Evolution of regulatory networks controlling adaptive traits in cichlids

Tarang K. Mehta[1], Christopher Koch[2], Will Nash[1], Sara A. Knaack[3], Padhmanand Sudhakar[1], Marton Olbei[1], Sarah Bastkowski[1], Luca Penso-Dolfin[1], Tamas Korcsmaros[1], Wilfried Haerty[1], Sushmita Roy[2,3,4]* and Federica Di-Palma[1,5,6]*

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

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

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

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

[5] Norwich Medical School, University of East Anglia, Norwich, UK

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

* Corresponding authors

Abstract

Seminal studies in vertebrate protein evolution concluded that gene regulatory changes likely drive anatomical innovations. Even in the unparalleled East African cichlid fish radiation, we demonstrate cis-regulatory divergence as a contributor to phenotypic diversity. To further investigate this mechanism, we extended the Arboretum algorithm, initially designed for yeast adaptation to study the evolution of regulatory expression divergence in complex vertebrate species. For the first time, we reconstruct tissue-specific gene regulatory networks underpinning evolutionary adaptations from multiple vertebrate species in a phylogeny. Our framework consists of identifying ancestral reconstructed and extant species gene co-expression modules and integrating their associated regulators to investigate gene regulatory network evolution contributing to traits of phenotypic diversity in the East African cichlids. Along the phylogeny, we identify tissue-specific co-expression patterns across six tissues of five cichlids species that were predicted to be regulated by divergent suites of regulators. We report striking cases of rapid network rewiring for adaptive trait genes, such as the visual system. In regulatory regions of visual opsin genes e.g. sws1, in vitro assays confirm that single nucleotide polymorphisms (SNPs) in transcription factor binding sites (TFBSs) have driven network rewiring between species sharing the same visual palette. SNPs overlapping TFBSs segregate according to phylogeny and ecology suggesting ecotype-associated network rewiring in East African cichlid radiations. Our unique integrative approach inferring multi-species regulatory networks allowed us to identify regulatory changes associated with traits of phenotypic diversity in radiating cichlids.

Introduction

Early studies analyzing protein evolution in vertebrates concluded that evolutionary changes in 'regulatory systems' likely drive anatomical innovation[1–3]. Central to this regulatory system is the binding of TFs to *cis*-regulatory elements (promoters and enhancers) that flank coding genes and control the spatial and temporal expression pattern of nearby/distant genes in developing organisms. As TFs can regulate several genes (nodes) independently, these regulatory interactions (edges) form gene regulatory networks (GRNs). Changes to GRNs can often be an important source of evolutionary innovation[4]; defined as 'GRN rewiring' and characterized as regulatory edges present in one or more (rewired in) species but lost in any of the other species where the orthologous TF has another regulatory edge. Rewiring can occur through mutations within *cis*-regulatory regions, altering the TF binding capability and target gene expression; this is opposed to *trans* changes affecting the protein-coding sequence of TFs that are under stronger purifying selection as it would impair function and regulation of downstream targets[5]. Given all forms of evidence, there is no doubt that the evolution of *cis*-regulatory modifications driving gene expression evolution and GRN rewiring events plays a major role in species adaptation. In order to unravel the genetic basis of functional and physiological diversification we need to determine the relative combined contribution of coding sequences, regulatory sequence evolution, and gene expression in regulatory network evolution. Several studies have integrated some of these datasets, largely focusing on evolution of gene regulatory network inference in unicellular prokaryotes, *E. coli*[6] and several non-vertebrate eukaryotes, including yeast[7,8], plants[9], fruit-fly[10] and echinoderms[10,11]. Whilst there are efforts to collate and integrate several datasets for model vertebrates, including human and mouse[12], very little is known about the phenotypic effect of genome-wide regulatory network divergence, especially amongst non-model vertebrates[13].

In vertebrates, ray-finned fishes are the largest radiation of any group and therefore a powerful model group to focus on genome-wide patterns of regulatory network evolution. Ray-finned fishes comprise ~30,000 species[14], the vast majority of which are teleosts (~29,000) exhibiting a large diversity of body plans, behaviors and habitats. Amongst teleosts and even all vertebrates, the East African cichlid radiations represent arguably the most dramatic examples of adaptive speciation. In the great lakes of East Africa (Victoria, Malawi and Tanganyika) and within the last few million years[15,16], one or a few ancestral lineages of haplochromine cichlid fish have given rise to over 1500 species.

These species occupy a large diversity of ecological niches and differ dramatically in phenotypic traits, including skeletal morphology, dentition, color patterning, and a range of behavioral traits. Such explosive phenotypic diversification of East African cichlids is unparalleled among vertebrates making cichlid fishes an ideal model to investigate the evolution of gene regulatory networks associated with phenotypes under selection. Sequencing and analyses of representative East African cichlid genomes and transcriptomes suggested that bursts of gene duplication events combined with the rapid evolution of regulatory elements and selected protein coding genes contributed to evolutionary innovations[17]. This, coupled with low levels of nucleotide diversity between cichlid species pairs (Lake Malawi cichlids, 0.1-0.25%)[18], implies high impact of low level regulatory sequence divergence to gene expression diversity. This likely fuels GRN rewiring, serving as a substrate for the evolution of phenotypic diversity in cichlids.

To investigate the functional implications of changes in gene regulation on phenotypic and ecotypic diversity in cichlids, we developed a framework to characterize changes in gene regulation at the sequence, regulatory edge, expression and entire network level. Our framework integrates both *cis* and *trans* regulators to gain a comprehensive view of gene regulation. Our approach first defines co-expressed modules of genes using Arboretum[7], taking multi-species phylogenies into account and clustering genes together based on similar transcriptional profiles. Next, using the modules as a backbone for network organization, we integrate several datasets to reconstruct species-specific GRNs. We apply our approach to transcriptome data of six tissues (brain, eye, heart, kidney, muscle and testis) from five representative East African Cichlid species (*M. zebra, P. nyererei, A. burtoni, N. brichardi* and *O. niloticus*) to investigate tissue-specific regulatory network evolution, to: 1) analyze the impact of nucleotide variation at regulatory binding sites linked to gene expression evolution; and 2) describe and compare the regulatory influences that genes e.g. TFs exert onto one another. Our results offer unique insights into the respective interlinked roles of various evolutionary processes contributing to phenotypic innovations in representatives from the explosive East African cichlid adaptive radiation.

