Evolution of miRNA binding sites and regulatory networks in cichlids

Tarang K. Mehta[1]*, Luca Penso-Dolfin[2], Will Nash[1], Sushmita Roy[3,4,5], Federica Di-Palma[6,7], Wilfried Haerty[1,6]

[1] Earlham Institute (EI), Norwich, UK

[2] Silence Therapeutics GmbH, Robert-Rössle-Straße 10, 13125 Berlin, Germany

[3] Dept. of Biostatistics and Medical Informatics, UW Madison, Madison, USA

[4] Wisconsin Institute for Discovery (WID), Madison, USA

[5] Dept. of Computer Sciences, UW Madison, Madison, USA

[6] School of Biological Sciences, University of East Anglia, Norwich, UK

[7] Genome British Columbia, Vancouver, Canada

* Corresponding author

Abstract

The divergence of regulatory regions and gene regulatory network (GRN) rewiring is a key driver of cichlid phenotypic diversity. However, the contribution of miRNA binding site turnover has yet to be linked to GRN evolution across cichlids. Here, we extend our previous studies by analysing the selective constraints driving evolution of miRNA and transcription factor binding sites (TFBSs) of target genes, to infer instances of cichlid GRN rewiring associated with regulatory binding site turnover. Comparative analyses identified increased species-specific networks that are functionally associated to traits of cichlid phenotypic diversity. The evolutionary rewiring is associated with differential models of transcription factor (TF) and miRNA binding site turnover, driven by a high proportion of fast-evolving polymorphic sites in adaptive trait genes compared to subsets of random genes. Positive selection acting upon discrete mutations in these regulatory regions is likely to be an important mechanism in rewiring GRNs in rapidly radiating cichlids. Regulatory variants of functionally associated TF and miRNA binding sites of visual opsin genes differentially segregate according to phylogeny and ecology of Lake Malawi species, identifying both rewired e.g. clade-specific and conserved network motifs of adaptive trait associated GRNs. Our approach revealed several novel candidate regulators, regulatory regions and three-node motifs across cichlid genomes with previously reported associations to known adaptive evolutionary traits.

Keywords

miRNA; gene regulatory network; cichlid; molecular evolution

Introduction

The molecular 'tinkering' of ancestral systems and divergence of gene regulatory processes is a hallmark of evolution, and has long been thought to be associated with morphological diversity [27,32,51,64]. Based on these theories, a number of studies have focused on gene regulatory networks (GRNs) with the aim of relating gene expression variation to phenotypic divergence [11,12,49]. With this aim, we recently developed an integrative approach to comparatively study GRN evolution across multiple tissues along a phylogeny [44]. However, our approach largely focused on gene co-expression and transcription factor binding site (TFBS) evolution, without assessing the contribution of other regulatory mechanisms towards GRN evolution, like posttranscriptional repression. This process generally occurs at the three prime untranslated region (3' UTR) of a gene, which can contain binding sites for both RNA-binding proteins (RBPs) and small non-coding RNAs (ncRNAs), such as microRNAs (miRNAs). miRNAs are key regulators of gene expression, and therefore fundamental to the evolution of novel phenotypes across the animal kingdom [6].

Vertebrate clades differ dramatically in species richness, and ray-finned fishes represent the largest radiation of any group (>32,000 species). Amongst this radiation, the East African cichlids are a diverse clade that arguably represents the most speciose example of adaptive radiations. In the three great lakes of East Africa (Tanganyika, Victoria and Malawi) and within the last 10 million years [19,63], one or a few ancestral lineages of cichlid fish have independently radiated into over 2000 species. These species have been able to explore a variety of ecological niches and partly as a result [63], have given rise to an explosive diversity of phenotypic traits [34]. Using genome and transcriptome sequences of five representative East African

species, we previously demonstrated that a number of molecular mechanisms may have contributed to diversification, including the rapid evolution of regulatory elements and the emergence of novel miRNAs that may alter gene expression programmes [8]. Recent studies, focused on genomic analysis of a wider range of lake species, identified low levels (0.1-0.25%) of genetic diversity between Lake Malawi species pairs [41], and link species richness in Lake Tanganyika tribes to variable heterozygosity, but not to the accelerated evolution of coding sequences [55]. Investigations of Lake Victoria species have also highlighted the role of ancient indel polymorphisms in non-coding regions towards species ecological divergence [42]. These findings largely report that the genomes are very similar within same lake species. This implies that discrete differences, like regulatory changes, are likely to have an important role in controlling gene expression and function, ultimately contributing to the large phenotypic differences among species. Indeed, our comparative approach to studying GRN evolution in six tissues of five East African cichlids was able to identify GRN changes along the phylogeny, including striking cases of network rewiring for visual genes [44]. We then experimentally confirmed that discrete TFBS mutations have disrupted regulatory edges across species, and segregate according to lake species phylogeny and ecology [44]. These findings suggested that GRN rewiring could be a key contributor to cichlid phenotypic diversity [44].

As there is evidence for the acquisition of between 36 and 1738 novel miRNAs in rapidly radiating cichlids [8,16,17,67], miRNAs have the potential to contribute to GRN evolution associated with cichlid phenotypic diversity. To that effect, signatures of purifying selection have been previously identified on cichlid miRNA binding sites

[16,30], and previous work found that on average, cichlid 3' UTRs were longer with more miRNA targets per gene than in non-cichlid teleost species [66]. Evidence for a higher evolutionary rate of 3' UTR divergence in the cichlid species has also been reported, with genes of the longest and most rapidly evolving 3' UTRs associated with translation and ribosomal pathways [66]. Additionally, conserved miRNAs differ across species in their expression levels, sequence, distribution and number of predicted binding sites [67]. Overall, these previous studies largely focused on mRNA/miRNA expression and sequence evolution at miRNA binding sites, identifying miRNAs as key gene regulators that could contribute to cichlid phenotypic diversity.

No previous study has explored the contribution of miRNAs and miRNA binding site turnover towards GRN rewiring events across cichlids. The turnover of miRNA binding sites can be defined as the rate at which an ancestrally conserved miRNAs acquire novel binding sites or lose existing ones along a phylogeny. Using our previously published genomic datasets [8] and predicted GRNs in five East African cichlids [44], here we integrate and analyse nucleotide conservation and/or variation at miRNA binding sites to better understand the selective constraints driving their evolution and infer instances of GRN rewiring associated with regulatory binding site turnover.

Results

miRNA binding site prediction in 3' UTRs of genes

To expand our previously characterised cichlid GRNs based on TFBSs [44], and to further assess the contribution of GRN rewiring towards cichlid phenotypic diversity, we used 992 cichlid miRNA mature sequences from 172 families [8] to predict miRNA binding sites in five cichlid species using Targetscan7 [1]. After filtering based on a context++ score threshold, we predicted 19,613,903 miRNA binding sites in the 3' UTRs of 21,871 orthogroups across five cichlid species (see *Materials and Methods*, Fig. 1a). We further filtered our data to only include 3' UTRs from 18,799 co-expressed orthogroups to match our previous data set [44], resulting in 15,390,993 predicted binding sites across the five species (Fig. 1a, Supplementary Fig. S1). We first compared the number of common and unique binding sites across orthologous 3' UTR sequences based on predicted miRNA family and target gene overlap (Fig. 1b). We note that there are 33,814 common sites between all species and that the three haplochromine species share the second most number (16,164) of binding sites (Fig. 1b). Unbiased by genome completeness or annotation quality (see *Supplementary information)*, between 31,186 (*P. nyererei*) and 128,831 (*A. burtoni*) unique binding sites were found to be unique to a species (Fig. 1b). In total, 3' UTR binding sites are predicted for 172 miRNA families (*M. zebra* – 118; *P. nyererei* – 117; *A. burtoni* – 151; *N. brichardi* – 115; and *O. niloticus* – 129), and variation in binding sites for these families across the five species supports differential targeting of genes in each species. For instance, *miR-15c* binding sites are under-represented in *N. brichardi* (Supplementary Fig. S2). This could be attributed to mutations of the

miR-15 seed sequence in *N. brichardi* (AGCAGCG) as compared to the other species (AGCAGCA).

Differential miRNA binding site usage highlights rewiring at the post-transcriptional level

To study miRNA binding site usage, we assess binding site conservation and divergence based on overlap of aligned 3' UTR regions (Supplementary Fig. S3). If at the same or overlapping positions in the alignment, a binding site has been predicted for more than one miRNA family between at least two species, then the ancestral binding site is predicted to be lost and functionally diverged (see *Supplementary information* and Supplementary Fig. S3). Compared to an average nucleotide identity of 95 – 99.7% across coding sequences, the average nucleotide identity across all 3' UTR alignments ranges from 83 - 95% across all pairwise species comparisons. By filtering targets based on complete positional overlap in at least two species, we retained a total of 1,626,489 3' UTR binding sites across all species (18,626/18,799 orthogroups represented). To predict functional divergence, we then assessed the number of shared sites utilised by either the same (Fig. 2a, Supplementary Fig. S4a) or different (Fig. 2b, Supplementary Fig. S4b) miRNA families between species. Consistent with the previous findings (Fig. 1b), most sites (50,212) are conserved across all species (Anc1 node, Fig. 2a). Following the phylogenetic relationships, the haplochromine species share the second highest number (Anc2 node: 32,087) of binding sites (Fig. 2a). Overall, binding sites are generally conserved and utilised by orthologous miRNA families along the whole phylogeny. Counter to this, compared to basal phylogenetic comparisons (Anc4:1 and Anc3:17 shared sites), there is more miRNA family divergence within the

haplochromine lineage (Anc2:3163 and Anc1:3200 shared sites) (Fig. 2b). For example, the developmental gene, *gata6*, has one miRNA binding site (*miR-27d*) shared between *N. brichardi* and *O. niloticus*, but in the haplochromines, has three miRNA binding sites (*miR-219, miR-128* and *miR-27*).

