, Y., Conith, A. J., et al. (2018). Genetic and developmental origins of a unique foraging adaptation in a Lake Malawi cichlid genus. *Proc Natl Acad Sci U S A, 115*(27), 7063-7068. doi:10.1073/pnas.1719798115

Conte, M. A., Gammerdinger, W. J., Bartie, K. L., et al. (2017). A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions. *BMC Genomics, 18*(1), 341. doi:10.1186/s12864-017-3723-5

Conte, M. A., Joshi, R., Moore, E. C., et al. (2019). Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes. *Gigascience, 8*(4). doi:10.1093/gigascience/giz030

Cortesi, F., Musilova, Z., Stieb, S. M., et al. (2015). Ancestral duplications and highly dynamic opsin gene evolution in percomorph fishes. *Proc Natl Acad Sci U S A, 112*(5), 1493-1498. doi:10.1073/pnas.1417803112

Crow, K. D., Wagner, G. P., & Investigators, S. T.-N. Y. (2006). Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. What is the role of genome duplication in the evolution of complexity and diversity? *Mol Biol Evol, 23*(5), 887-892. doi:10.1093/molbev/msj083

Daido, Y., Hamanishi, S., & Kusakabe, T. G. (2014). Transcriptional co-regulation of evolutionarily conserved microRNA/cone opsin gene pairs: implications for photoreceptor subtype specification. *Dev Biol, 392*(1), 117-129. doi:10.1016/j.ydbio.2014.04.021

Danley, P. D., & Kocher, T. D. (2001). Speciation in rapidly diverging systems: lessons from Lake Malawi. *Mol Ecol, 10*(5), 1075-1086.

Davies, W. I., Collin, S. P., & Hunt, D. M. (2012). Molecular ecology and adaptation of visual photopigments in craniates. *Mol Ecol, 21*(13), 3121-3158. doi:10.1111/j.1365-294X.2012.05617.x

Dennenmoser, S., Sedlazeck, F. J., Schatz, M. C., et al. (2019). Genome-wide patterns of transposon proliferation in an evolutionary young hybrid fish. *Mol Ecol, 28*(6), 1491-1505. doi:10.1111/mec.14969

Dion-Cote, A. M., & Barbash, D. A. (2017). Beyond speciation genes: an overview of genome stability in evolution and speciation. *Curr Opin Genet Dev, 47*, 17-23. doi:10.1016/j.gde.2017.07.014

Dion-Cote, A. M., Renaut, S., Normandeau, E., et al. (2014). RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. *Mol Biol Evol, 31*(5), 1188-1199. doi:10.1093/molbev/msu069

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics, 5*, 113. doi:10.1186/1471-2105-5-113

Fan, S., & Meyer, A. (2014). Evolution of genomic structural variation and genomic architecture in the adaptive radiations of African cichlid fishes. *Front Genet, 5*, 163. doi:10.3389/fgene.2014.00163

Feller, A. F., Haesler, M. P., Peichel, C. L., et al. (2020). Genetic architecture of a key reproductive isolation trait differs between sympatric and non-sympatric sister species of Lake Victoria cichlids. *Proc Biol Sci, 287*(1924), 20200270. doi:10.1098/rspb.2020.0270

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet, 9*(5), 397-405. doi:10.1038/nrg2337

Feulner, P. G. D., Schwarzer, J., Haesler, M. P., et al. (2018). A Dense Linkage Map of Lake Victoria Cichlids Improved the Pundamilia Genome Assembly and Revealed a Major QTL for Sex-Determination. *G3 (Bethesda), 8*(7), 2411-2420. doi:10.1534/g3.118.200207

Gammerdinger, W. J., & Kocher, T. D. (2018). Unusual Diversity of Sex Chromosomes in African Cichlid Fishes. *Genes (Basel), 9*(10). doi:10.3390/genes9100480

Goldsmith, T. H. (2013). Evolutionary tinkering with visual photoreception. *Vis Neurosci, 30*(1-2), 21-37. doi:10.1017/S095252381200003X

Green, R. E., Krause, J., Briggs, A. W., et al. (2010). A draft sequence of the Neandertal genome. *Science, 328*(5979), 710-722. doi:10.1126/science.1188021

Hoekstra, H. E., & Coyne, J. A. (2007). The locus of evolution: evo devo and the genetics of adaptation. *Evolution Int J Org Evolution, 61*(5), 995-1016.

Hofmann, C. M., O'Quin, K. E., Marshall, N. J., et al. (2009). The eyes have it: Regulatory and structural changes both underlie cichlid visual pigment diversity. *PLoS Biol, 7*(12), e1000266.

Husemann, M., Nguyen, R., Ding, B., et al. (2015). A genetic demographic analysis of Lake Malawi rock-dwelling cichlids using spatio-temporal sampling. *Mol Ecol, 24*(11), 2686-2701. doi:10.1111/mec.13205

Jordan, R., Kellogg, K., Howe, D., et al. (2006). Photopigment spectral absorbance of Lake Malawi cichlids. *J Fish Biology, 68*(4), 1291-1299.

Joyce, D. A., Lunt, D. H., Genner, M. J., et al. (2011). Repeated colonization and hybridization in Lake Malawi cichlids. *Curr Biol, 21*(3), R108-109. doi:10.1016/j.cub.2010.11.029

Kocher, T. D. (2004). Adaptive evolution and explosive speciation: the cichlid fish model. *Nat Rev Genet, 5*(4), 288-298.

Kohany, O., Gentles, A. J., Hankus, L., et al. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics, 7*, 474. doi:10.1186/1471-2105-7-474

Konings, A. (2007). *Malawi cichlids in their natural habitat, 4th ed.* (2nd ed.). El Paso, TX: Cichlid Press.