Results

Gene co-expression is tissue-specific and highlights functional evolutionary trajectories across cichlids

Using RNA-Seq data of five species and across 6 tissues in the Arboretum[7] algorithm, we identified ten modules of 12,051-14,735 co-expressed genes, with a per species average of 1,205-1,474 genes per module (Fig. 1a). Modules of co-expressed genes across the five species have varying expression levels ranging from strongly induced e.g. module 1 (eye), no change e.g. module 5 and weakly to strongly repressed e.g. module 3 (heart, kidney and muscle) (Fig. 1a). Five modules (0, 1, 3, 8 and 9) with a per species average of 1,001-1,353 genes, are tissue-specific, being strongly induced across one to two tissues, whereas the other five modules, with a per species average of 1,410-1,594 genes, have variable expression across the six tissues (Fig. 1a). Consistent with phylogeny and divergence times, gene orthologs of the three closely related haplochromines (*P. nyererei, M. zebra* and *A. burtoni*) and *N. brichardi* are generally more conserved in module assignment than *O. niloticus* e.g. module 2, 4 and 6 (Fig. S-R1a, blue off-diagonal elements), indicating differential gene expression evolution across the species. Between the haplochromines, gene orthologs are distributed across similar tissue-specific expression modules e.g. 0 and 8 (Fig. S-R1a, blue off-diagonal elements in haplochromines), suggesting tissue-specific gene expression is driven by differential transcriptional networks between cichlid species. The functional landscape of modules can be related to tissue-specific co-expression (Fig. S-R1b); for example, module 3 with strong brain induction (Fig. 1a) is significantly enriched for neural processes, reflecting core co-regulated networks of genes associated with signal transduction and synaptic activity (FDR <0.05, Fig. S-R1b). Modules with variable tissue co-expression e.g. module 2 (Fig. 1a) have divergent functional enrichment across species, suggesting that proteolysis and ribosomal activity (FDR <0.05, Fig. S-R1b) in kidney and heart physiological function is different among cichlid species.

State changes and variation in gene expression across tissues and species underpins divergence of regulatory networks

In our approach, Arboretum[7] module assignment is drawn from a prior distribution at the last common ancestor (LCA), allowing us to follow the evolution of gene co-expression and regulation based on ancestral assignments of the species tree[7]. Orthologous genes of each species can be assigned to non-orthologous modules (referred to as 'state changes') (Fig S-R1a), indicative of potential co-expression divergence and transcriptional rewiring from the LCA. By examining module state changes along the phylogeny, we identified 655 unique state changes along ancestral nodes (Fig. 1b). In recently rapidly radiating haplochromines (*M. zebra* and *P. nyererei*) and cichlid lineages, this highlights expression divergence of regulatory TFs and developmental genes such as *rxrb* (nuclear receptor)*, shh and ihh* (morphogenesis). Many important cellular and developmental TFs e.g. *foxo1, hoxa11* and *lbx1* have switched transcriptional programmes between the ancestral nodes (51 - Anc4/3; 20 - Anc3/2; 34 - Anc2/1). Using a measure of gene expression tissue-specificity, *tau* [54], we show that genes with no state change in module assignment (green bars) have an even, narrow to mid-intermediate breadth of expression whereas state changed genes (red bars) have a narrow to broad expression breadth (Fig. 1c), representative of orthologs clustering in non-orthologous modules (state changes). C*is*-regulation underlying some of these transcriptional programmes was analyzed based on enrichment of TFBS motifs in gene promoter regions. Regulatory TFs are enriched in module gene promoters according to tissue-specific function; for example, promoters of module 1 genes (eye-specific expression) are significantly enriched (FDR <0.05) for TF motifs involved in retina- and lens- related development/functions e.g. CRX, PITX3 and OTX1[19] (Fig. S-R1c). Variability in enrichment of retina/lens related TFs motifs e.g. RAR*α/β/γ* and RXR*α/β/γ*[20] in all species module 1 gene promoters except *N. brichardi*, highlights differential gene regulatory programmes and networks underlying traits under selection in cichlids, like the visual system[21].

Fine scale nucleotide variation at binding sites likely drives functional regulatory divergence in cichlids

Central to gene expression regulation are *cis*-regulatory elements (including promoters and enhancers) that harbor several TFBSs and mutations thereof, can drive GRN evolution, altering target gene transcription without affecting the expression pattern of

genes co-regulated by the same TF. The impact of noncoding sequence variation on gene expression was tested based on the evolutionary rate of 4622 1:1 orthologous gene promoter sequences against synonymous (fourfold degenerate) sites of protein coding regions, used as a proxy for neutral evolution. In the five cichlid genomes, there is no significant increase in evolutionary rate at promoter regions compared to fourfold-degenerate sites (Fig. S-R2aA, C, E, and F), with also no difference in promoter evolutionary rate between 1) state-changed and non-state changed genes; and 2) genes with conserved and divergent expression (Fig. S-R2b). We identify very few outlier genes with significantly higher evolutionary rate at promoter regions than corresponding fourfold sites at ancestral nodes (12-351 genes, Fig. S-R2aB) and within species (29-352 genes, Fig. S-R2aD), indicative of small-scale changes in promoter regions. Given the lack of significant evolutionary rate in the majority of gene promoter regions, we hypothesize that discrete changes that could otherwise alter *cis*-regulatory binding sites, could drive gene expression variation in the five cichlids.