Comparative analysis of three-node motifs identifies increased novel network architecture between the five cichlid species

Since a GRN can be composed of both transcriptional activation and repression, we extend our previous findings [44] by instead focusing on 'three-node motifs' [3]. As previously shown for mammals [59], the study of such motifs may serve as a reliable indicator of evolutionary conserved and diverged network signatures across species. Owing to the input dataset and our aim of focusing on the impact of miRNA associated GRN rewiring in five cichlids, we focus on a topology representative of a miRNA feed-forward loop (miRNA-FFL) (Fig. 3a). In this model, the TF is predicted to regulate a target gene (TG) and a miRNA is predicted to directly regulate either the TF or TG (Fig. 3a). According to this model and to avoid any bias of gene/miRNA loss or mis-annotations in motif/binding site comparisons across all species, we filtered a starting set of 37,320,950 three-node motif edges for 1-to-1 orthologous TFs, TGs and miRNA families. This resulted in a final set of 17,987,294 three-node motif edges across the five species (see *Supplementary information* and Supplementary Fig. S6). In this set, 467,279 (3%) three-node motif edges are conserved across all five species (Supplementary Fig. S8, Supplementary Table S2). Instead, 1,321,875 (7%) – 3,124,263 (17%) three-node motifs are unique to each species (Supplementary Fig. S8, Supplementary Table S2). Using unique TF:TG (429,197) and miRNA:TG (366,302) edges in the three-node motifs across the five

species, we note that on average, 56% of miRNA:TG edges are lost compared to 46% of all TF:TG edges across five species (see *Supplementary information* and Supplementary Fig. S9).

Using the same unique edges of each TG, we identified 115,031 unique TF:miRNA relationships and assessed their frequency to identify co-regulatory conservation and divergence along the phylogeny (see *Materials and Methods*). Of these TF:miRNA relationships, 25,209 (22%) are conserved across all five species. An example of one such conserved relationship is *miR-18*, a miRNA with negatively correlated expression with mRNA pairs in Midas cichlids [17], being paired with NR2C2, a TF that we previously implicated in visual opsin GRN rewiring in cichlids [44]. On the other hand, 35,137 (31%) TF:miRNA relationships are unique to any one species and target genes associated with phenotypic diversity from previous studies (Supplementary Table S3). For example, IRF7:*miR-27b* is a unique co-regulatory relationship of *M. zebra,* and targets the fast-evolving [8] morphogenesis gene, *bmpr1* (Supplementary Fig. S10). Overall, by looking at three-node motifs, we identify evolutionary conserved signatures as well as much more novel species-specific network architecture that can be associated with traits of cichlid phenotypic diversity.

Network rewiring is associated with different models of regulatory binding site turnover in three-node motifs across species

The previous section focused on the evolution of whole network motifs. Here, we determine whether species differences in edges of these motifs are due to regulatory binding site turnover associated with previously described GRN rewiring events [44]. Using the unique TF:TG (429,197) and miRNA:TG (366,302) edges of 6,802 1-to-1

TG orthogroups, we note variation in TF or miRNA binding site gain or loss along the phylogeny (Fig. 3b). In the haplochromines, both *M. zebra* (4,195) and *P. nyererei* (4,799) have more TGs with miRNA binding site losses, whereas in *A. burtoni*, there are more TGs with either TFBS (3,937) or miRNA (3,866) gain (Fig. 3b). On the other hand, *N. brichardi* has more TGs (5,132) with TFBS and miRNA binding site loss (Fig. 3b).

Our previously described GRNs, that are based on TFBSs only [44], are therefore likely rewired due to binding site turnover in the three-node motifs. To test this, we implemented eight models of binding site evolution, including all combinations of TFBS/miRNA gain, loss, and 'no change', using *O. niloticus* as a reference. This allowed us to associate (p-val <0.05) the models of binding site evolution contributing to either 6,542 significantly rewired (degree-corrected D_n score[20] >0.17) or 260 low to non-rewired (degree-corrected D_n score[20] ≤0.17) 1-to-1 orthogroups (see *Materials and Methods*) from our previous study [44]. Since rewiring was measured according to TFBS evolution only [44], we confirmed that TFBS gain/loss, instead of miRNA binding site gain/loss, had the largest effect on significantly rewired (degree-corrected D_n score[20] >0.17) orthologs (Fig. 4a, Supplementary Fig. S11). The most associated models of rewired orthologs are TFBS loss in *A. burtoni* (p-value =0.009) and TFBS gain in *M. zebra, P. nyererei* and *N. brichardi* (p-value =0.0007-0.03) (Supplementary Table S5). However, all low to non-rewired orthologs (D_n score ≤0.17) that should be impervious to TFBS-based rewiring, are expectedly most associated with no change in TFBS, but miRNA binding site loss in all four species (p-value =0.000005-0.05) (Supplementary Table S6). This therefore indicates a discrete impact of GRN rewiring based on miRNA binding site loss.

Overall, this suggests that different models of regulatory binding site evolution have impacted GRN rewiring in the studied cichlid lineages.

Regulatory binding site turnover in three-node motifs is associated with network rewiring of adaptive trait genes

Further examination of the rewired orthologs with either of the eight models of binding site evolution identifies teleost and cichlid trait genes associated with phenotypic diversity from previous studies (Fig. 4b, Supplementary Fig. S13, Supplementary Table S3, see *Supplementary information*). Compared to all orthologs (mean D_n rewiring score = 0.17), we previously showed that four visual opsin genes (*sws1*, *rho*, *sws2a*, and *rh2b*) have considerably rewired networks (D_n score = 0.23-0.28) in species utilising the same wavelength visual palette and opsin genes [44]. The evolution of GRNs and utilisation of diverse palettes of co-expressed opsins is able to induce shifts in adaptive spectral sensitivity of adult cichlids [9] and thus, we previously demonstrated that opsin expression diversity could be the result of TF regulatory divergence in cichlids [44]. *Sws1* (ultraviolet-sensitive) opsin, utilised as part of the short-wavelength palette in *M. zebra* and *N. brichardi*, has TFBS gain and no change in miRNA binding site in *M. zebra* and *A. burtoni*, but TFBS and miRNA binding site loss in the other two species (Fig. 4b); *Sws2a* (short-wave-sensitive) opsin, utilised as part of the long-wavelength palette in *P. nyererei* and *A. burtoni*, has TFBS gain and miRNA binding site loss in *P. nyererei*, TFBS and miRNA binding site gain in *A. burtoni*, but TFBS and miRNA binding site loss in the other two species (Fig. 4b); *Rh2b* (middle-wave-sensitive) opsin, utilised as part of the short-wavelength palette in *M. zebra* and *N. brichardi*, has TFBS loss and no change in miRNA binding site in *M. zebra*, but TFBS and miRNA binding site loss in

the other three species (Fig. 4b); and *rhodopsin* (*rho*), associated with dim light vision in all species, has TFBS and miRNA binding site gain in *M. zebra* and *A. burtoni*, but TFBS and miRNA binding site loss in the other two species (Fig. 4b). These patterns of TF and miRNA regulatory divergence could therefore contribute to differential expression of adaptive trait genes (see *Supplementary information*), including visual opsins.

Discrete changes at regulatory sites are fast-evolving and associated with binding site turnover

To study the evolution of TF and miRNA regulatory divergence in the five cichlids, we assessed whether regulatory binding site turnover in three-node motifs is occurring at regions with a different rate of evolution than that expected under a neutral model. We did this by 1) determining the rate of evolution at fourfold degenerate sites and regulatory regions (3' UTR, up to 5kb gene promoter, miRNA binding sites and TFBSs); 2) identifying between species variation at regulatory sites and test for accelerated evolution; and 3) assessing corresponding regions in the context of phylogeny and ecology of radiating lake species. We started with 20,106 - 24,559 (3' UTR), 19,706 – 24,123 (up to 5kb gene promoter), 232,050 – 478,796 (miRNA binding sites), and 3,790,407 – 7,064,048 (TFBSs) unique regulatory regions across the five species, and as a putatively neutrally evolving comparison, 5,292,087 – 6,539,362 fourfold degenerate sites (Supplementary Table S9). The rate of substitutions in whole genome pairwise comparisons was calculated using phyloP [50]. In total, 86 - 98% of the nucleotides investigated had mapped conservation-acceleration (CONACC) scores (Supplementary Table S9). Across all five species pairwise comparisons, 92% of the fourfold degenerate sites are conserved, which is

consistent with an average of ~6% pairwise divergence at fourfold sites between *O. niloticus* and the other four species [8], whereas 3% are evolving at a faster rate than that expected (Supplementary Fig. S14, Supplementary Table S10). On the other hand, 81% of the regulatory regions are conserved, and 4% are exhibiting accelerated evolution (Supplementary Fig. S14, Supplementary Table S10). Since our previous study found that discrete regulatory mutations are driving GRN rewiring events [44], we hypothesised that such mutations could account for some of the accelerated regulatory sites. Using pairwise polymorphic nucleotide sites in each of the four regulatory regions (Supplementary Table S12), we identified that 81-87% (3' UTR), 69-77% (up to 5kb gene promoter), 83-99% (miRNA binding sites), and 6-8% (TFBSs) of accelerated sites are accounted for by variation in a single species (Supplementary Fig. S15, Supplementary Table S14). Notably, the proportion of these accelerated sites are significantly different (Wilcoxon rank sum test, adjusted *p-value* <0.05), especially between TF and miRNA binding sites both within, and between species (Supplementary Table S15). These results support the notion that discrete mutations in TFBSs[44] and miRNA binding sites are fast evolving i.e. fast-evolving regulatory mutations, and drive regulatory binding site turnover in three-node motifs of the five cichlids.

Discrete changes at regulatory sites are associated with regulatory binding site turnover in adaptive trait genes

Our previous study identified an abundance of adaptive trait genes with comparatively higher rewired (D_n score >0.17) networks (based on TFBSs), compared to all orthologs [44]. As a measure of regulatory binding site turnover, we therefore sought to test the frequency of association of fast-evolving regulatory

mutations in 90 adaptive trait genes (Supplementary Table S3) compared to those in corresponding regulatory regions of 90 random 'no to low rewired' genes (D_n score ≤0.17) from our previous study [44] (see *Supplementary information*). We used the 'no to low rewired' genes to ensure that this test is not biased towards genes that have rewired GRNs based on TF divergence only (see *Materials and Methods*). By comparing the proportion of fast-evolving regulatory mutations in corresponding regions of 90 adaptive trait genes and 90 random no to low rewired genes, the most notable differences (>950/1000 Wilcoxon rank sum tests, adjusted *p-value* <0.05) are found in the proportion of accelerated nucleotides in TFBSs of 90 adaptive trait gene promoter regions (Supplementary Table S16). We identified 17 adaptive trait genes with significant turnover between TF and miRNA binding sites (Supplementary Table S17-18). In *M. zebra, P. nyererei,* and *O. niloticus*, this includes genes associated with brain development and neurogenesis e.g. *neurod1*, morphogenesis e.g. *bmpr1*, and visual opsins e.g. *rho* and *sws1* (Supplementary Table S17). Furthermore, fast evolving regulatory mutations of miRNAs and TFs could be associated with the function of adaptive trait genes like, for example, ATF3 associated with neuroprotection of the retina[35] and *miR-99* implicated in retinal regulatory networks[4] are both predicted to target the visual opsin *sws1*, and MXI1 associated with neurogenesis [33] and *miR-212* associated with synaptic plasticity and function[54] is predicted to target the dim-light visual opsin, *rho* (Supplementary Table S18). Discrete mutations in regulatory binding sites of cichlid adaptive trait genes could therefore be driving GRN evolution associated with traits of cichlid phenotypic diversity.