Kratochwil, C. F., Liang, Y., Gerwin, J., et al. (2018). Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations. *Science, 362*(6413), 457-460. doi:10.1126/science.aao6809

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Levine, J. S., MacNichol, E. F., Jr., Kraft, T., et al. (1979). Intraretinal distribution of cone pigments in certain teleost fishes. *Science, 204*(4392), 523-526.

Luo, S., & Lu, J. (2017). Silencing of Transposable Elements by piRNAs in Drosophila: An Evolutionary Perspective. *Genomics Proteomics Bioinformatics, 15*(3), 164-176. doi:10.1016/j.gpb.2017.01.006

Maan, M. E., Seehausen, O., & Groothuis, T. G. (2017). Differential Survival between Visual Environments Supports a Role of Divergent Sensory Drive in Cichlid Fish Speciation. *Am Nat, 189*(1), 78-85. doi:10.1086/689605

Malinsky, M., Challis, R. J., Tyers, A. M., et al. (2015). Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science, 350*(6267), 1493-1498. doi:10.1126/science.aac9927

Malinsky, M., & Salzburger, W. (2016). Environmental context for understanding the iconic adaptive radiation of cichlid fishes in Lake Malawi. *Proc Natl Acad Sci U S A, 113*(42), 11654-11656. doi:10.1073/pnas.1614272113

Malinsky, M., Svardal, H., Tyers, A. M., et al. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat Ecol Evol, 2*(12), 1940-1955. doi:10.1038/s41559-018-0717-x

McGee, M. D., Borstein, S. R., Meier, J. I., et al. (2020). The ecological and genomic basis of explosive adaptive radiation. *Nature*. doi:10.1038/s41586-020-2652-7

McGee, M. D., Borstein, S. R., Neches, R. Y., et al. (2015). A pharyngeal jaw evolutionary innovation facilitated extinction in Lake Victoria cichlids. *Science, 350*(6264), 1077-1079. doi:10.1126/science.aab0800

McKenna, A., Hanna, M., Banks, E., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Meier, J. I., Marques, D. A., Mwaiko, S., et al. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat Commun, 8*, 14363. doi:10.1038/ncomms14363

Meier, J. I., Sousa, V. C., Marques, D. A., et al. (2017). Demographic modelling with whole-genome data reveals parallel origin of similar Pundamilia cichlid species after hybridization. *Mol Ecol, 26*(1), 123-141. doi:10.1111/mec.13838

Meier, J. I., Stelkens, R. B., Joyce, D. A., et al. (2019). The coincidence of ecological opportunity with hybridization explains rapid adaptive radiation in Lake Mweru cichlid fishes. *Nat Commun, 10*(1), 5391. doi:10.1038/s41467-019-13278-z

Mills, R. E., Luttig, C. T., Larkins, C. E., et al. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res, 16*(9), 1182-1190. doi:10.1101/gr.4565806

Naciri, Y., & Linder, H. P. (2020). The genetics of evolutionary radiations. *Biol Rev Camb Philos Soc*. doi:10.1111/brv.12598

Nagai, H., Terai, Y., Sugawara, T., et al. (2011). Reverse evolution in RH1 for adaptation of cichlids to water depth in Lake Tanganyika. *Mol Biol Evol, 28*(6), 1769-1776. doi:10.1093/molbev/msq344

Nandamuri, S. P. (2018). *Mechanisms contributing to opsin expression divergence in the visual system of African cichlids.* (Ph.D.), University of Maryland, College Park, MD.

Nandamuri, S. P., Conte, M. A., & Carleton, K. L. (2018). Multiple trans QTL and one cis-regulatory deletion are associated with the differential expression of cone opsins in African cichlids. *BMC Genomics, 19*(1), 945. doi:10.1186/s12864-018-5328-z

Nandamuri, S. P., Dalton, B. E., & Carleton, K. L. (2017). Determination of the Genetic Architecture Underlying Short Wavelength Sensitivity in Lake Malawi Cichlids. *J Hered, 108*(4), 379-390. doi:10.1093/jhered/esx020

O'Leary, N. A., Wright, M. W., Brister, J. R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res, 44*(D1), D733-745. doi:10.1093/nar/gkv1189

O'Quin, K. E., Hofmann, C. M., Hofmann, H. A., et al. (2010). Parallel evolution of opsin gene expression in African cichlid fishes. *Mol Biol Evol, 27*(12), 2839-2854.

O'Quin, K. E., Schulte, J. E., Patel, Z., et al. (2012). Evolution of cichlid vision via trans-regulatory divergence. *BMC Evol Biol, 12*, 251. doi:10.1186/1471-2148-12-251

O'Quin, K. E., Smith, A. R., Sharma, A., et al. (2011). New evidence for the role of heterochrony in the repeated evolution of cichlid opsin expression. *Evol Dev, 13*(2), 193-203.

O'Quin, K. E., Smith, D., Naseer, Z., et al. (2011). Divergence in cis-regulatory sequences surrounding the opsin gene arrays of African cichlid fishes. *BMC Evol Biol, 11*, 120.

Ohno, S. (1970). *Evolution by Gene Duplication*. New York, NY: Springer-Verlag.

Oliver, K. R., & Greene, W. K. (2009). Transposable elements: powerful facilitators of evolution. *Bioessays, 31*(7), 703-714. doi:10.1002/bies.200800219

Otto, S. P., & Whitton, J. (2000). Polyploid incidence and evolution. *Annu Rev Genet, 34*, 401-437. doi:10.1146/annurev.genet.34.1.401

Parry, J. W., Carleton, K. L., Spady, T., et al. (2005). Mix and match color vision: tuning spectral sensitivity by differential opsin gene expression in Lake Malawi cichlids. *Curr Biol, 15*(19), 1734-1739.