Owing to the discrete nucleotide variations observed in various regulatory regions, including selected promoter regions (Fig. S-R2a), we expect that some of the variation may occur at TFBSs (Fig. S-R1c). We identified several pairwise SNPs between the five cichlids that overlap various genomic regions (Table S-R2a), including state-changed and non-state changed gene promoters and 3' UTRs (Fig. S-R2c). A large proportion of pairwise species SNPs (12 to 25 million) overlap predicted TFBSs in promoter regions, constituting 14-22% of all pairwise SNPs in the five species (Table S-R2a, Fig. S-R2c). GO enrichment analysis of cichlid pairwise SNPs overlapping gene regulatory regions highlight associations with key molecular processes e.g. signal transduction - non-state changed promoter TFBSs (Fig. S-R2d). These findings imply that discrete nucleotide variation at regulatory binding sites likely drive functional gene co-expression variation in cichlids through GRN rewiring events.

Cichlid adaptations are likely driven through gene regulatory network rewiring events facilitated by discrete changes in regulatory binding sites

To study patterns of core and divergent regulatory interactions that could be associated with cichlid phenotypic diversity, we applied an integrated framework (Fig. S-M1) to reconstruct species-specific gene regulatory networks (Table S-R3a). There are variances in network structure for all five closely-related cichlid species (Fig. S-R3a) and as such, we focus on 3,295,212-5,900,174 transcription factor - target gene (TF-TG)

predicted interactions across the five species (Table S-R3a) as 1) A large proportion of all pairwise species SNPs (14-22%) overlap TFBSs (Table S-R2a, Fig. S-R2c) and hence, disrupted binding sites will offer insights into TF-TG rewiring between species; 2) gene orthology is well characterized (as opposed to other regulators, like miRNAs); and 3) direct correlation to tissue co-expression patterns can be made (Fig. 1a). TF-TG sets of 1-to-1 orthologs allow network comparison along the whole phylogeny and are >59% conserved between species (Table S-R3b-c), likely representing core functional GRNs. In comparison, when all TF-TG edges are included, the percentage conservation is lower (<41%) between species (Table S-R3d), highlighting unique edges and likely GRN rewiring events between species. GO term enrichment of module constrained network edges both recapitulate tissue co-expression patterns e.g. module 1 (activated in eye) associated with visual perception and also exhibit variation e.g. module 1 edges in *A. burtoni* and *N. brichardi* are associated with response to stimulus (Fig. S-R3b).

To determine whether differentially co-expressed (state changed) TFs are rewired in TF-TG interactions, we first focus on 1-to-1 orthologous genes in TF-TG interactions, termed 'TF-TG 1-to-1 edges', along the five cichlid tree. These edges and focusing on TFs in particular, are highly associated with morphogenesis and cichlid traits under selection e.g. eye and brain development (Fig. S-R3cA). Many of the TFs in TF-TG 1-to-1 edges are rewired and state changed in module assignment (4060-9423/215810 edges) across the five species (Fig. 2a), indicating rewiring of transcriptional programs associated with signaling pathways, cell differentiation and embryonic development (Fig. 2b). Further examination of network rewiring rates using the DyNet[22] degree-corrected rewiring (D_n) score (Extended Data Table S-R3A) identifies several (12/31) candidate teleost and cichlid trait genes associated with morphogenesis from previous studies (Extended Data Table S-R3B), each with highly rewired GRNs when compared to the average (0.14) degree corrected D_n rewiring score (Extended Data Table S-R3C). Examples of which include *gdf10b* associated with axonal outgrowth and fast evolving in cichlids[17] and the visual opsin gene, *rh2* (Extended Data Table S-R3C). We extend our analyses to compensate for gene loss or lack of biological function, by studying GRN rewiring of 6,844 (1-to-1 orthologs) and 7,746 (non-1-to-1 orthogroups where orthologs confirmed as absent in genome, see *Methods*). This set of 843,168 'TF-TG all edges' are highly associated with morphogenesis e.g. retina development (Fig. 2b) and many rewired TFs are also state changed (in 2421-7447/843168 edges, Fig. 2d), indicating rewired transcriptional programs of associated signaling pathways, stress response and brain development (Fig. 2e). To study the extent of regulator (TF) rewiring, we then focus on the gain and loss rates of TF-TG edges along the five cichlid tree. Out of the 345

predicted regulators (TFs) in TF-TG all edges and when any given ancestor and child species branches along the five cichlid tree are compared; 133/186 (72%) TFs are predicted to have a higher rate of edge gain than loss e.g. DLX5 and NEUROD2, possibly acting as recruited regulators of gene expression in each branch from their last common ancestor (LCA) (Extended Data Table S-R3D); whereas 53/186 (28%) TFs have a higher loss of edges than gains e.g. OLIG2 and NR2C2, implying loss of gene expression regulatory activity from their LCA (Extended Data Table S-R3D). This suggests TFs and their binding sites are evolving along the five cichlid tree towards acquiring regulatory activity of genes more so than loss from their LCA. This extent of rewired TG networks is highlighted against a comparison to the average degree corrected rewiring (D_n) score (0.23) of 14536 orthologs in TF-TG all edges, where we identify several highly rewired genes (60/90) associated with cichlid phenotypic diversity (Extended Data Table S-R3E, Fig. 2f). This includes most visual opsins e.g. *rho, sws2* and *sws1,* genes associated with photoreceptor cell differentiation, *actr1b*[23], and eye development, *pax6a*[19] (Extended Data Table S-R3F).