Discrete changes at regulatory regions of adaptive traits gene segregate according to phylogeny and ecology of radiating cichlids

In our previous study, we identified that discrete TFBS mutations driving GRN evolution of visual opsin genes, also segregate according to the phylogeny and ecology of radiating lake species [44]. Here, we extend this approach to study TF and miRNA binding site variation of three-node motifs in the context of phylogeny and ecology of lake species. Using the Lake Malawi species, *M. zebra,* as a reference, we assess whether regulatory binding site turnover in three-node motifs of this species could be genotypically associated with the ecology of sequenced Lake Malawi species [41]. For this, we started with 827 nucleotide sites that 1) have identified variation between *M. zebra* and any of the other four cichlid species; 2) are located in binding sites of either TFs (709 nucleotide sites) or miRNAs (118 nucleotide sites) of *M. zebra* adaptive trait genes, that also have a significant difference (adjusted *p-value* <0.05) in the proportion of accelerated nucleotides, indicative of regulatory binding site turnover in their associated three-node motifs; and 3) are evolving at a significantly faster rate (adjusted *p-value* <0.05) than expected under a neutral model (Supplementary Table S18). We identified that 94 out of 827 accelerated nucleotide sites with between species variation across 73 Lake Malawi species, also exhibit flanking sequence conservation, representative of shared ancestral variation. Of the 94 accelerated nucleotide sites, 21 are found in miRNA binding sites, and 73 are found in TFBSs of which, 55 were not identified in our previous study [44] due to not incorporating substitution rates. Amongst the 76 accelerated nucleotide sites uniquely identified in this study, 15 (20%) include TF and miRNA binding site variation of visual opsin genes. Given the variability and importance of visual systems towards cichlid foraging habits, we therefore focus on

variation at accelerated regulatory regions of visual opsin genes. If the TF and miRNA binding sites are likely functional, we hypothesise that radiating species with similar foraging habits would share conserved regulatory genotypes, to possibly regulate and tune similar spectral sensitivities; whereas distally related species with dissimilar foraging habits would segregate at the corresponding regulatory site.

We first focus on a three-node motif of the *M. zebra* short wavelength palette visual opsin gene, *sws1*, that is predicted to be regulated by *miR-99a* and ATF3 (Fig. 5a). The homozygous variant (C|C) that predicts binding of *miR-99a* to *M. zebra sws1* 3' UTR (Fig. 5a) is 1) conserved in 60/134 (45%) Lake Malawi individuals, including the fellow algae eater, *T. tropheops,* and other distantly related species e.g. *D. kiwinge* and *N. polystigma*, that utilise the same short wavelength palette; but 2) lost in the other four species due to the A/A homozygous variant (Fig. 5a) and also homozygous segregated (A|A) in 38/134 (28%) Lake Malawi individuals, including its most closely related Mbuna species *(P. genalutea*) and *A. calliptera* (Fig. 5b and Supplementary Fig. S17). Another homozygous variant (C|C), that predicts binding of ATF3 to *M. zebra sws1* gene promoter, but is lost in *O. niloticus,* due to the T/T homozygous variant (Fig. 5a), is 1) conserved in all closely related Mbuna species and 102/116 (88%) Lake Malawi individuals, including the closely related *A. calliptera* clade; but 2) heterozygous or homozygous segregating in distantly related Lake Malawi species that utilise the same short wavelength palette, but occupy different habitats and foraging habits e.g. *D. kiwinge* – T|T and *N. polystigma* – T|C (Fig. 5b and Supplementary Fig. S16). Overall, this suggests that whilst *miR-99a* could be core regulator of *sws1* in nearly half of the studied Lake Malawi species, it is 1) unlikely to be a co-regulator of *sws1* (with ATF3) in either distantly related Lake

Malawi species utilising the short wavelength palette e.g. *D. kiwinge* and *N. polystigma*, or the *A. calliptera* clade; but 2) likely to co-regulate *sws1* (with ATF3) in most members of the rock-dwelling Mbuna clade (Fig. 5b and Supplementary Fig. S16-17). In another example, we show that a three-node motif of the dim-light vision gene, *rho,* consisting of *miR-212* and MXI1 has conserved regulatory genotypes in all studied Lake Malawi species, but has segregated and therefore not predicted in the other four cichlids (Supplementary Fig. S18-19). Phylogenetic independent contrast analysis [15] of the *sws1* (Supplementary Fig. S20-21) and *rho* (Supplementary Fig. S22-23) genotypes against visual traits and ecology of each of the 73 Lake Malawi species, highlights very little change in correlation once the phylogeny is taken into account and a regression model fitted (see *Materials and Methods*). In summary, we identified three-node motifs of visual systems that segregate according to phylogeny and ecology of lake species. Regulatory binding site turnover of three-node motifs is therefore a key contributing mechanism of GRN evolution associated with adaptive innovations in East African cichlid radiations.

Discussion

Evolutionary changes of regulatory systems and GRN rewiring events can contribute to the evolution of phenotypic diversity and rapid adaptation [38]. This is particularly the case for East African cichlid diversification that has been shaped by complex evolutionary and genomic forces. These include divergent selection acting upon regulatory regions that can alter gene expression programmes [8], rapid evolution of noncoding RNA expression [60] and ancient polymorphisms in noncoding regions [42], contrasted against a background of low between-species genetic diversity [41,42,55]. All of these findings imply that discrete differences at regulatory regions could contribute

to phenotypic differences and indeed, through discrete changes in TFBSs, we previously showed that GRN rewiring could be a key contributor to cichlid phenotypic diversity [44]. However, our previous study did not explore and integrate other genetic mechanisms, like the contributions of miRNAs towards cichlid GRN evolution. Given that miRNAs are key regulators of post-transcriptional gene expression, and that novel miRNAs have evolved in rapidly radiating cichlids [8,16,17,67], they could therefore contribute to GRN evolution associated with cichlid phenotypic diversity.

Across the five cichlid species, we identified a comparable number (218) of miRNA binding sites per 3' UTR as that previously identified (222) in Midas cichlids [16]. Across the five species, 3' UTR binding sites are differentially predicted for up to 172 miRNA families. The under-representation of certain families in a species can be attributed to mutations of the seed sequence and *arm* switching [6,8]. The largest number of conserved miRNA families are across all five species and include binding sites in 3' UTRs of genes associated with jaw development [7] and deep-water adaptation [22]. This supports an important regulatory role of miRNAs to cichlid adaptive traits [8] over a divergence time of ~ 19 MYRs [25]. We identified more miRNA family divergence within the haplochromine lineage, particularly in 3' UTRs of developmental genes; a finding that is consistent with rapid evolutionary changes of noncoding RNA expression [60] and noncoding regions [42] in corresponding Lake Tanganyika and Victoria species. Our results suggest a deeply conserved role of miRNA regulation in the five cichlids however, binding site divergence of miRNA families is likely to have an important gene regulatory role in the rapid (~ 6 MYRs [25]) phenotypic divergence of haplochromines.

Since a GRN can be composed of both transcriptional activation and repression, we extended our previous study [44] to focus on a miRNA feed-forward loop 'three-node motif'. Using this three-node motif as a measure of network divergence and evolutionary constraint, we identified increased novel/species-specific three-node motifs overall, reflected by a higher rate of miRNA edge loss (than TF edge loss) along the phylogeny. This is consistent with previous findings in Midas cichlids where miRNAs and concomitantly, their binding sites, can be rapidly lost between related groups [67]. In support, we tested the association of eight models of TFBS and/or miRNA binding site evolution, including 'no change', on TG edges previously defined as low to non-rewired [44] based on TFBSs only. We found that the most associations were expectedly with no change in TFBS, and miRNA binding site loss in all four species compared to *O. niloticus* as a reference. This indicates that miRNA binding site loss is having a discrete impact on GRN rewiring, but overall, different models of regulatory binding site evolution have impacted GRN rewiring in the cichlid lineages studied here. This included identifying that the most associated model of four highly rewired visual opsin genes (*sws1*, *rho*, *sws2a*, and *rh2b*) [44] was generally TFBS (in 50%) and miRNA binding site (in 66%) loss across the species. This supports our previous work demonstrating that opsin expression diversity could be the result of TFBS divergence in cichlids [44] and thus, regulatory divergence is likely to accommodate for heterochronic shifts in opsin expression [10,48]. Overall, these findings suggest that differential patterns of TF and miRNA regulatory divergence are likely to be associated with three-node motif and GRN rewiring of cichlid adaptive traits.

Across all five species pairwise comparisons, we find that regulatory divergence i.e. binding site turnover in three-node motifs is occurring at regions with a different rate of evolution than that expected under a neutral model. This is supported by a previous study that also identified evolutionary-accelerated 3' UTRs in the same five cichlid species and overall, suggested this as a contributory mechanism for speciation [52]. However, we extend all previous work to show that on average, nearly a third of all fast-evolving nucleotide sites in the four regulatory regions (3' UTR, up to 5kb gene promoter, miRNA binding sites and TFBSs) can be explained by pairwise polymorphisms in a single species. Whilst more than 83% of fast-evolving nucleotides in miRNA binding sites are accounted for variation in a single species, less than 8% of TFBSs are accounted for by the same type of fast-evolving variation. This supports our previous finding of discrete mutations in TFBSs driving GRN rewiring events [44], as well as elevated SNP densities in predicted miRNA binding sites, compared to flanking 3' UTR regions, of five Lake Malawi species [40]. Positive selection acting upon these regulatory regions is therefore likely to be an important evolutionary force in rapidly radiating cichlids. This is especially the case for adaptive trait genes such as the visual opsins e.g. *rho* and *sws1*, that we show to exhibit a higher proportion of fast-evolving nucleotides in their TF and miRNA binding sites, compared to subsets of random genes. Furthermore, these TFs and miRNAs are generally functionally associated with their target gene in predicted three-node motifs like, for example, the visual opsin gene, *sws1*, is predicted to be co-regulated by the TF, ATF3, that is associated with neuroprotection of the retina[35] and *miR-99* implicated in retinal regulatory networks[4]. The regulatory variants of this three-node motif (ATF3 > *sws1* < *miR-99a*) in *M. zebra* also appear to differentially segregate according to phylogeny and ecology of Lake Malawi species [41]. We find that

ATF3:*miR-99a* could be an important regulator of *sws1* in the rock-dwelling Mbuna clade, but unlikely to co-regulate *sws1* as part of the short-wavelength palette in the *A. calliptera* clade and distantly related Lake Malawi species. For another opsin gene, we identified that the possible neural co-regulation of *rho*, and therefore dim-light vision response by MXI1:*miR-212*, could be a Lake Malawi specific regulatory innovation. Overall, differential binding of miRNAs and TFs associated with retinal sensory modalities [40] and visual tuning [57] is likely to be an important genetic mechanism contributing to Lake Malawi species visual adaptations. Whilst these results significantly expand our previously characterised visual opsin GRNs [44] and provide insights into their evolution in radiating cichlids, we also provide support for the hypothesis that the evolution of cichlid visual tuning has been facilitated by regulatory mutations that are constrained by mutational dynamics [45,57] . Differential regulation of opsin genes in three-node motifs between cichlid species and their implications towards visual tuning could correspond to diversity of foraging habits, diet, habitat choice and also nuptial colouration. Fitting the Lake Malawi phylogeny had little effect on the correlations between regulatory genotypes, and visual/ecological characteristics, and therefore suggests covariance between TF/miRNA regulatory genotypes and traits. However, similar to our previous study [44], weak correlation suggests that ecotype-associated three-node motif and GRN rewiring requires additional testing. This analysis would further benefit from 1) supplementing any missing data (of wavelength palette, habitat and/or foraging habit/diet); 2) adding species data from any lowly represented clades e.g. Mbuna; and 3) experimental testing of the predicted sites.