Parsons, K. J., Trent Taylor, A., Powder, K. E., et al. (2014). Wnt signalling underlies the evolution of new phenotypes and craniofacial variability in Lake Malawi cichlids. *Nat Commun, 5*, 3629. doi:10.1038/ncomms4629

Penso-Dolfin, L., Man, A., Haerty, W., et al. (2018). *Analysis of structural variants in four African Cichlids highlights an association with developmental and immune related genes*. bioRXiv.

Ricci, M., Peona, V., Guichard, E., et al. (2018). Transposable Elements Activity is Positively Related to Rate of Speciation in Mammals. *J Mol Evol, 86*(5), 303-310. doi:10.1007/s00239-018-9847-7

Roberts, R., Ser, J., & Kocher, T. D. (2009). Genetic basis of a sexual conflict in Lake Malawi cichlids. *Science, e-pub*.

Roberts, R. B., Hu, Y., Albertson, R. C., et al. (2011). Craniofacial divergence and ongoing adaptation via the hedgehog pathway. *Proc Natl Acad Sci U S A, 108*(32), 13194-13199. doi:10.1073/pnas.1018456108

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., et al. (2011). Integrative genomics viewer. *Nat Biotechnol, 29*(1), 24-26. doi:10.1038/nbt.1754

Salzburger, W., Baric, S., & Sturmbauer, C. (2002). Speciation via introgressive hybridization in East African cichlids? *Mol Ecol, 11*(3), 619-625.

Sandkam, B. A., Campello, L., O'Brien, C., et al. (2020). *Tbx2a* modulates switching of RH2 and LWS opsin gene expression. *Mol Biol Evol, 37*(7), 2002-2014.

Santos, M. E., Braasch, I., Boileau, N., et al. (2014). The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. *Nat Commun, 5*, 5149. doi:10.1038/ncomms6149

Schulte, J. E., O'Brien, C. S., Conte, M. A., et al. (2014). Interspecific Variation in Rx1 Expression Controls Opsin Expression and Causes Visual System Diversity in African Cichlid Fishes. *Mol Biol Evol, 31*(9), 2297-2308. doi:10.1093/molbev/msu172

Seehausen, O. (1996). *Lake Victoria rock cichlids*. Germany: Verduijn Cichlids.

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends Ecol Evol, 19*(4), 198-207.

Seehausen, O. (2006). African cichlid fish: a model system in adaptive radiation research. *Proc Biol Sci, 273*(1597), 1987-1998.

Seehausen, O., Terai, Y., Magalhaes, I. S., et al. (2008). Speciation through sensory drive in cichlid fish. *Nature, 455*, 620-626.

Sefc, K. M. (2011). Mating and Parental Care in Lake Tanganyika's Cichlids. *Int J Evol Biol, 2011*, 470875. doi:10.4061/2011/470875

Serrato-Capuchina, A., & Matute, D. R. (2018). The Role of Transposable Elements in Speciation. *Genes (Basel), 9*(5). doi:10.3390/genes9050254

Servedio, M. R., Van Doorn, G. S., Kopp, M., et al. (2011). Magic traits in speciation: 'magic' but not rare? *Trends Ecol Evol, 26*(8), 389-397. doi:10.1016/j.tree.2011.04.005

Smit, A. F. A., & Hubley, R. (2010). RepeatModeler: www.repeatmasker.org.
Smit, A. F. A., Hubley, R., & Green, P. (2010). RepeatMasker: www.repeatmasker.org.
Smith, P. F., Konings, A., & Kornfield, I. (2003). Hybrid origin of a cichlid population in Lake Malawi: implications for genetic variation and species diversity. *Mol Ecol, 12*(9), 2497-2504.
Spady, T. C., Parry, J. W., Robinson, P. R., et al. (2006). Evolution of the cichlid visual palette through ontogenetic subfunctionalization of the opsin gene arrays. *Mol Biol Evol, 23*(8), 1538-1547.
Spady, T. C., Seehausen, O., Loew, E. R., et al. (2005). Adaptive molecular evolution in the opsin genes of rapidly speciating cichlid species. *Mol Biol Evol, 22*(6), 1412-1422.
Stern, D. L., & Orgogozo, V. (2008). The loci of evolution: how predictable is genetic evolution? *Evolution, 62*(9), 2155-2177.
Sugawara, T., Terai, Y., Imai, H., et al. (2005). Parallelism of amino acid changes at the RH1 affecting spectral sensitivity among deep-water cichlids from Lakes Tanganyika and Malawi. *Proc Natl Acad Sci U S A, 102*(15), 5448-5453.
Suh, A., Smeds, L., & Ellegren, H. (2018). Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. *Mol Ecol, 27*(1), 99-111. doi:10.1111/mec.14439
Svardal, H., Quah, F. X., Malinsky, M., et al. (2020). Ancestral Hybridization Facilitated Species Diversification in the Lake Malawi Cichlid Fish Adaptive Radiation. *Mol Biol Evol, 37*(4), 1100-1113. doi:10.1093/molbev/msz294
Symonova, R., Majtanova, Z., Sember, A., et al. (2013). Genome differentiation in a species pair of coregonine fishes: an extremely rapid speciation driven by stress-activated retrotransposons mediating extensive ribosomal DNA multiplications. *BMC Evol Biol, 13*, 42. doi:10.1186/1471-2148-13-42
Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol, 10*(3), 512-526. doi:10.1093/oxfordjournals.molbev.a040023
Terai, Y., Seehausen, O., Sasaki, T., et al. (2006). Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. *PLoS Biol, 4*(12), e433.
Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform, 14*(2), 178-192. doi:10.1093/bib/bbs017
Turner, T. L., & Hahn, M. W. (2010). Genomic islands of speciation or genomic islands and speciation? *Mol Ecol, 19*(5), 848-850. doi:10.1111/j.1365-294X.2010.04532.x
van der Meer, H. J., & Bowmaker, J. K. (1995). Interspecific variation of photoreceptors in four co-existing haplochromine cichlid fishes. *Brain Behav Evol, 45*(4), 232-240.
Volff, J. N. (2005). Genome evolution and biodiversity in teleost fish. *Heredity (Edinb), 94*(3), 280-294. doi:10.1038/sj.hdy.6800635
Won, Y. J., Sivasundar, A., Wang, Y., et al. (2005). On the origin of Lake Malawi cichlid species: a population genetic analysis of divergence. *Proc Natl Acad Sci U S A, 102 Suppl 1*, 6581-6586.
Yokoyama, S. (2008). Evolution of dim-light and color vision pigments. *Annu Rev Genomics Hum Genet, 9*, 259-282.