To determine the phylogenetic and ecological association of discrete regulatory changes, we analyzed patterns of SNP divergence in candidate gene promoter TFBSs identified in the five cichlids and 73 Lake Malawi species[18] (Fig. S-R3d). Depending on flanking sequence conservation, species with the same genotype as *M. zebra* may harbor the same gene promoter TFBS and if different, likely diverged. For example, we note scenarios of 1) SNP-TFBS genotype divergence in distantly-related clades e.g. NKX2.1-*sws1*; and 2) SNP-TFBS genotype conservation in mainly same/closely-related clades e.g. EGR2-*cntn4*. This shows that genes associated with traits under selection and cichlid phenotypic diversity e.g. visual systems[21] and morphogenesis[17], harbor SNP genotypes overlapping TFBSs that segregate according to phylogeny and ecology of radiating lake species.

Regulatory networks are rewired in genes associated with traits under selection and underpinned by discrete *cis*-regulatory changes

Through our integrative approach, we examine the regulatory network topology of several genes associated with unique cichlid traits[24,25] that overlap our six studied tissues. One such trait involves visual system adaptation through the differential utilization of diversely expressed complements of seven cone opsin genes responsible for color vision[21]. The evolution of largely unexplored cichlid GRNs and diverse palettes

of co-expressed opsins can induce large shifts in adaptive spectral sensitivity of adult cichlids and thus, we hypothesize that opsin expression diversity is the result of rapid adaptive evolution in cichlids. Indeed, by focusing on species utilizing the same wavelength visual palette and opsin genes, we previously note that several visual opsin genes (*rh2b, sws1, sws2a* and *rho*) have highly rewired regulatory networks (Extended Data Table S-R3F). Across the predicted transcriptional networks of cichlid visual opsins, there are several visual-system associated TF regulators of opsin genes (*sws2a, rh2b* and *rho*) that are common e.g. STAT1A and CRX, as well as unique e.g. IRF1 and MAFA (Fig. S-R4a-c). Some TF regulators, like *satb1*, are common between two opsins of the same, short wavelength palette in *N. brichardi* (Fig. S-R4a-c). Other TFs, like GATA2, are common to dim-light vision but a duplicate, GATA2A, is a unique regulator of *rho* in a single species (Fig. S-R4c). Such patterns of TF regulatory divergence contributing to differential opsin expression are likely to correlate with peak absorption spectra of the regulated opsin and ecology of each species.

Sws1 (ultraviolet) opsin is utilized as part of the short-wavelength sensitive palette in *N. brichardi* and *M. zebra* and in both species networks, we identify common regulators associated with retinal ganglion cell patterning e.g. SATB1[26] as well as several unique regulators associated with nuclear receptor signaling e.g. RXRB and NR2C2[27] and retinal neuron synaptic activity e.g. ATRX[28] (Fig. 3a). Overall, there are substantially more predicted unique TF regulators of *sws1* in *M. zebra* (38 TFs) as compared to *N. brichardi* (6 TFs) (Fig. 3a). Such diverse regulation can increase *sws1* expression and in turn, increase spectral sensitivity to UV light and the ability for *M. zebra* to detect/feed on UV-absorbing phytoplankton and algae, as previously shown for Lake Malawi cichlids[29]. Also, tight TF-based regulation of *N. brichardi sws1* could induce rapid shifts in expression and spectral shift sensitivities between larger peak λ_{max} of 417 nm in *N. brichardi* single cones[30] compared to 368 nm of *M. zebra* Sws1[31]. Diverse regulation of *sws1* is likely based on TFBS variation as we identify a SNP that has likely broken the *M. zebra* NR2C2/RXRB shared motif that is otherwise predicted 2kb upstream of the *N. brichardi sws1* transcription start site (TSS) (Fig. 3b). Functional validation via EMSA confirms that NR2C2 and not RXRB binds to the predicted motif (Fig. 3b) in the *N. brichardi sws1* promoter, forming a complex, and the SNP has likely disrupted binding, and possibly regulation of *M. zebra sws1* (Fig. 3c-d). Based on these results and regulatory mutation effect on cichlid opsin expression[32], point mutations in TFBSs are likely driving their evolution and GRN rewiring events in traits that are under selection in radiating cichlids.

Regulatory network rewiring events are linked to phylogeny and ecology of East African cichlid radiations

To validate whether GRN rewiring, as a result of TFBS variation, can be associated with phylogeny and ecology of lake species, we looked further into examples of SNP-TFBS genotypes that are diverged between *M. zebra* and other available Lake Malawi species[18] (Fig. S-R3d). The homozygous SNP (T|T) that breaks binding of NR2C2 to *M. zebra sws1* promoter (Fig. R4a) is 1) conserved with the fellow algae eater, *T. tropheops* that also utilizes the same short-wavelength palette; 2) heterozygous segregating *(P. genalutea* - C|T and *I. sprengerae* - T|C) in closely related Mbuna species; and 3) homozygous segregated (C|C) in distantly related Mbuna species (*C. afra, C. axelrodi* and *G. mento*) and most other Lake Malawi species of which, some utilize the same short-wavelength palette and are algae eaters e.g. *H. oxyrhynchus* (Fig. 4). This suggests that in species closely related to *M. zebra* with similar diet and habitat, *sws1* may not be regulated by NR2C2, whilst other species could be, similar to *N. brichardi* (Fig. 3, Fig. 4). In another example, regulation of *rho* by GATA2*,* and not its duplicate, GATA2A*,* may be sufficient for dim-light vision response in some rock dweller species (*M. zebra* and possibly *P. genulatea, T. tropheops* and *I. sprengerae*) but both *gata2* copies could regulate *rho* in many other Lake Malawi species (79% with C|C genotype) as well as *O. niloticus* and *A. burtoni* (Fig. S-R4d). This highlights potential differential usage of a duplicate TF in dim-light vision regulation. Based on these examples of SNPs overlapping TFBSs that segregate according to phylogeny and ecology, it appears that ecotype-associated network rewiring is a key driver of adaptation in East African cichlid radiations.