Alongside our previous study [44], the three-node motifs and extended GRNs generated here represent a unique resource for the community; powering further molecular and evolutionary analysis of cichlid adaptive traits. For example, further examination of the three-node motifs predicted for the visual systems, that could co-regulate opsin expression diversity, could further shed light on previous preliminary studies [10,23,45,47,57]. This could involve functional validations of three-node motifs to observed trait variation by 1) high-throughput miRNA-mRNA complex and protein-DNA assays to confirm binding of thousands of sites; 2) reporter and/or cell-based assays to demonstrate transcriptional regulation; and 3) genome editing e.g. CRISPR mutations of regulatory variants to test for any observed phenotypic effect. Nonetheless, by studying the impact of miRNA regulation in three-node motifs, this work extends the first genome-wide exploration of GRN evolution in cichlids [44], and the same computational framework can be applied to study GRN evolution in other phylogenies. However, the combined framework could be extended further by 1) analysing the impact of either more, or all of the 104 three-node motif models [2] through the integration of epigenetic and co-immunoprecipitation assay data to gain regulatory directionality; and 2) including relevant datasets to study the regulatory effect of other mechanisms e.g. lncRNAs and enhancers on network topology, that could also contribute towards the evolution of cichlid phenotypic diversity [8,56]. Whilst many of the predicted three-node motifs could be false positives, the approach applied here and previously [44] ensured for rigorous filtering at each step; this included stringent statistical significance measures, and all whilst accounting for any node loss and mis-annotations in selected species (see *Materials and Methods*).

In summary, cichlids appear to utilise an array of genetic mechanisms that also contribute towards phenotypic diversity in other organisms [13,26,28,61,65,68]. However, here we provide support of TF and miRNA co-regulatory rewiring in three-node motifs of genes associated with adaptive traits in cichlids. This is further supported by large-scale genotyping studies of the predicted regulatory sites in rapidly radiating cichlid species [41]. This potential link between the evolution of three-node motifs as part of GRNs associated with cichlid adaptive traits requires further experimental verification. This is beyond that described for *cis*-regulatory sites previously [44], as well as support based on large-scale genotyping [41,42,55] and transcriptome evolution [60]; epigenetic divergence [39,62]; transgenesis assays [8,58]; population studies and CRISPR mutant assays [37]; and transcriptomic/*cis*-regulatory assays [23,45,46,57] of cichlid species.

Materials and Methods

Genomic and transcriptomic resources

Genomes and transcriptomes of the five cichlid species were obtained from NCBI and corresponding publication [8]: *P. nyererei* - PunNye1.0, NCBI BioProject: PRJNA60367; BROADPN2 annotation; *M. zebra* - MetZeb1.1, NCBI BioProject: PRJNA60369; BROADMZ2 annotation; *A. burtoni* - AstBur1.0, NCBI BioProject: PRJNA60363; BROADAB2 annotation; *N. brichardi* - NeoBri1.0, NCBI BioProject: PRJNA60365; BROADNB2 annotation; *O. niloticus* - Orenil1.1 (NCBI BioProject: PRJNA59571; BROADON2 annotation.

MicroRNA (miRNA) target prediction

We used the microRNAs that were previously sequenced from whole embryo for five cichlid species (*O. niloticus, N. brichardi*, *A. burtoni*, *P. nyererei* and *M. zebra*) [8]. The miRNA mature sequences and hairpin structures have been characterised as described previously [8] and deposited in miRBase [36]. A total of 992 (On-198, Nb-183, Ab-243, Mz-185, Pn-183) cichlid miRNA mature sequences and annotated 3' UTRs of 21,871 orthogroups (On-22411, Nb-20195, Ab-22662, Mz-21918, Pn-21599) in all five species [8] were used for target prediction. We used TargetScan7 [1] to predict species-specific genes targeted by the sequenced microRNAs (miRNAs). We used *mafft-7.271* [29] to generate gene specific multiple alignments of the annotated 3' UTRs across all five cichlid species. Target predictions were obtained by running TargetScan7 [1] using the reformatted alignments and the sequenced mature miRNA sequences as input. We selected all targets with a context++ score lower or equal to -0.1 to filter out low quality predictions; these were the binding sites used for analyses. The multiple alignments of annotated 3' UTRs and positions of predicted

sites in each species were used to identify overlapping miRNA binding sites of miRNA families between species.

Gene ontology (GO) enrichment

To assess enrichment of Gene Ontology (GO) terms in a given gene set, we use the Benjamini-Hochberg [5] False Discovery Rate (FDR) corrected hypergeometric *P*-value (*q*-value). The background (control set) for the enrichment analysis is composed of all co-expressed genes (18,799 orthogroups) from our previous study [44]. GO terms for the five cichlids were extracted from those published previously [8].

Transcription factor (TF) motif scanning

To study transcription factor (TF) – target gene (TG) associations in three-node motifs, we used predicted transcription factor binding sites (TFBSs) from our previous study [44]. Briefly, we used the aforementioned published assemblies and associated gene annotations [8] for each species to extract gene promoter regions, defined as up to 5 kb upstream of the transcription start site (TSS) of each gene. We used a combination of 1) JASPAR vertebrate motifs; 2) extrapolated cichlid-species specific (CS) Position Specific Scoring Matrices (PSSMs) [44]; and 3) aggregated generic cichlid-wide (CW) PSSMs [44] to identify TF motifs. Using FIMO [21], the gene promoter regions of each species were scanned for each TF motif using either 1) an optimal calculated *p-value* for each TF PSSM, as calculated using the *matrix quality* script from the RSAT tool suite [43]; or 2) FIMO [21] default *p-value* (1e-4) for JASPAR [31] PSSMs and PSSMs for which an optimal *p-value* could not be determined. Statistically significant TFBS motifs (FDR<0.05) were associated with their proximal target gene (TG) and represented as two nodes and one TF-TG edge. In total, there

were 3,295,212-5,900,174 predicted TF-TG edges (FDR<0.05) across the five species [44]. This was encoded into a matrix of 1,131,812 predicted TF-TG edges (FDR<0.05), where each edge is present in at least two species [44]. To enable accurate analysis of GRN rewiring and retain relevant TF-TG interactions, all collated edges were pruned to a total of 843,168 TF-TG edges (FDR<0.05) where 1) the edge is present in at least two species; 2) edges are not absent in any species due to node loss or mis-annotation; and 3) based on the presence of nodes in modules of co-expressed genes in our previous study [44].

Three-node motif generation

Three-node motifs in our study are defined as a miRNA feed-forward loop (miRNA-FFL), where a TF is predicted to regulate a TG and a miRNA is predicted to directly regulate either the TF or TG (Fig. 3a). Three-node motifs (TF:TG:miRNA) were encoded by merging all combinations of predicted TF and miRNA interactions of a TG.

Three node motif analysis

For each species three-node motifs, all TF:miRNA nodes were extrapolated for all TGs and their frequency recorded (based on the same TF orthogroup and miRNA family classification). By reverse ranking the frequency of all TF:miRNA nodes in each species, the top 100 relationships were classified to test for any significant overlap of TFs and miRNAs in species-specific three-node motifs.

A presence-absence matrix of three-node motifs in each species was generated, and the number of TFBS and miRNA binding site gains and losses, against predictions in

O. niloticus, were calculated for each species TG. The degree-corrected rewiring (D_n) score of TF-TG interactions in each orthogroup, as inferred by the DyNet-2.0 package [20] implemented in Cytoscape-3.7.1 [18], was then mapped for GRN rewiring analysis.

Hypergeometric tests for regulatory site gain and loss enrichment

The *phyper* function in R (v4.0.2) was used to test for enrichment of rewired (degree-corrected D_n score >0.17) or low to non-rewired (degree-corrected D_n score ≤0.17) genes in each of the eight models of TFBS and/or miRNA binding site gains, losses or no change. The D_n score threshold of 0.17 (for rewired vs low to non-rewired) was set based on the mean D_n score for all orthogroups and as a measure of significantly rewired genes based on our previous study [44].

Calculating substitution rate at regulatory regions

To identify loci evolving at a faster rate than that expected under a neutral model, we used phyloP [50] from the Phylogenetic Analysis with Space/Time Models (PHAST) v1.5 package [24]. Using the previously published 5-way *multiz* multiple alignment file (MAF) centred on *O. niloticus* v1.1 [8], a neutral substitution model was constructed using the previously published five cichlid phylogeny [8] in phyloFit from PHAST v1.5 [24] by fitting a time reversible substitution 'REV' model. The multiple alignment was split by chromosome/scaffold and phyloP [50] ran using the likelihood ratio test (LRT) and the 'all branches' test to predict conservation-acceleration (CONACC) scores for each site in the five species multiple alignment.