Zeng, L., Pederson, S. M., Kortschak, R. D., et al. (2018). Transposable elements and gene expression during the evolution of amniotes. *Mob DNA, 9*, 17. doi:10.1186/s13100-018-0124-5

Data Accessibility: Sanger sequencing results are deposited in Genbank (accession numbers in Supp Table S6). Illumina sequences are deposited in the NBCI Short Read Archive based on previous publications.

Author contributions: This study was conceived by KC, OS, MM, and TK. Genomic analyses were performed by MC, MM, JIM and KC. Laboratory analyses were done by SPN, BS, KC, and SM. All authors contributed to and approved the manuscript.

Figure legends

Figure 1. Genomic variation in putative causative mutations underlying changes in opsin expression. Genomic regions for three long palette outgroup species (in black: the tilapia, *Oreochromis niloticus, Astatotilapia burtoni,* and *Pundamilia nyererei*) are compared to species from Lake Malawi (in gray: *Metriaclima zebra*, *Aulonocara baenschi, Dimidiochromis compressiceps* and *Tramitichromis intermedius*). Dots are used to show which opsin genes are expressed in each visual palette in adult fish. These include the short (SWS1, RH2B, RH2A), medium (SWS2B, RH2B, RH2A) and long (SWS2A, RH2A, LWS) palettes. Deletions are shown as dashed boxes while insertions are shown as triangles. For simplicity, we show only the deletions related to our eQTL. However, there are other indels among these taxa as shown in Supp Fig S3.

Figure 2. Phylogenetic relationships of cichlid taxa studied along with the state of the four indels across 239 individuals from 209 species (207 haplochromines plus two non-haplochromine outgroups). Insertions are noted in color with blue (Rx1), green (Mitfa), and red (Tbx2a). Deletions are shown in gray or yellow (Rx1). Ancestral alleles having neither an insertion nor a deletion are shown in black. If known from PCR, both alleles for each individual are shown next to each other. Both homozygous (alleles same color) and heterozygous (alleles in different color) individuals occur. For species outside of Lake Malawi, the number of individuals sampled in particular clades are noted in parentheses. Within Lake Malawi, known opsin expression palettes for individuals are shown by the species name color coded as the short, medium or long palettes, as shown by the photoreceptors with their associated expressed genes. Names shown in black have unknown expression palettes. The tree is based on Meier et al (2017; Supp. Fig. 1).

Figure 3. Top three families of transposable elements in cichlids, based on their genomic location. The data either comes from the tilapia, *O. niloticus* (Til) or *M. zebra* (Mz). The TE families include the DNA transposons Tc1 Mariner (TcMar) and hAT, and the retrotransposon, Rex1. Therefore, Mz_TcMar is the number of transposable elements in the *M. zebra* genome in the Tc1 Mariner family. Data provided in Supp. Table S5.

Legends for Supplementary Figures and Tables

Supp Fig S1. Tree from Meier et al 2017 RAD tree with nodes labeled by groups for comparison with cichlid taxa in Supp Fig S2.

Supp Fig S2. Tree of 58 taxa including 52 cichlid samples with genome sequences searched in this project. Groups are labeled by the homologous group that they match from Supp Fig S1.

Supp Fig S3A. Transposable element and indel locations for the broader Rx1 region. Species include tilapia *O. niloticus* (Til), *M. zebra* (Mz), *P. nyererei* (Pny), *A. burtoni* (Aburt), *Tramitichromis intermedius* (Tint), *Aulonocara baenschi* (Abae), *Melanochromis auratus* (Maur), *Neolamprologus brichardi* (Nbrich) and *Dimidiochromis compressiceps* (Dcomp). Translational start site is at 3795 bp in tilapia (yellow box). The key regulatory variation is the insertion in *D. compressiceps* and the corresponding deletion (green box). However, there is an additional large insertion in *T. intermedius* and smaller indels (*) in tilapia and *P. nyererei.*
Supp Fig S3B. Transposable element and indel locations for the broader Tbx2a region. Species include tilapia *O. niloticus* (Til), *A. burtoni* (Aburt), *M. zebra* (Mz), *Aulonocara baenschi* (Aulonocara) and *Tramitichromis intermedius* (Tramitichromis). Tbx2a exons are shown as black boxes in the bottom row, which start near 18731 bp of tilapia (yellow arrow). The key regulatory variation is the insertion in *M. zebra (*gray arrow) and the corresponding deletion in *Aulonocara* (green arrow). However, there are small insertions in *A. burtoni,* and tilapia (*).

Supp Fig S3C. Transposable element and indel locations for the broader Mitf region. Species include tilapia *O. niloticus* (Til), *M. zebra* (Mz), *A. burtoni* (Aburt), *P. nyererei* (Pnye), and *Tramitichromis intermedius* (Tint). Mitf exons are shown in yellow starting at tilapia base 117. The key regulatory variation is the insertion in *T. intermedius* in intron 1 (gray arrow). However, there are other indels in *A. burtoni* and tilapia (*).