Discussion

Gene regulatory network divergence can serve as a substrate for the evolution of phenotypic diversity and adaptation. In both unicellular and multicellular organisms, various mechanisms of regulatory and gene expression divergence underlying phenotypic diversity have been elucidated: horizontal gene transfer and regulatory reorganization in bacteria[33]; gene duplication in fungi[34]; *cis*-regulatory expression divergence in flies[35]; variable gene co-expression in worms[36]; dynamic rewiring of TFs in plant leaf shape[9]; *cis*-regulatory mutations in stickleback fish[37]; alternative splicing[38] and differential rate of gene expression evolution shaped by various selective pressures[39,40] in mammals. Despite these changes to genomic and regulatory architecture, some of these groups have remained virtually unvaried for millions of years of evolution, whereas certain organisms, like the near 1,500 species of East African cichlid fish, have rapidly radiated and diversified in an explosive manner. Alongside ecological opportunity[16], East African cichlid diversification has been shaped by complex evolutionary and genomic forces largely based on a canvas of low genetic diversity between species[18], gene duplications and divergent selection acting upon regulatory regions[17]; all of which imply the rapid evolution of regulatory networks underlying traits under selection. However, very little is known about the evolution of regulatory networks (genotype) and their potential phenotypic effect across ecologically-diverse cichlid species (ecotype)[41]. To study the various mechanisms of regulatory divergence towards phenotypic diversity as shown in other organisms[9,33–40,42] and cichlids[17,41], we developed a novel framework to 1) identify ancestral and extant species co-expression modules and 2) integrate associated regulators (*cis*-regulatory elements, transcription factors and miRNAs) to dissect gene expression and regulatory contribution to cichlid phenotypic diversity.

Along the phylogeny, our analyses identify gene co-expression modules with tissue-specific patterns and differential trajectories across six tissues of five cichlids; a trend previously noted in non-developmental programmes of worms[36] and shown to be under regulatory control in cichlids[17]. The differential gene co-expression trajectories across cichlids are predicted to be regulated by divergent suites of regulators, including TFs that are state-changed in co-expression module assignment. Similar to the rewiring of TFs associated with plant leaf shape[9], this suggests transcriptional rewiring events and differential gene expression evolution linked to cichlid phenotypic diversity.

Cis-regulatory elements (including promoters and enhancers) are central to cichlid gene expression regulation (Brawand et al. 2014) and harbor several TFBSs that when mutated, can drive GRN evolution. Similar to that shown in the adaptive evolution of threespine sticklebacks[37], we show that discrete nucleotide variation at binding sites likely drives functional regulatory divergence through gene regulatory network rewiring events in cichlids. When this is analyzed across species networks, we report striking cases of rapid network rewiring for genes known to be involved in traits under natural and/or sexual selection, such as the visual system, shaping cichlid adaptation to a variety of ecological niches. In regulatory regions of visual opsin genes e.g. *sws1*, *in vitro* assays confirm that SNPs in TFBSs (NR2C2) have driven network rewiring between species sharing the same visual palette. The impact of gene duplications, also implicated in fungi adaptation[34] and cichlid evolutionary divergence through differences in duplicate TF gene expression[17], is shown for the regulation of *rho* by GATA2*,* common to dim-light vision but a duplicate TF, GATA2A*,* being a unique regulator in *M. zebra* only. Furthermore, certain *M. zebra* SNPs overlapping gene promoter TFBSs e.g. *sws1* (NR2C2) and *rho* (GATA2A) segregate according to phylogeny and ecology of Lake Malawi species[18], suggesting ecotype-associated network rewiring events of traits under selection in East African cichlid radiations.

Our unique integrative approach has created a rich resource documenting gene co-expression, regulatory and network divergence as drivers of cichlid adaptive evolution and possibly speciation events. In this study we largely focus on *cis*-regulatory mechanisms of GRN rewiring however, our resource will allow for studies on the regulatory effect of other mechanisms e.g. miRNAs and gene duplications on network topology during cichlid evolution. Whilst it appears that cichlids utilize an array of regulatory mechanisms that are also shown to drive phenotypic diversity in other organisms[9,34–37,42], we conclude that discrete changes at regulatory binding sites, driving functional regulatory divergence and network rewiring events is likely to be a major source of evolutionary innovation, contributing to phenotypic diversity in radiating cichlid species. Beyond visual systems, we identify network rewiring of genes associated with several cichlid adaptive traits, including *runx2* associated with jaw morphology[43]; *ednrb1* in pigmentation and egg spots[17,44]; and *egr1* implicated in behavioral phenotypes[45]. Based on this 'catalogue' of trait-associated GRNs, genotype-phenotype-ecotype relationships contributing to cichlid phenotypic diversity can be functionally validated by 1) high-throughput *in vitro* assays - testing polymorphic TF-TG interactions; 2) *tol2* transgenesis[46,47] - testing *cis*-regulatory ability to recapitulate and promote divergence of TG expression; 3) genome editing techniques[48] - phenotyping the regulatory genotypes

linked to ecology; and 4) genotyping parents and mutants - confirm, compare and correct trait-associated genotypes in GRNs.