To obtain pairwise phyloP scores, we 1) created MAFs centred on each species by reordering using mafOrder from UCSC kent tools v333; 2) removed all alignments that excluded the reference species using mafFilter from UCSC kent tools v333; 3) created sorted MAFs for all pairwise species combinations using the mafFilter function in mafTools v0.1 [14]; 4) constructed a neutral substitution model for each pairwise combination using phyloFit from PHAST v1.5 [24] by fitting a time reversible substitution 'REV' model; 5) split each pairwise MAF by chromosome/scaffold; and 6) calculated substitution rates in phyloP [50] using the likelihood ratio test (LRT) and the 'all branches' test to predict conservation-acceleration (CONACC) using each corresponding pairwise neutral substitution model. To compare CONACC scores of regulatory regions to neutrally evolving regions, fourfold degenerate sites were extracted from each genome using an in-house perl script that takes a gene annotation as gene transfer format (GTF) file, whole genome FASTA file and fourfold degenerate codon table as input. The phyloP scores were then mapped to fourfold degenerate sites and the four regulatory regions (3' UTR excluding miRNA binding sites, up to 5kb gene promoter excluding TFBSs, 3' UTR miRNA binding sites and up to 5kb gene promoter TFBSs) of each species using *bedtools-2.25.0* intersect [53].

Identification of pairwise variation between the five species

Pairwise variation between all five species was identified based on an *M. zebra* v1.1 assembly centred 5-way *multiz* alignment [8]. Pairwise (single-nucleotide) variants were mapped to the phyloP scores of four regulatory regions (3' UTR, up to 5kb gene promoter, miRNA binding sites and TFBSs) using *bedtools-2.25.0* intersect [53].

Testing the significance of CONACC scores in regulatory regions of adaptive trait genes

The significance of CONACC scores of pairwise polymorphic nucleotide sites in regulatory regions of 90 adaptive trait genes was tested by: 1) Randomly picking up to 90 no to low rewired genes (D_n score ≤0.17) from our previous study [44], 1000 times; and 2) testing (*Wilcoxon* rank sum) the difference in CONACC scores of pairwise polymorphic nucleotide sites in each regulatory region of the 90 random genes to the corresponding regulatory region of all 90 adaptive trait genes. The number of times (from 1000 tests) was recorded as either having a significant (Wilcoxon rank sum test, adjusted *p-value* <0.05) or insignificant (Wilcoxon rank sum test, adjusted *p-value* >0.05) difference in CONACC scores of pairwise polymorphic nucleotide sites in each regulatory region of all five species. The adjusted *p-values* derived from Wilcoxon rank sum tests, between CONACC scores of polymorphic nucleotide sites in the regulatory region of 90 adaptive trait genes were ranked, and reverse sorted, to identify significant (adjusted *p-value* <0.05) regulatory binding site turnover.

Identification of segregating variants within binding sites

Pairwise variants of *M. zebra* were overlapped with single nucleotide polymorphisms (SNPs) in Lake Malawi species [41] using *bedtools-2.25.0* intersect [53]. The pairwise variants overlapping binding sites and lake species SNPs were then filtered based on the presence of the same pairwise variant in orthologous 3' UTR alignments. This ensured concordance of whole-genome alignment derived variants with variation in 3' UTR alignments and predicted binding sites. At each step, complementation of reference and alternative alleles was accounted for to ensure correct overlap. This

analysis was not carried out to distinguish population differentiation due to genetic structure, but to instead map 3' UTR regulatory variants onto a number of radiating cichlid species to link to phylogenetic and ecological traits.

Phylogenetic independent contrasts

Phylogenetic independent contrasts (PICs) were carried out to statistically test the effect of fitting the 73 Lake Malawi species phylogeny [41] to the covariance of segregating TFBSs and miRNA binding sites, visual (wavelength palette) and ecological traits (habitat and foraging habit/diet). This involved 1) categorically coding the genotypes of segregating regulatory sites, visual trait and ecological measurements for each of the 73 Lake Malawi species (119 individuals), and 2) using the *ape* package (v5.4.1) in R (v4.0.2) to apply the PICs test [15] on all correlations with the binding site genotype (genotype vs wavelength palette, genotype vs habitat, and genotype vs foraging habit/diet). PICs assumes a linear relationship and a process of Brownian motion between traits, and thus, for each combination of data, scatterplots were first generated. To test for any change in the correlation owing to phylogenetic signal, the regression model was compared between the relationships both excluding and including the Lake Malawi phylogeny [41].

Declarations

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We thank the BROAD institute and the Cichlid Genome Consortium for providing full access to genomic data.

Funding

TKM, LPD, WN, WH and FDP were supported and the project strategically funded by the Biotechnological and Biosciences Research Council (BBSRC), part of UK Research and Innovation, Institute Strategic Programme BB/J004669/, Core Strategic Programme Grants BB/P016774/1; BBS/E/T/000PR9817; BB/CSP17270/1; and BB/CCG1720/1 at the Earlham Institute, and acknowledge the work delivered via the Scientific Computing group, as well as support for the physical HPC infrastructure and data centre delivered via the NBI Computing infrastructure for Science (CiS) group. SR was supported by a National Science Foundation (NSF) career award (DBI: 1350677) and the McDonnell foundation at The Wisconsin Institute for Discovery.

Authors' contributions

TKM ran gene ontology (GO) enrichment, three-node motif generation, analysed three-node motifs, tested regulatory site gain and loss, calculated substitution rates, identified pairwise variants, tested the significance of CONACC scores, identified segregating binding sites, and ran phylogenetic independent contrast (PICs) tests;

LPD ran miRNA target prediction; WN ran TF motif scanning; TKM and WH wrote the manuscript with input from LPD, WN, SR and FDP.

Authors' information

Twitter handles: @TK_mehta (Tarang K. Mehta); @LucaPensoDolfin (Luca Penso-Dolfin); @nashalselection (Will Nash); @sroyyors (Sushmita Roy); @ScienceisGlobal (Federica Di-Palma); @WHaerty (Wilfried Haerty).

Corresponding authors

Correspondence to Tarang.Mehta@earlham.ac.uk

References

1. Agarwal V, Bell GGW, Nam JJ-W, Bartel DDP, Ameres S, Martinez J, Schroeder R, Anders G, Mackowiak S, Jens M, et al. Predicting effective microRNA target sites in mammalian mRNAs. eLife. 2015;4:101–112.

2. Ahnert SE, Fink TMA. Form and function in gene regulatory networks: The structure of network motifs determines fundamental properties of their dynamical state space. Journal of the Royal Society Interface. 2016;13:20160179.

3. Alon U. Network motifs: Theory and experimental approaches. Nature Reviews Genetics. 2007;8(6):450–461.

4. Andreeva K, Cooper NGF. MicroRNAs in the Neural Retina. Pasyukova E, editor. International Journal of Genomics. 2014;2014:165897.

5. Benjamini Y, Hochberg Y. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B. 1995;57(1):289–300.

6. Berezikov E. Evolution of microRNA diversity and regulation in animals. Nature Reviews Genetics. 2011;12(12):846–860.

7. Bloomquist RF, Fowler TE, Sylvester JB, Miro RJ, Streelman JT. A compendium of developmental gene expression in Lake Malawi cichlid fishes. BMC Developmental Biology. 2017;17(1):3.

8. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng a. Y, Lim ZW, Bezault E, et al. The genomic substrate for adaptive radiation in African cichlid fish. Nature. 2014;2(c):17–19.

9. Carleton K. Cichlid fish visual systems: mechanisms of spectral tuning. Integrative zoology. 2009;4(1):75–86.

10. Carleton KL, Spady TC, Streelman JT, Kidd MR, McFarland WN, Loew ER. Visual sensitivities tuned by heterochronic shifts in opsin gene expression. BMC Biology. 2008;6:22.

11. Carroll SB. Endless forms: the evolution of gene regulation and morphological diversity. Cell. 2000;101(6):577–580.

12. Carroll SB. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell. 2008;134(1):25–36.

13. Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitxl enhancer. Science (New York, N.Y.). 2010;327(5963):302–305.

14. Earl D, Nguyen N, Hickey G, Harris RS, Fitzgerald S, Beal K, Seledtsov I, Molodtsov V, Raney BJ, Clawson H, et al. Alignathon: a competitive assessment of whole-genome alignment methods. Genome research. 2014;24(12):2077–89.

15. Felsenstein J. Phylogenies and the Comparative Method. The American

Naturalist. 1985;125(1):1–15.

16. Franchini P, Xiong P, Fruciano C, Meyer A. The Role of microRNAs in the Repeated Parallel Diversification of Lineages of Midas Cichlid Fish from Nicaragua. Genome biology and evolution. 2016;8(5):1543–1555.

17. Franchini P, Xiong P, Fruciano C, Schneider RF, Woltering JM, Hulsey CD, Meyer A. MicroRNA Gene Regulation in Extremely Young and Parallel Adaptive Radiations of Crater Lake Cichlid Fish. Molecular Biology and Evolution. 2019;36(11):2498–2511.

18. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: A graph theory library for visualisation and analysis. Bioinformatics. 2015;32(2):309–311.

19. Genner MJ, Seehausen O, Lunt DH, Joyce DA, Shaw PW, Carvalho GR, Turner GF. Age of cichlids: New dates for ancient lake fish radiations. Molecular Biology and Evolution. 2007;24(5):1269–1282.

20. Goenawan IH, Bryan K, Lynn DJ. DyNet: Visualization and analysis of dynamic molecular interaction networks. Bioinformatics. 2016;32(17):2713–2715.

21. Grant CE, Bailey TL, Noble WS. FIMO: Scanning for occurrences of a given motif. Bioinformatics (Oxford, England). 2011;27(7):1017–1018.

22. Hahn C, Genner MJ, Turner GF, Joyce DA. The genomic basis of cichlid fish adaptation within the deepwater “twilight zone” of Lake Malawi. Evolution Letters. 2017;1(4):184–198.

23. Hofmann CM, O’Quin KE, Marshall NJ, Cronin TW, Seehausen O, Carleton KL, Justin Marshall N, Cronin TW, Seehausen O, Carleton KL. The eyes have it: regulatory and structural changes both underlie cichlid visual pigment diversity. PLoS biology. 2009;7(12):e1000266.

24. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Briefings in Bioinformatics. 2011;12(1):41–51.

25. Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, Arcila D, Betancur R, Li C, Becker L, et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proceedings of the National Academy of Sciences of the United States of America. 2018;115(24):6249–6254.

26. Ichihashi Y, Aguilar-Martinez JA, Farhi M, Chitwood DH, Kumar R, Millon L V., Peng J, Maloof JN, Sinha NR. Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape. Proceedings of the National Academy of Sciences. 2014;111(25):2616–2621.

27. Jacob F. Evolution and tinkering. Science (New York, N.Y.). 1977;196(4295):1161–1166.

28. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484(7392):55–61.

29. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution. 2013;30(4):772–780.