Supp Fig S3D. Indel locations for the broader SWS1 region. Species include the tilapia *O. niloticus* (Til), *P. nyererei* (Pnye), *A. burtoni* (Abur), *M. zebra* (Mz), *D. compressiceps* (Dim compressiceps), *Tramitichromis intermedius* (Tintermed) and *A. baenschi* (Aul baenschi). The first SWS1 exon is shown in yellow starting at tilapia base 2443. The key regulatory change is the deletion in *A. baenschi* which overlaps the key regulatory elements shown in red, miRNA-729 and the CNE. However, there is a large insertion in tilapia (*).

Supp Fig S4A. Transposable element and indel locations for the Rx1 region for *D. compressiceps* (Dcomp) aligned with *M. zebra* (Mz), *A. burtoni* (Ab), *P. nyererei* (Pny) and the tilapia, *O. niloticus* (Til). Translational start site is at 4189 bp (not shown).

Supp Fig S4B. Transposable element and indel locations for the Tbx2a upstream regulatory region for tilapia, *O. niloticus* (Til), *A. burtoni* (Aburt), *P. nyererei* (Pny), *Aulonocara baenschi* (Aulonocara), *Tramitichromis intermedius* (Tramitichromis), and *M. zebra* (Mz). Translational start site is at 27,130 bp (not shown).

Supp Fig S4C. Transposable element and indel locations for the Mitf region for *Tramitichromis intermedius* (Tint), *Aulonocara baenschi* (Aulonocara), and *M. zebra* (Mzebra), *A. burtoni* (Aburt), *P. nyererei* (Pnye), and tilapia, *O. niloticus* (Til). Exons shown as yellow arrows.

Supp Fig S4D. TE vs indel locations. SWS1 region for the tilapia *O. niloticus* (Til), *A. burtoni* (Aburt), *P. nyererei* (Pnye), *M. zebra* (Mz), *Tramitichromis intermedius* (Tintermedius), and *Aulonocara baenschi* (Aul baenschi). Exon 1 is shown as yellow arrow while miRNA-729 is shown as red arrow.

Supp Fig S5. Comparison of deletion in the Rx1 enhancer. Reads for *Ctenochromis pectoralis* (Cpect), *Aulonocara baenschi* (Ab411), and *Lithochromis rubripinnis* (species 2801) aligned against the *Pundamilia nyererei* genome. Region on LG10: 38,887,100-38,887,750).
A) Read mapping overview showing the contiguous region for *L. rubripinnis* as opposed to the unique deletions for *C. pectoralis* (323bp) as compared to *A baenschi* (413 bp).

Supp Fig S5B. Close up of read mapping that shows the left hand side of the deletion where the *C. pectoralis* sequence has 94 bp that are missing in *A baenschi.*

Supp Fig S5C. Close up read mapping that shows the right hand side of the deletion where the *A baenschi* sequence has 4 bp that are missing in *C. pectoralis.*

Supp Fig S6A. Transposable elements annotated in the Rx1 sequence surrounding the indel used to date the indel origin (note indel sequences are removed in this alignment). A subset of the 18 species studied are shown here including *Rhamphochromis esox, Placidochromis johnstoni, Trematocranus placodon, Labidichromis gigas, Protomelas fenestratus, Cynotilapia afra and Aulonocara baenschi.* The TE sequences are annotated in *D. compressiceps* (Dcomp), shown in gray, and shared across species.

Supp Fig S6B. Transposable elements annotated in the Tbx2a sequence surrounding the indel used to date the deletion origin (note indel sequences are removed in this alignment). Region shown for *P. nyererei* (Pny), *M. zebra* (Mz), *Aulonocara baenschi* (Aulonocara), and *Tramitichromis intermedius* (Tramitichromis). The TEs are noted in gray and shared across species.

Supp Fig S6C. Transposable elements annotated in the Mitf sequence surrounding the indel used to date the insertion origin (note inserted sequences are removed in this alignment). The alignment includes *Pundamilia nyererei* and 12 Lake Malawi species including *Placidochromis johnstoni, Lethrinops aurita, Nimbochromis linni, Tyrannochromis macrostoma, Aristichromis chrystyi, Dimidiochromis compressiceps, Hemichromis oxyrhynchus, Melanochromis auratus, Labeotropheus trewavasae, M 'black & white' johanni, Labidochromis gigas, Metriaclima zebra.* Exons shown as yellow arrows and TE elements in gray.

Supp Fig S6D. Transposable elements annotated in the SWS1 sequence surrounding the indel used to date the deletion's origin (note deleted region are removed in this alignment). SWS1

region alignment for *P. nyererei* (Pnye), and 10 of the studied Lake Malawi cichlid species including *Melanochromis auratus, M 'black & white' johanni, Aristichromis chrystyi, Hemichromis oxyrhynchus, Trematocranus placodon, Lethrinops aurita, Dimidiochromis compressiceps, Rhamphochromis esox, and Cynotilapia afra.* The one TE is noted in gray and shared across all species.

Supp Fig S7A. Tree based on 1400 bp of sequence surrounding the Rx1 indel. This enables us to compare the distance between species either with an insertion (INS; boxed species) or with the deletion (DEL) to the ancestral *A calliptera* sequence. The ratio of this distance to the pelagic Malawi species or the Victorian species, *P. nyererei* is then determined.

Supp Fig S7B. Tree based on 1800 bp of sequence downstream of the Mitf insertion. This enables us to compare the distance between species with an insertion (Ins; boxed species) to those without (wild type, WT). The ratio of this distance to the pelagic Malawi species or the Victorian species, *P. nyererei* is then determined.
Supp Fig S7C. Tree based on 1000 bp of sequence surrounding the SWS1 deletion in *A. baenschi* (in box) to other Lake Malawi species. The ratio of this distance to the pelagic Malawi species or the Victorian species, *P. nyererei* is then determined.