References

1. Wilson, A. C., Maxson, L. R. & Sarich, V. M. Two types of molecular evolution: evidence from studies of interspecific hybridization. *Proc. Natl. Acad. Sci. U. S. A.* **71**, 2843–2847 (1974).
2. Prager, E. M. & Wilson, A. C. Slow evolutionary loss of the potential for interspecific hybridization in birds: a manifestation of slow regulatory evolution. *Proc. Natl. Acad. Sci. U. S. A.* **72**, 200–4 (1975).
3. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–16 (1975).
4. Peter, I. S. & Davidson, E. H. Evolution of gene regulatory networks controlling body plan development. *Cell* **144**, 970–985 (2011).
5. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* **8**, 206–216 (2007).
6. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.* **31**, 64–68 (2002).
7. Roy, S. *et al.* Arboretum: Reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Res.* **23**, 1039–1050 (2013).
8. Koch, C. *et al.* Inference and Evolutionary Analysis of Genome-Scale Regulatory Networks in Large Phylogenies. *Cell Syst.* **4**, 543–558.e8 (2017).
9. Ichihashi, Y. *et al.* Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape. *Proc. Natl. Acad. Sci.* **111**, E2616–E2621 (2014).
10. Levine, M. & Davidson, E. Gene regulatory networks for development. *Pnas* **102**, 4936–4942 (2005).
11. Israel, J. W. *et al.* Comparative Developmental Transcriptomics Reveals Rewiring of a Highly Conserved Gene Regulatory Network during a Major Life History Switch in the Sea Urchin Genus Heliocidaris. *PLoS Biol.* **14**, (2016).
12. Yevshin, I., Sharipov, R., Valeev, T., Kel, A. & Kolpakov, F. GTRD: A database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.* **45**, D61–D67 (2017).
13. Pfennig, D. W. & Ehrenreich, I. M. Towards a gene regulatory network perspective on phenotypic plasticity, genetic accommodation and genetic assimilation. *Molecular Ecology* **23**, 4438–4440 (2014).

14. Froese, R. & Pauly, D. Fishbase. *FishBase* (2017). Available at: www.fishbase.org.
15. Genner, M. J. *et al.* Age of cichlids: New dates for ancient lake fish radiations. *Mol. Biol. Evol.* **24**, 1269–1282 (2007).
16. Wagner, C. E., Harmon, L. J. & Seehausen, O. Ecological opportunity and sexual selection together predict adaptive radiation. *Nature* **487**, 366–369 (2012).
17. Brawand, D. *et al.* The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **93**, 17–19 (2014).
18. Malinsky, M. *et al.* Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat. Ecol. Evol.* **2**, 1940–1955 (2018).
19. Bloomquist, R. F., Fowler, T. E., Sylvester, J. B., Miro, R. J. & Streelman, J. T. A compendium of developmental gene expression in Lake Malawi cichlid fishes. *BMC Dev. Biol.* **17**, (2017).
20. Browman, H. I. & Hawryshyn, C. W. Retinoic Acid Modulates Retinal Development in the Juveniles Op a Teleost Fish. *J. Exp. Biol.* **193**, 191–207 (1994).
21. Carleton, K. Cichlid fish visual systems: mechanisms of spectral tuning. *Integrative zoology* **4**, 75–86 (2009).
22. Goenawan, I. H., Bryan, K. & Lynn, D. J. DyNet: Visualization and analysis of dynamic molecular interaction networks. in *Bioinformatics* **32**, 2713–2715 (2016).
23. Whited, J. L. Dynactin is required to maintain nuclear position within postmitotic Drosophila photoreceptor neurons. *Development* **131**, 4677–4686 (2004).
24. Kocher, T. D. Adaptive evolution and explosive speciation: The cichlid fish model. *Nature Reviews Genetics* **5**, 288–298 (2004).
25. Henning, F. & Meyer, A. The Evolutionary Genomics of Cichlid Fishes: Explosive Speciation and Adaptation in the Postgenomic Era. *Annu. Rev. Genomics Hum. Genet.* **15**, 417–441 (2014).
26. Peng, Y. R. *et al.* Satb1 Regulates Contactin 5 to Pattern Dendrites of a Mammalian Retinal Ganglion Cell. *Neuron* **95**, 869–883.e6 (2017).
27. Evans, R. M. & Mangelsdorf, D. J. Nuclear receptors, RXR, and the big bang. *Cell* **157**, 255–266 (2014).
28. Medina, C. F. *et al.* Altered visual function and interneuron survival in Atrx knockout mice: Inference for the human syndrome. *Hum. Mol. Genet.* **18**, 966–977 (2009).
29. Hofmann, C. M. *et al.* The eyes have it: Regulatory and structural changes both underlie cichlid visual pigment diversity. *PLoS Biol.* **7**, (2009).
30. O'Quin, K. E., Hofmann, C. M., Hofmann, H. A. & Carleton, K. L. Parallel Evolution of opsin gene expression in African cichlid fishes. *Mol. Biol. Evol.* **27**, 2839–2854

(2010).

31. Carleton, K. L., Hárosi, F. I. & Kocher, T. D. Visual pigments of African cichlid fishes: Evidence for ultraviolet vision from microspectrophotometry and DNA sequences. *Vision Res.* **40**, 879–890 (2000).
32. OQuin, K. E. *et al.* Divergence in cis-regulatory sequences surrounding the opsin gene arrays of African cichlid fishes. *BMC Evol. Biol.* **11**, (2011).
33. McAdams, H. H., Srinivasan, B. & Arkin, A. P. The evolution of genetic regulatory systems in bacteria. *Nature Reviews Genetics* **5**, 169–178 (2004).
34. Thompson, D. A. *et al.* Evolutionary principles of modular gene regulation in yeasts. *eLife* **2013**, (2013).
35. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Regulatory changes underlying expression differences within and between Drosophila species. *Nat. Genet.* **40**, 346–350 (2008).
36. Yanai, I. & Hunter, C. P. Comparison of diverse developmental transcriptomes reveals that coexpression of gene neighbors is not evolutionarily conserved. *Genome Res.* **19**, 2214–2220 (2009).
37. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
38. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science (80-.).* **338**, 1587–1593 (2012).
39. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
40. Chen, J. *et al.* A quantitative model for characterizing the evolutionary history of mammalian gene expression. *bioRxiv* (2017). doi:10.1101/229096
41. Salzburger, W. Understanding explosive diversification through cichlid fish genomics. *Nature Reviews Genetics* (2018). doi:10.1038/s41576-018-0043-9
42. Chan, Y. F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitxl enhancer. *Science (80-.).* **327**, 302–305 (2010).
43. Fraser, G. J. *et al.* An ancient gene network is co-opted for teeth on old and new jaws. *PLoS Biol.* **7**, 0233–0247 (2009).
44. Santos, M. E. *et al.* Comparative transcriptomics of anal fin pigmentation patterns in cichlid fishes. *BMC Genomics* **17**, (2016).
45. Burmeister, S. S., Jarvis, E. D. & Fernald, R. D. Rapid behavioral and genomic responses to social opportunity. *PLoS Biol.* **3**, 1996–2004 (2005).
46. Fujimura, K. & Kocher, T. D. Tol2-mediated transgenesis in tilapia (Oreochromis niloticus). *Aquaculture* **319**, 342–346 (2011).
47. Juntti, S. A., Hu, C. K. & Fernald, R. D. Tol2-Mediated Generation of a Transgenic