30. Kautt AF, Kratochwil CF, Nater A, Machado-Schiaffino G, Olave M, Henning F, Torres-Dowdall J, Härer A, Hulsey CD, Franchini P, et al. Contrasting signatures of genomic divergence during sympatric speciation. Nature. 2020;588(7836):106–111.

31. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Chèneby J, Kulkarni SR, Tan G, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework.

Nucleic Acids Research. 2017;46:D260–D266.

32. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science (New York, N.Y.). 1975;188(4184):107–16.

33. Klisch TJ, Souopgui J, Juergens K, Rust B, Pieler T, Henningfeld KA. Mxi1 is essential for neurogenesis in Xenopus and acts by bridging the pan-neural and proneural genes. Developmental Biology. 2006;292(2):470–485.

34. Kocher TD. Adaptive evolution and explosive speciation: the cichlid fish model. Nature reviews. Genetics. 2004;5(4):288–298.

35. Kole C, Brommer B, Nakaya N, Sengupta M, Bonet-Ponce L, Zhao T, Wang C, Li W, He Z, Tomarev S. Activating Transcription Factor 3 (ATF3) Protects Retinal Ganglion Cells and Promotes Functional Preservation After Optic Nerve Crush. Investigative Ophthalmology & Visual Science. 2020;61(2):31.

36. Kozomara A, Griffiths-Jones S. MiRBase: Annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Research. 2014;42(D1).

37. Kratochwil CF, Liang Y, Gerwin J, Woltering JM, Urban S, Henning F, Machado-Schiaffino G, Hulsey CD, Meyer A. Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations. Science. 2018;362(6413):457 LP – 460.

38. Kratochwil CF, Meyer A. Evolution□: Tinkering within Gene Regulatory Landscapes. Current Biology. 2015;25(7):R285–R288.

39. Kratochwil CF, Meyer A. Mapping active promoters by ChIP-seq profiling of H3K4me3 in cichlid fish - a first step to uncover cis-regulatory elements in ecological model teleosts. Molecular Ecology Resources. 2014 Nov 18:n/a-n/a.

40. Loh Y-HE, Yi S V, Streelman JT. Evolution of microRNAs and the diversification of species. Genome biology and evolution. 2011;3:55–65.

41. Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nature Ecology & Evolution. 2018;2(12):1940–1955.
42. McGee MD, Borstein SR, Meier JI, Marques DA, Mwaiko S, Taabu A, Kishe MA, O'Meara B, Bruggmann R, Excoffier L, et al. The ecological and genomic basis of explosive adaptive radiation. Nature. 2020;586(7827):75–79.
43. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, et al. RSAT 2015: Regulatory sequence analysis tools. Nucleic Acids Research. 2015;43(W1):W50–W56.
44. Mehta TK, Koch C, Nash W, Knaack SA, Sudhakar P, Olbei M, Bastkowski S, Penso-Dolfin L, Korcsmaros T, Haerty W, et al. Evolution of regulatory networks associated with traits under selection in cichlids. Genome Biology. 2021;22(1):25.
45. Nandamuri SP, Conte MA, Carleton KL. Multiple trans QTL and one cis-regulatory deletion are associated with the differential expression of cone opsins in African cichlids. BMC Genomics. 2018;19:945.
46. O'Quin KE, Hofmann CM, Hofmann HA, Carleton KL. Parallel Evolution of opsin gene expression in African cichlid fishes. Molecular Biology and Evolution. 2010;27(12):2839–2854.
47. O'Quin KE, Schulte JE, Patel Z, Kahn N, Naseer Z, Wang H, Conte MA, Carleton KL. Evolution of cichlid vision via trans-regulatory divergence. BMC evolutionary biology. 2012;12(1):251.
48. O'Quin KE, Smith D, Naseer Z, Schulte J, Engel SD, Loh Y-HHE, Streelman JT, Boore JL, Carleton KL. Divergence in cis-regulatory sequences surrounding the opsin gene arrays of African cichlid fishes. BMC Evolutionary Biology. 2011;11(1):120.

49. Peter IS, Davidson EH. Evolution of gene regulatory networks controlling body plan development. Cell. 2011;144(6):970–985.

50. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome research. 2010;20(1):110–21.

51. Prager EM, Wilson AC. Slow evolutionary loss of the potential for interspecific hybridization in birds: a manifestation of slow regulatory evolution. Proceedings of the National Academy of Sciences of the United States of America. 1975;72(1):200–4.

52. Puntambekar S, Newhouse R, San-Miguel J, Chauhan R, Vernaz G, Willis T, Wayland MT, Umrania Y, Miska EA, Prabakaran S. Evolutionary divergence of novel open reading frames in cichlids speciation. Scientific Reports. 2020;10(1):21570.

53. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–842.

54. Remenyi J, van den Bosch MWM, Palygin O, Mistry RB, McKenzie C, Macdonald A, Hutvagner G, Arthur JSC, Frenguelli BG, Pankratov Y. miR-132/212 Knockout Mice Reveal Roles for These miRNAs in Regulating Cortical Synaptic Transmission and Plasticity. PLOS ONE. 2013;8(4):e62509.

55. Ronco F, Matschiner M, Böhne A, Boila A, Büscher HH, El Taher A, Indermaur A, Malinsky M, Ricci V, Kahmen A, et al. Drivers and dynamics of a massive adaptive radiation in cichlid fishes. Nature. 2021;589(7840):76–81.

56. Salzburger W. Understanding explosive diversification through cichlid fish genomics. Nature Reviews Genetics. 2018;19:705–717.

57. Sandkam BA, Campello L, O’Brien C, Nandamuri SP, Gammerdinger W, Conte M, Swaroop A, Carleton KL. Tbx2a modulates switching of RH2 and LWS opsin

gene expression. Molecular Biology and Evolution. 2020;37(7):2002–2014.
58. Santos ME, Braasch I, Boileau N, Meyer BS, Sauteur L, Böhne A, Belting H-G, Affolter M, Salzburger W, Santos E, et al. The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. Nature communications. 2014;5(1):5149.
59. Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T, Stelhing-Sun S, Lee K, et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature. 2014;515(7527):365–370.
60. El Taher A, Böhne A, Boileau N, Ronco F, Indermaur A, Widmer L, Salzburger W. Gene expression dynamics during rapid organismal diversification in African cichlid fishes. Nature Ecology & Evolution. 2021;5(2):243–250.
61. Thompson DA, Roy S, Chan M, Styczynski MP, Pfiffner J, French C, Socha A, Thielke A, Napolitano S, Muller P, et al. Evolutionary principles of modular gene regulation in yeasts. eLife. 2013;2:e00603.
62. Vernaz G, Malinsky M, Svardal H, Du M, Tyers AM, Santos ME, Durbin R, Genner MJ, Turner GF, Miska EA. Mapping epigenetic divergence in the massive radiation of Lake Malawi cichlid fishes. Nature Communications. 2021;12(1):5870.
63. Wagner CE, Harmon LJ, Seehausen O. Ecological opportunity and sexual selection together predict adaptive radiation. Nature. 2012;487(7407):366–369.
64. Wilson AC, Maxson LR, Sarich VM. Two types of molecular evolution: evidence from studies of interspecific hybridization. Proceedings of the National Academy of Sciences of the United States of America. 1974;71(7):2843–2847.
65. Wittkopp PJ, Haerum BK, Clark AG. Regulatory changes underlying expression differences within and between Drosophila species. Nature Genetics. 2008;40(3):346–350.
66. Xiong P, Hulsey CD, Meyer A, Franchini P. Evolutionary divergence of 3’ UTRs

in cichlid fishes. BMC Genomics. 2018;19(1):433.

67. Xiong P, Schneider RF, Hulsey CD, Meyer A, Franchini P. Conservation and novelty in the microRNA genomic landscape of hyperdiverse cichlid fishes. Scientific Reports. 2019;9(1):13848.

68. Yanai I, Hunter CP. Comparison of diverse developmental transcriptomes reveals that coexpression of gene neighbors is not evolutionarily conserved. Genome Research. 2009;19(12):2214–2220.

Figures

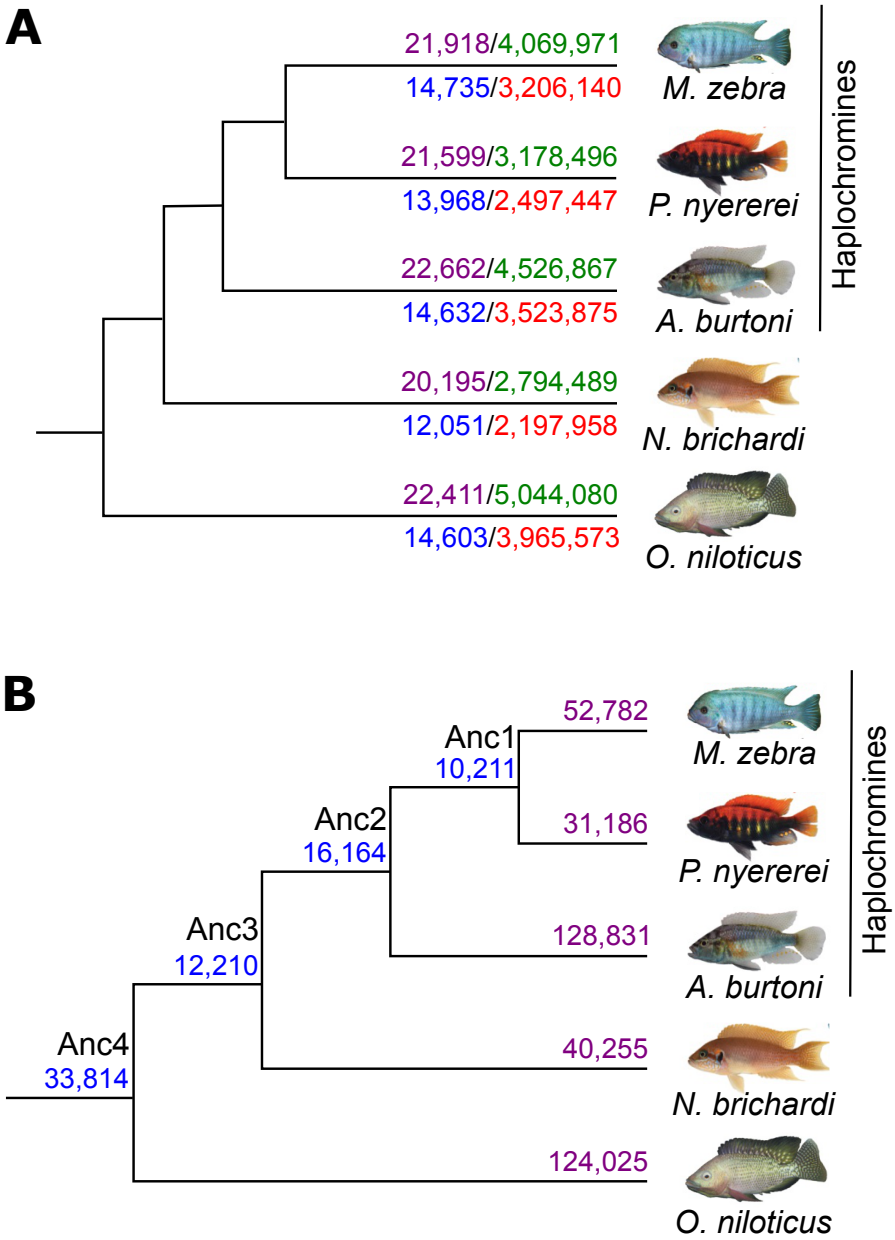

Fig. 1 – miRNA target prediction in five cichlid species.