Supp Fig S7D. Tree based on 1800 bp of sequence surrounding the Tbx2a deletion in *A. baenschi (*(in box) as compared to *T intermedius* and *M zebra.* The ratio of this distance to the pelagic Malawi species or the Victorian species, *P. nyererei* is then determined.

Supp Table S1. Genotypes at four indel loci for all 239 individuals.

Supp Table S2. Primers for PCR screening for indels and for sequencing nearby regions

Supp Table S3. Location of regulatory sequences in *Metriaclima zebra* UMD2a genome (Conte

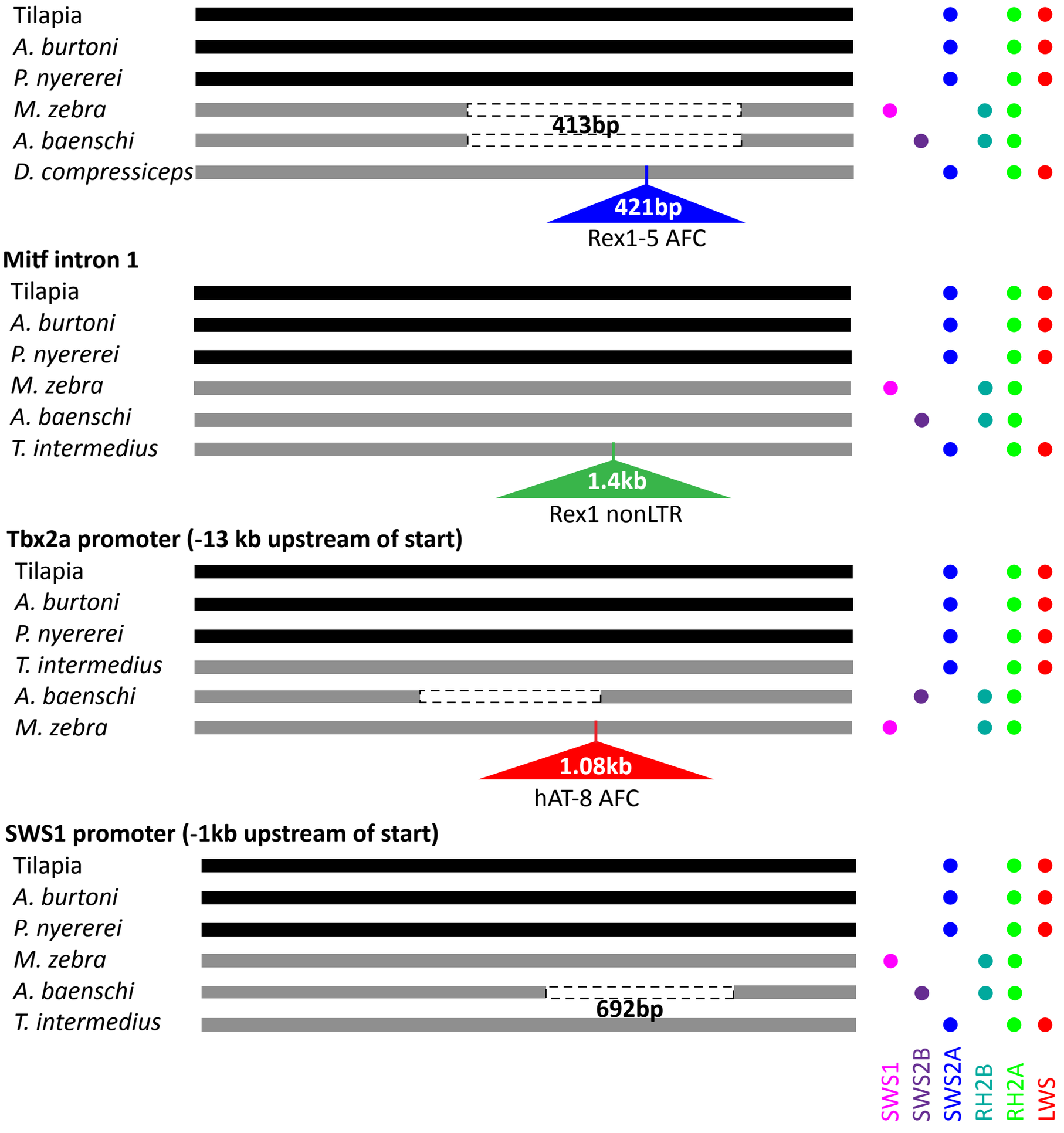
et al 2019).