Haplochromine Cichlid, Astatotilapia burtoni. *PLoS One* **8**, (2013).

48. Kratochwil, C. F. *et al.* Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations. *Science (80-.).* **362**, 457 LP-460 (2018).
49. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
50. Wu, Y.-C., Rasmussen, M. D., Bansal, M. S. & Kellis, M. TreeFix: Statistically Informed Gene Tree Error Correction Using Species Trees. *Syst. Biol.* **62**, 110–120 (2013).
51. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx1126
52. Hume, M. A., Barrera, L. A., Gisselbrecht, S. S. & Bulyk, M. L. UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **43**, D117–D122 (2015).
53. Kulakovskiy, I. V. *et al.* HOCOMOCO: A comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.* **41**, (2013).
54. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res.* **43**, W39–W49 (2015).
55. Marshall, H. *et al.* A conserved retinoic acid response element required for early expression of the homeobox gene Hoxb-1. *Nature* **370**, 567–571 (1994).
56. Aparicio, S. *et al.* Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 1684–8 (1995).
57. Jones, P. *et al.* InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
58. Medina-Rivera, A. *et al.* RSAT 2015: Regulatory sequence analysis tools. *Nucleic Acids Res.* **43**, W50–W56 (2015).
59. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
60. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
61. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).
62. Kent, W. J. BLAT - The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

63. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
64. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
65. Siahpirani, A. F. & Roy, S. A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic Acids Res.* **45**, 2221–2221 (2017).
66. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* **1695**, 1–9 (2006).
67. Hobolth, A. & Jensen, J. L. Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Stat. Appl. Genet. Mol. Biol.* **4**, (2005).
68. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. in *Bioinformatics* **25**, (2009).
69. Franz, M. *et al.* Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics* **32**, 309–311 (2015).
70. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, (2009).
71. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

Acknowledgments

We thank the BROAD institute and the Cichlid Genome Consortium for providing full access to genomic data. This work was strategically funded by the BBSRC, Core Strategic Programme Grant BB/CSP17270/1 at the Earlham Institute and National Science Foundation (NSF) career award (DBI: 1350677) and the McDonnell foundation at The Wisconsin Institute for Discovery.

Author contributions

CK, SAK and SR constructed gene trees, ran Arboretum and gene ontology (GO) enrichment; TKM, WN and PS developed and ran transcription factor (TF) motif prediction and enrichment; TKM analyzed co-expression modules, enrichment and breadth of gene expression; WN, TKM and WH calculated and analyzed evolutionary rates; SAK and SR generated co-expression edges; TKM reconstructed networks and carried out GO enrichment and analyses; SB and LPD analyzed network structure; MO, TKM and TK analyzed network rewiring; TKM and WN analyzed SNPs overlapping TFBSs; TKM carried out EMSA; TKM, WH, SR and FDP wrote the manuscript with input from SAK, WN, PS, SB and TK.

Author information

The authors declare no competing interests. Correspondence and requests for materials should be addressed to Federica.Di-Palma@earlham.ac.uk and Tarang.Mehta@earlham.ac.uk

Figure legends

Figure 1 - Evolution of gene expression in five cichlids. (A) 10 (0–9, heatmaps) co-expression modules identified by Arboretum[7] in six tissues of five cichlid species. Color bar denotes expression relative to subsets across each tissue - (red) induced; (green) repressed; and (black) no change. Each heatmap shows the expression profile of genes assigned to that module in a given species and height is proportional to number of genes in module (on *bottom*). **(B)** Number of state changes in module assignment along the five cichlid phylogeny[17]. Blue numbers (*left*, on ancestral nodes) - ancestral node genes assigned to modules; Green numbers (*middle*, on ancestral nodes) - state changes compared to the deepest common ancestor (Anc4); Red numbers (*right*, on ancestral nodes) - state changes from last common ancestor (LCA) along the phylogeny; Purple numbers (on branches) - state changes in the species compared to all other species. **(C)** Breadth of expression, calculated as Tau (demarcated in plots), shown for each of five species module genes that are switch/state changed (*left*, red violin bars) and no switch/non state changed (*right*, green violin bars). *P* values describing difference between state changed and non-state changed genes breadth of expression calculated using Mann-Whitney test.

Figure 2 - **Number of TFs switching/state changing co-expression module assignment.** This is shown for 215,810 TF-TG 1-to-1 edges: **(A)** on each branch of the five cichlid phylogeny; **(B)** GO term enrichment (FDR <0.05) of edge comparison in each species where *log10* fold enrichment shown as grid heatmap (legend on *right*); **(C)** uncorrected (*left,* red violin bars) and corrected (*right,* green violin bars) *log2* DyNet rewiring score of all gene edges (*left* plot) and candidate gene edges (*right* plot). Analyses of 843,168 TF-TG all edges: **(D)** TFs switching/state changing co-expression module assignment on each branch of the five cichlid phylogeny; **(E)** GO term enrichment (FDR <0.05) of edge comparison in each species where *log10* fold enrichment shown as grid heatmap (legend on *right*); **(F)** *log2* score (y-axis) of all gene edges (*left* plot) and candidate gene edges (*right* plot) of uncorrected (*left,* red violin bars) and corrected (*right,* green violin bars) DyNet rewiring scores.