(A) Number of miRNA target sites predicted across 3' UTR sequences in each species. Number of all input orthogroup 3' UTR sequences for each species (purple numbers) and predicted miRNA target sites from TargetScan7 after filtering for low quality predictions (green numbers) are shown for each species above the branch. Number of 3' UTR sequences across 18,799 co-expressed orthogroups for each

species (blue numbers) and predicted miRNA target sites (red numbers) are shown for each species below the branch. **(B)** Number of common and unique miRNA target sites across 3' UTR sequences of co-expressed orthogroups. Number of common miRNA target sites across 3' UTR sequences of 18,799 co-expressed orthogroups are shown at ancestral nodes (blue numbers) and unique target sites in each species (purple numbers). Common and unique sites are simply defined based on overlap of miRNA family and target gene between species.

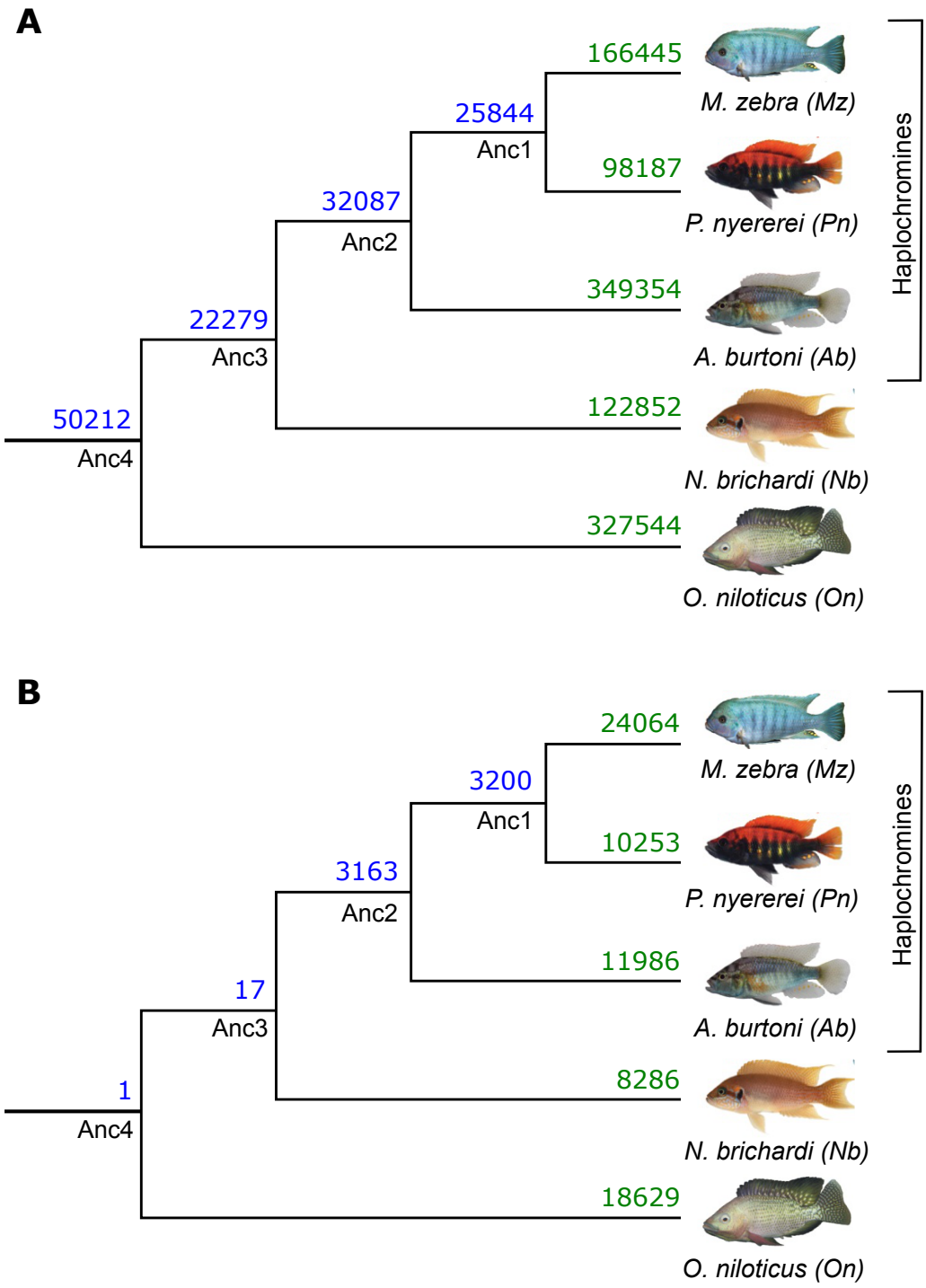

Fig. 2 – Evolution of miRNA binding sites along the five cichlid phylogeny. Number of shared and non-shared target sites based on miRNA binding site overlap in multiple 3' UTR alignments are shown at ancestral nodes (in blue) and branches (in green) for **(A)** same miRNA families and **(B)** different miRNA families.

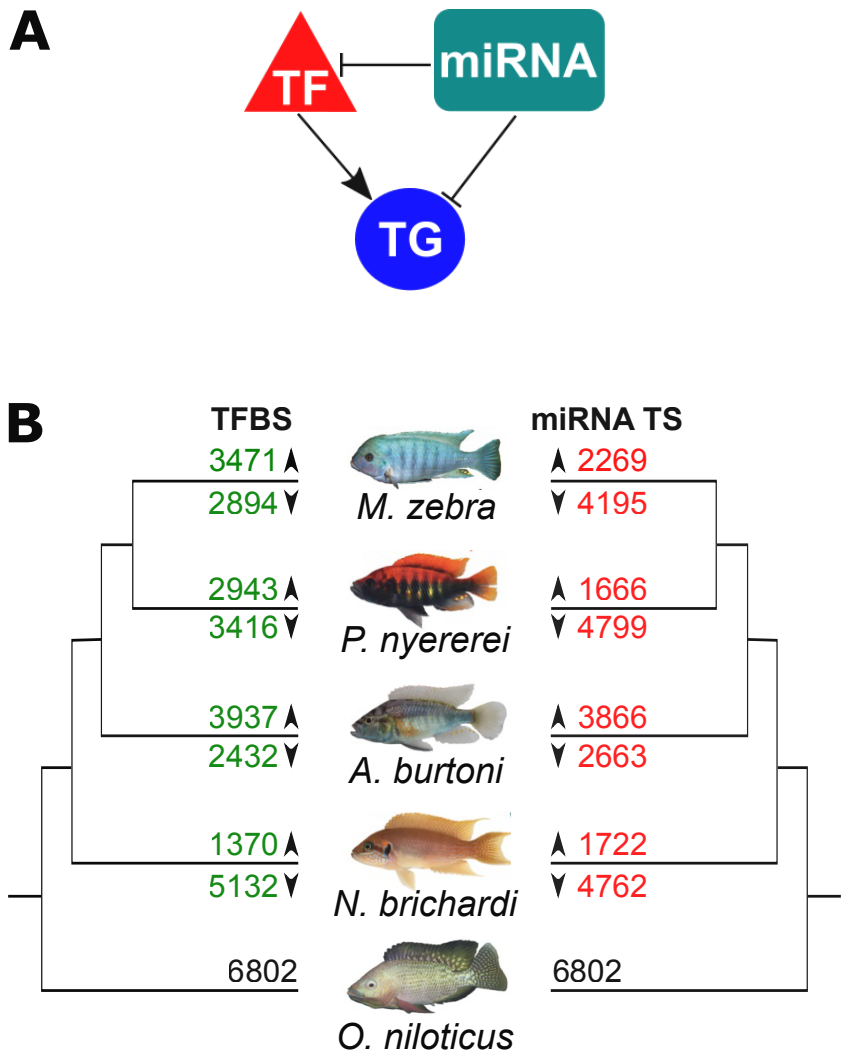

Fig. 3 – Evolution of three-node motifs (TF-TG-miRNA) in cichlids

(A) Three-node motif model used to assess network architecture. The three-node motif model used is representative of a miRNA feed-forward loop (miRNA-FFL). TF – Transcription Factor; TG – Target Gene. **(B)** TFBS and miRNA target site gain and loss in edges of 1-to-1 orthogolous target genes in three-node motifs of four cichlids. Five cichlid phylogeny showing number of 1-to-1 target gene orthologs with either TFBS (on left, in green) or miRNA (on right, in red) gain (above branch) or loss (below branch) vs *O. niloticus*. Binding sites in *O. niloticus* were used as reference for calculating gains and losses in the other species for 1-to-1 orthogroups (black numbers).