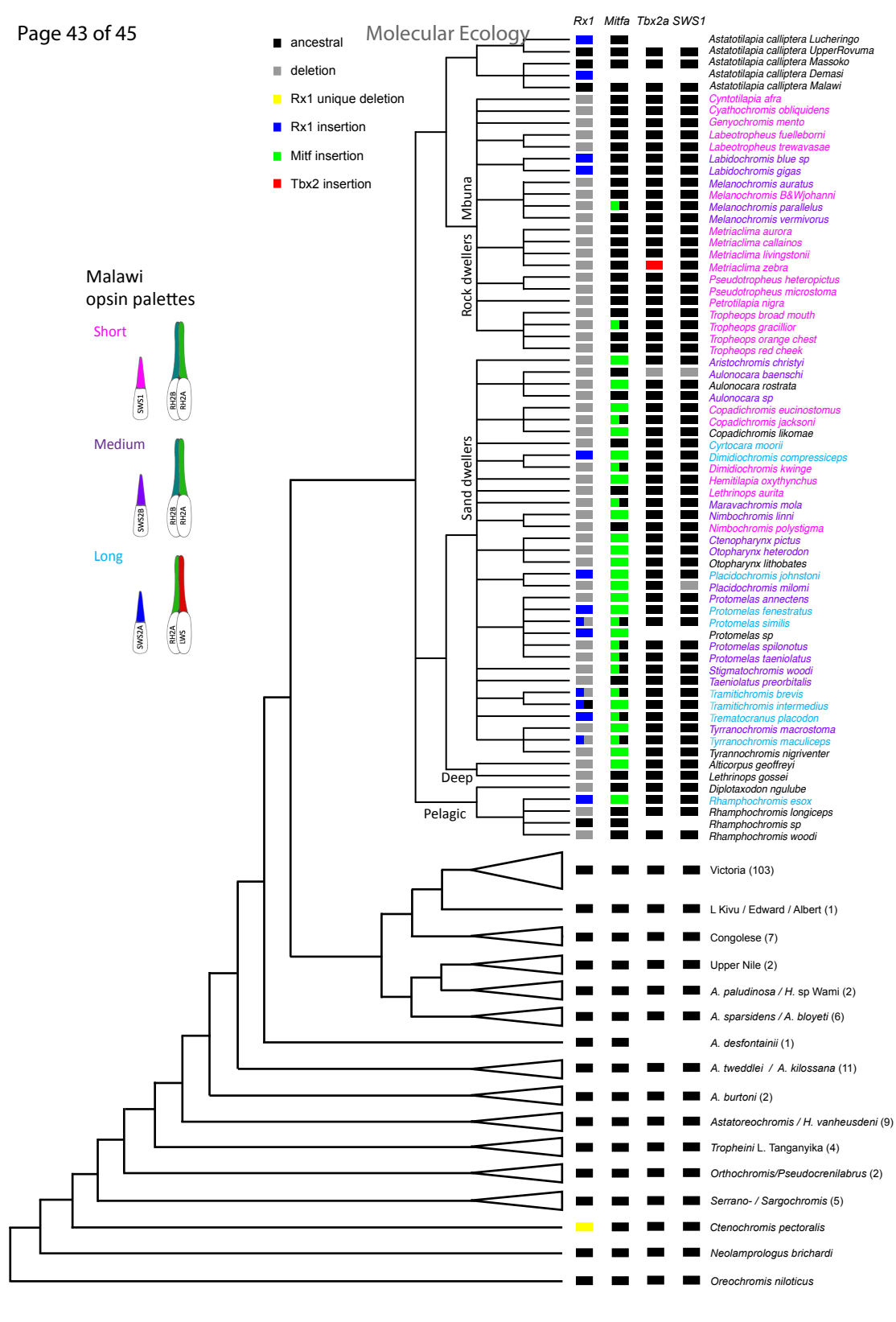
Supp Table S4. Sequences for four genomic indel regions. For the SWS1 and Tbx2a loci, this lists the sequence that is present in *M. zebra* genome (UMD2a; Conte et al 2019) but missing in *A. baenschi*. For the Mitf locus, it denotes the location of the insertion in different species including *T. intermedius*. For the Rx1 locus, it shows the location of the deletion which is shared by *M. zebra* and other short or medium palette species.

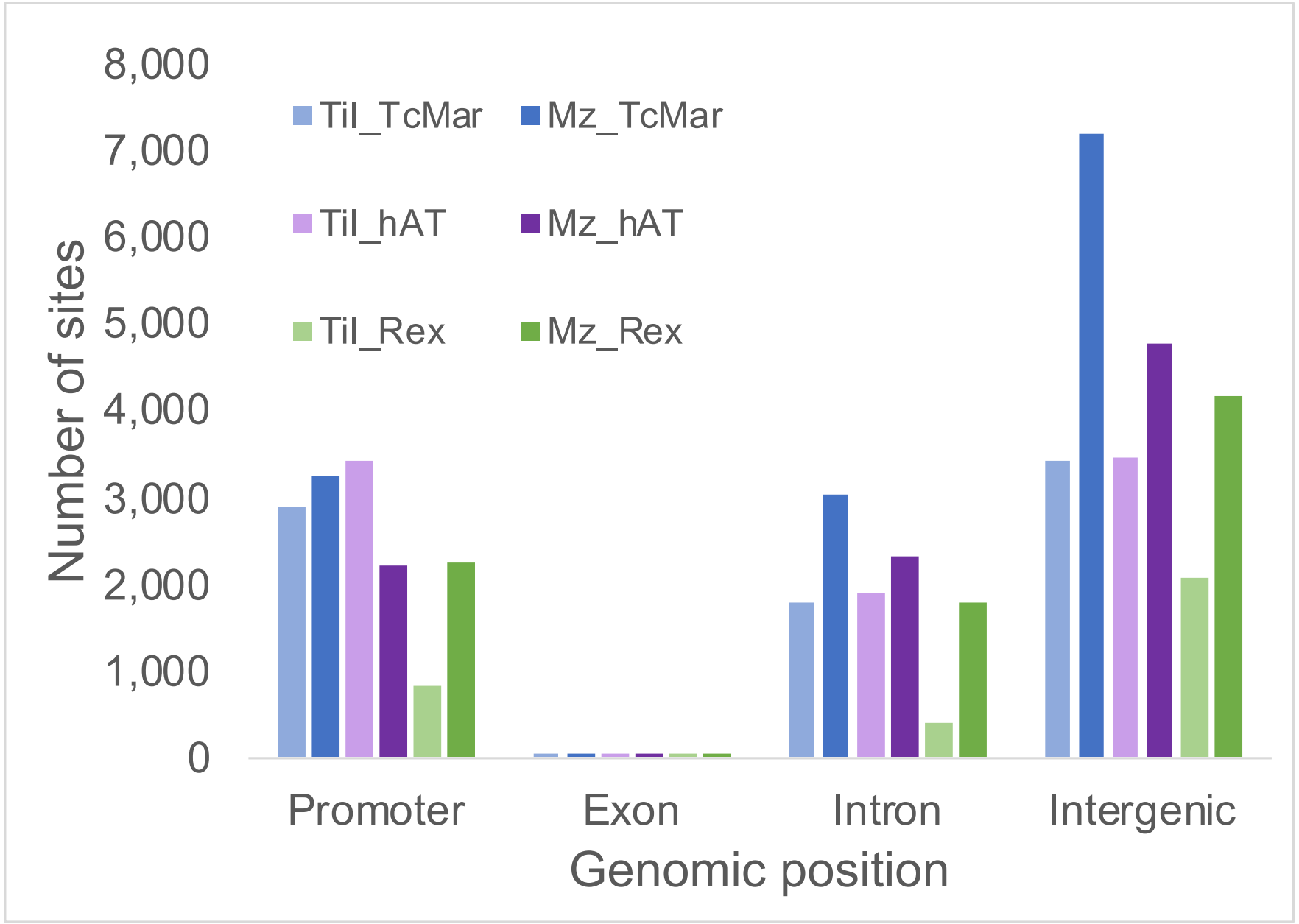
Supp Table S5. Frequency and location of recent insertions for three TE families in the genomes of *Oreochromis niloticus* (On) and *Metriaclima zebra* (Mz).