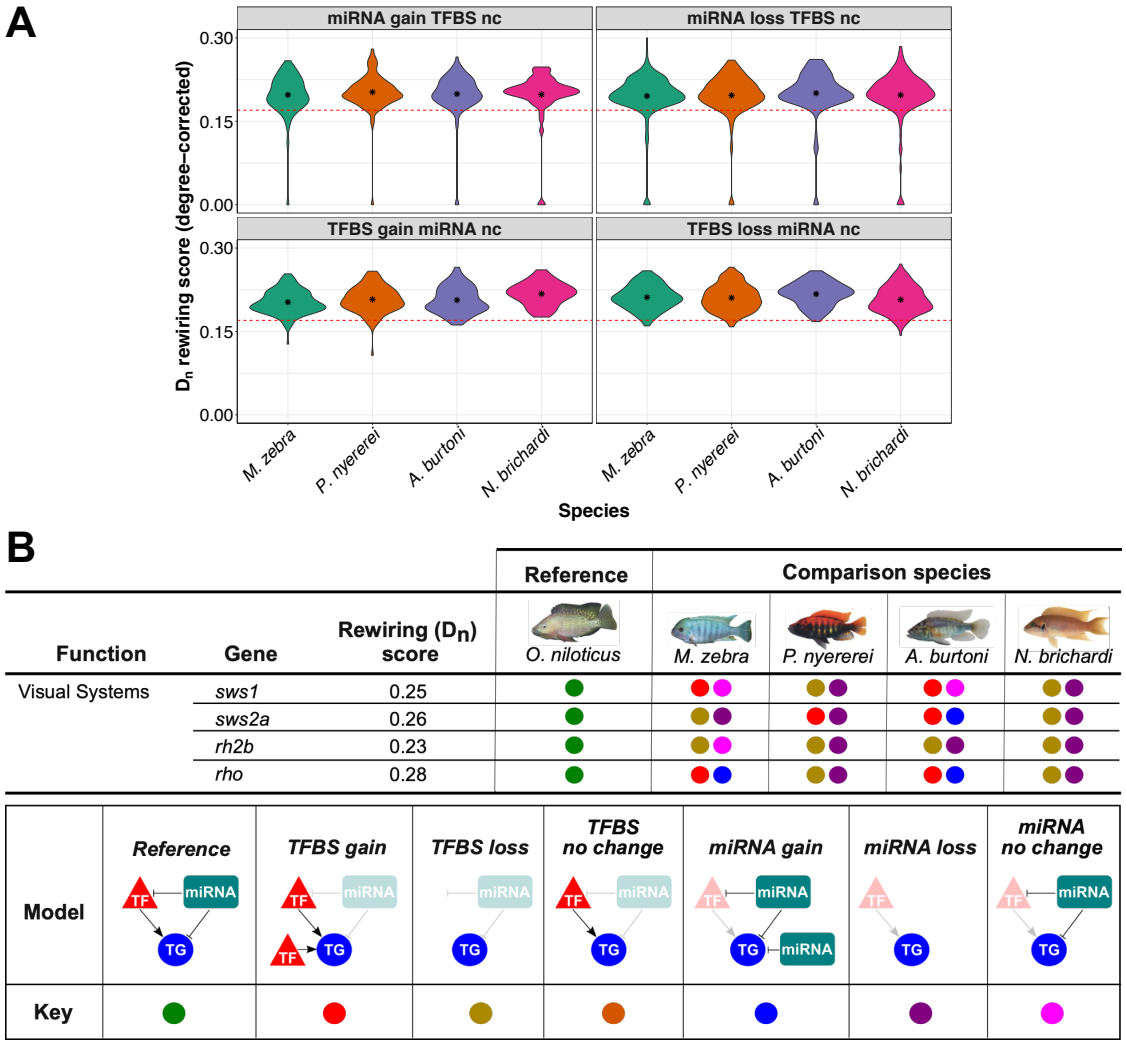

Fig. 4 – Binding site evolution in three node motifs of cichlid genes and their association with rewiring events.

(A) Different models of TFBS and miRNA binding site evolution with associated rewiring rates of 1-to-1 orthogroups in four cichlids. Violin plots of 4/8 models of binding site evolution in each species (x-axis) with DyNet rewiring score of each 1-to-1 orthogroup as degree corrected D_n score (y-axis). Red dotted line demarcates a D_n score threshold of 0.17 (for rewired vs low to non-rewired genes), which was set based on the mean D_n score for all orthogroups in our previous study [2]. The term 'nc' refers to no change and mean values are shown as internal asterisk. All statistics are included in Table S5 and violin plots of other models in Fig. S11. **(B)** Binding site evolution of four cichlid visual system genes. DyNet rewiring (D_n) score for all genes obtained from our previous study [2]. For the four comparison species, each genes model of TFBS and miRNA target site evolution in three-node motifs is calculated using the orthologous *O. niloticus* gene as a reference and demarcated as per the 'model' and 'key' in legend. All statistics are included in Supplementary Table S4-5.

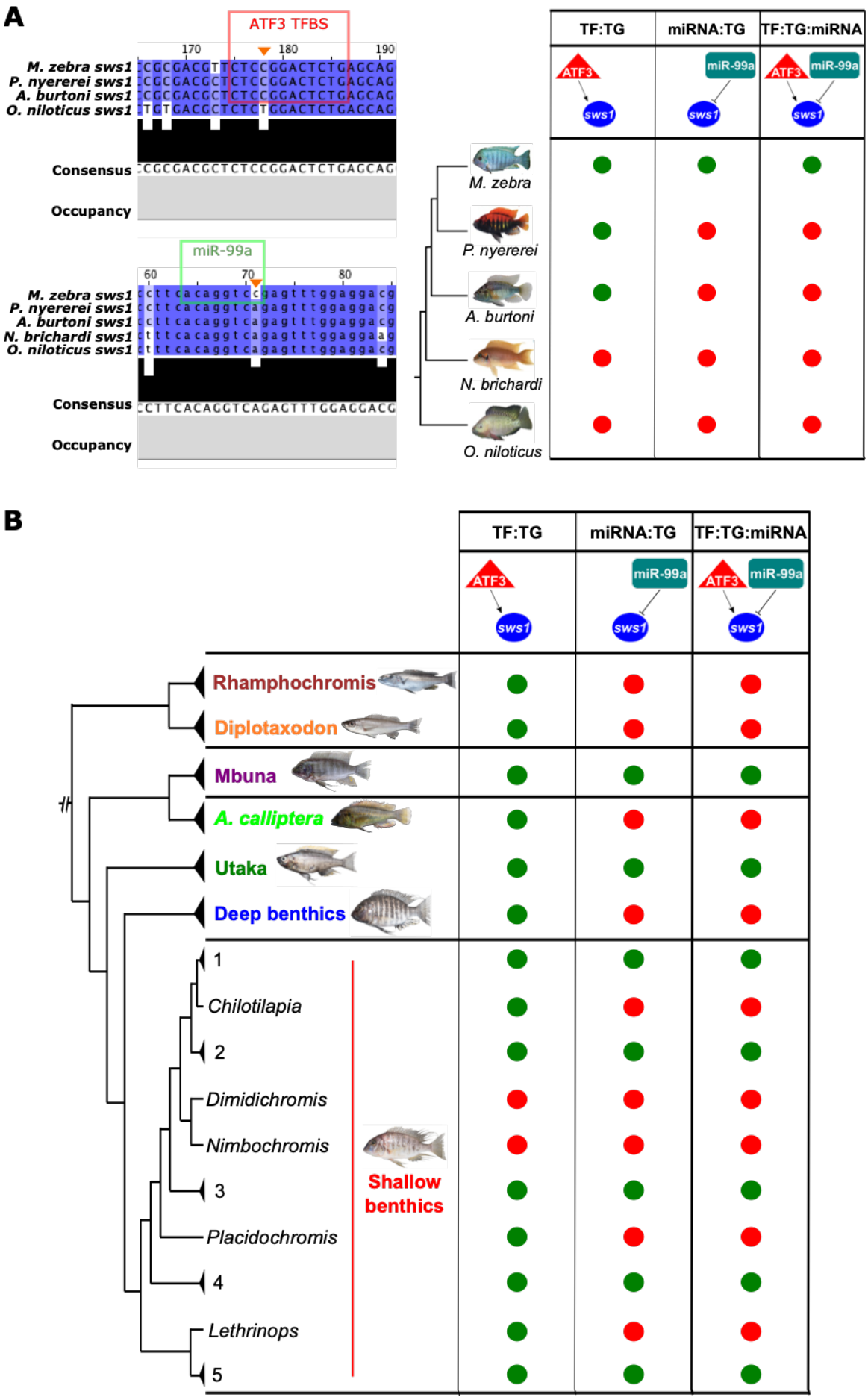
A
ATF3 TFBS
170
180
190
M. zebra sws1
P. nyererei sws1
A. burtoni sws1
O. niloticus sws1
Consensus
Occupancy
miR-99a
60
70
80
M. zebra sws1
P. nyererei sws1
A. burtoni sws1
N. brichardi sws1
O. niloticus sws1
Consensus
Occupancy
M. zebra
P. nyererei
A. burtoni
N. brichardi
O. niloticus
TF:TG
miRNA:TG
TF:TG:miRNA
ATF3
miR-99a
sws1
B
TF:TG
miRNA:TG
TF:TG:miRNA
ATF3
miR-99a
sws1
Rhamphochromis
Diplotaxodon
Mbuna
A. calliptera
Utaka
Deep benthics
1
Chilotilapia
2
Dimidichromis
Nimbochromis
3
Placidochromis
4
Lethrinops
5
Shallow benthics

Fig. 5 – Evolution of the ATF3:*sws1*:miR-99a three-node motif in the five cichlids and Lake Malawi species

(A) On the *top left,* ATF3 motif prediction in *M. zebra, P. nyererei* and *A. burtoni sws1* gene promoter (red box) and substitution demarcated in *O. niloticus sws1* gene promoter (orange arrow). On the *bottom left,* miR-99a binding site prediction in *M. zebra sws1* 3' UTR (green box) and substitution demarcated in *P. nyererei, A. burtoni, N. brichardi* and *O. niloticus sws1* 3' UTR (orange arrow). According to these predicted sites, evolution of the ATF3:*sws1*:miR-99a three-node motif in the five cichlid phylogeny is depicted based on presence (green circle) and absence (red circle). **(B)** Simplified presence (green circle) and absence (red circle) of the ATF3:*sws1*:miR-99a three-node motif in Lake Malawi species based on SNP genotypes overlapping ATF3 TFBS and miR-99a binding sites in *M. zebra sws1* gene promoter and 3' UTR (orange arrows, Fig.5a). Lake Malawi phylogeny reproduced from published ASTRAL phylogeny [1]. Phylogenetic branches labelled with genus, species or clade identifiers. Within the shallow benthics, species within some clades are summarised by numbers: 1 – *Hemitaeniochromis, Protomelas*; 2 – *Hemitilapia, Otopharnyx, Mylochromis*; 3 – *Champsochromis, Tyrannochromis, Trematocranus, Otopharnyx, Mylochromis, Stigmatochromis, Taeniochromis, Buccochromis, Ctenopharynx;* 4 – *Mylochromis;* 5 – *Taeniolethrinops.* Expanded genotype, phenotype and ecotype phylogeny in Supplementary Fig. S16-17.

References

1. Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nature Ecology & Evolution. 2018;2(12):1940–1955.

2. Mehta TK, Koch C, Nash W, Knaack SA, Sudhakar P, Olbei M, Bastkowski S, Penso-Dolfin L, Korcsmaros T, Haerty W, et al. Evolution of regulatory networks associated with traits under selection in cichlids. Genome Biology. 2021;22(1):25.