Supp Table S6: The average divergence between indel alleles and wild type alleles. These are compared to known divergences for deep divergences within Lake Malawi and to differences between Lakes Malawi and Victoria.

Supp Table S7. Significant correlations between opsin genes from previous work. Correlational data includes ontogenetic series for *Oreochromis niloticus*[1], F_2 genetic crosses for *Tramitichromis intermedius x Aulonocara baenschi*[2] and *Metriaclima mbenjii* x *Aulonocara baenschi*[3], and data for 54 species from Lake Malawi[4]. For each gene combination, the correlation coefficient (R) and the p value are given. Only correlations with a p value <0.0033 (0.05 / 15 comparisons) were included to correct for 15 multiple tests within each dataset.


Rx1 promoter (-2.5 kb upstream of start)
Opsin palette
Tilapia
A. burtoni
P. nyererei
M. zebra
A. baenschi
D. compressiceps
413bp
421bp
Rex1-5 AFC
Mitf intron 1
Tilapia
A. burtoni
P. nyererei
M. zebra
A. baenschi
T. intermedius
1.4kb
Rex1 nonLTR
Tbx2a promoter (-13 kb upstream of start)
Tilapia
A. burtoni
P. nyererei
T. intermedius
A. baenschi
M. zebra
1.08kb
hAT-8 AFC
SWS1 promoter (-1kb upstream of start)
Tilapia
A. burtoni
P. nyererei
M. zebra
A. baenschi
T. intermedius
692bp
SWS1
SWS2B
SWS2A
RH2B
RH2A
LWS


ancestral
deletion
Rx1 unique deletion
Rx1 insertion
Mitf insertion
Tbx2 insertion
Rx1
Mitfa
Tbx2a
SWS1
Malawi
opsin palettes
Short
SWS1
RH2B
RH2A
Medium
SWS2B
RH2B
RH2A
Long
SWS2A
RH2A
LWS
Mbuna
Rock dwellers
Sand dwellers
Deep
Pelagic
Astatotilapia calliptera Lucheringo
Astatotilapia calliptera UpperRovuma
Astatotilapia calliptera Massoko
Astatotilapia calliptera Demasi
Astatotilapia calliptera Malawi
Cyntotilapia afra
Cyathochromis obliquidens
Genyochromis mento
Labeotropheus fuelleborni
Labeotropheus trewavasae
Labidochromis blue sp
Labidochromis gigas
Melanochromis auratus
Melanochromis B&Wjohanni
Melanochromis parallelus
Melanochromis vermivorus
Metriaclima aurora
Metriaclima callainos
Metriaclima livingstonii
Metriaclima zebra
Pseudotropheus heteropictus
Pseudotropheus microstoma
Petrotilapia nigra
Tropheops broad mouth
Tropheops gracillior
Tropheops orange chest
Tropheops red cheek
Aristochromis christyi
Aulonocara baenschi
Aulonocara rostrata
Aulonocara sp
Copadichromis eucinostomus
Copadichromis jacksoni
Copadichromis likomae
Cyrtocara moorii
Dimidiochromis compressiceps
Dimidiochromis kwinge
Hemitilapia oxythynchus
Lethrinops aurita
Maravachromis mola
Nimbochromis linni
Nimbochromis polystigma
Ctenopharynx pictus
Otopharynx heterodon
Otopharynx lithobates
Placidochromis johnstoni
Placidochromis milomi
Protomelas annectens
Protomelas fenestratus
Protomelas similis
Protomelas sp
Protomelas spilonotus
Protomelas taeniolatus
Stigmatochromis woodi
Taeniolatus preorbitalis
Tramitichromis brevis
Tramitichromis intermedius
Trematocranus placodon
Tyrranochromis macrostoma
Tyrranochromis maculiceps
Tyrannochromis nigriventer
Alticorpus geoffreyi
Lethrinops gossei
Diplotaxodon ngulube
Rhamphochromis esox
Rhamphochromis longiceps
Rhamphochromis sp
Rhamphochromis woodi
Victoria (103)
L Kivu / Edward / Albert (1)
Congolese (7)
Upper Nile (2)
A. paludinosa / H. sp Wami (2)
A. sparsidens / A. bloyeti (6)
A. desfontainii (1)
A. tweddlei / A. kilossana (11)
A. burtoni (2)
Astatoreochromis / H. vanheusdeni (9)
Tropheini L. Tanganyika (4)
Orthochromis/Pseudocrenilabrus (2)
Serrano- / Sargochromis (5)
Ctenochromis pectoralis
Neolamprologus brichardi
Oreochromis niloticus


8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
Number of sites
Til_TcMar
Mz_TcMar
Til_hAT
Mz_hAT
Til_Rex
Mz_Rex
Promoter
Exon
Intron
Intergenic
Genomic position