Figure 3 - Evolution of the *sws1* opsin regulatory networks in *N. brichardi* and *M. zebra.* (A) Reconstructed regulatory networks of *sws1* opsin shown for *N. brichardi* (*left)* and *M. zebra* (*right)*: circular layout nodes are common regulators (unless missing); grid layout nodes are unique regulators. Node shape, annotation and edge color denoted in

legend. **(B)** On *left,* NR2C2 and RXRB position weight matrices (PWM) and motif prediction in *N. brichardi sws1* gene promoter (red box) and single nucleotide polymorphism (SNP) in *M. zebra sws1* gene promoter (red arrow). On *right,* NR2C2 and RXRB partial protein alignment showing DNA-binding domain (DBD) annotation in human, mouse, *M. zebra* and *N. brichardi.* **(C)** EMSA validation of NR2C2 and RXRB DBD binding to *N. brichardi* and *M. zebra sws1* gene promoter. Table denotes combinations of DNA probe and expressed DBD in EMSA reactions that include negative controls (lane 1 to 4); *N. brichardi* binding positive control (lane 5 and 6); *M. zebra* binding positive control (lane 7 and 8); kit negative (lane 9) and binding positive control (lane 10). **(D)** EMSA validation of increasing Nr2c2 DBD concentrations and binding to predicted TFBS in *N. brichardi sws1* gene promoter.

Figure 4 - SNP genotypes overlapping NR2C2 TFBS in *M. zebra sws1* promoter, other Lake Malawi species and *N. brichardi* outgroup. Lake Malawi species ordered into clades based on least controversial and all included species published ASTRAL phylogeny[18]. Phylogenetic branches labelled with species sample name and clade according to legends (*right*): A) Species foraging/diet habit (color)[29] and phased SNP genotype (shape)[18]; B) Adult opsin wavelength palette utilized[29]; and C) species habitat[29,14].

Methods

A systematic comparative framework to study the evolutionary dynamics of tissue-specific regulatory networks in cichlids

We developed a systematic comparative framework (Fig. S-M1) to infer evolutionary gene regulatory networks across five representative East African cichlid species - *O. niloticus (On), N. brichardi (Nb), A. burtoni (Ab), P. nyererei (Pn)* and *M. zebra (Mz)*. Our framework comprises: (1) identifying modules of co-expressed genes from multi-tissue/multi-species and single-tissue/multi-species data; (2) integrating several datasets (co-expression, *cis* regulatory elements, transcription factor binding site (TFBS) and miRNA profiles) to reconstruct gene regulatory networks (GRNs) refined with gene expression data to find fine-grained tissue-specific network modules; (3) examining factors driving evolutionary innovation in cichlids i.e. impact of gene duplication, *cis*-regulatory elements, novel miRNAs and nucleotide divergence within binding sites of regulatory regions, and determining their mechanistic roles towards regulatory network and module divergence; and (4) using the reconstructed networks, co-expression modules and functional landscape analysis to interpret GRN evolution of candidate genes and gene sets associated with traits under natural and/or sexual selection in cichlids.

Construction of cichlid gene trees

By considering the gene tree of 18,799 orthologous groups (orthogroups), Arboretum[7] is able to generate module assignments reflecting many-to-many relationships between orthologs resulting from gene duplication and loss. To construct gene trees with different levels of duplication, we obtained protein sequences of longest transcripts from five cichlids as well as stickleback, spotted gar and zebrafish as outgroups. Spotted gar was added as it predates the teleost-specific genome duplication event (3R) and zebrafish, as a model teleost to leverage known molecular interactions as an initial prediction of functional interactions/associations in cichlids based on orthology. We applied OrthoMCL[49] followed by TreeFix[50] to learn the reconciled gene trees. We noticed that several of the trees exhibited incomplete lineage sorting for the cichlid specific subtree but disappeared once the tree was relearned using the cichlid only species. We therefore relearned gene trees for the cichlid only species - in total, we reconstructed 17858 gene families of which, 108 had gene duplication events. A fraction of these (29 gene families) also exhibited incomplete lineage sorting. Incomplete lineage sorting was also observed

for gene groups without gene duplications, of the 17756 gene families that had no duplication, 810 exhibited incomplete lineage sorting (ILS).

Inference of multi- and single- tissue transcriptional modules in five cichlids

We ran Arboretum[7], an algorithm for identifying modules of co-expressed genes on gene expression values of six tissues (brain, eye, heart, kidney, muscle, testis) from five cichlid species, namely *O. niloticus (On), N. brichardi (Nb), A. burtoni (Ab), P. nyererei (Pn)* and *M. zebra (Mz)*[17]. Gene expression values were based on RNA sequencing of several adult individuals; the sample source, tissue isolation, RNA and library preparation, sequencing, assembly and annotation are described previously[17]. To ensure equality in *n*-fold change of expression, the gene expression values were log transformed as: log(x+1), where x is the raw expression value[17], and "log" is the natural logarithm, and then expression was normalized across each gene to have mean zero to be used as input for Arboretum[7]. Selection of the six tissues allowed us to study tissue-specific associated traits under natural and/or sexual selection in cichlids: Brain (development, behavior and social interaction); Eye (adaptive water depth/turbidity vision); Heart (blood circulation and stress response); Kidney (hematopoiesis and osmoregulation associated with water adaptation); Muscle (size, shape and movement associated with dimorphism and agility); and Testis (sexual systems associated with behavior and dimorphism).

In total, 18,799 orthogroups (see Methods: Construction of cichlid gene trees) and their associated expression data and gene tree information were inputted into Arboretum[7]. In total, this represents 59-68% of all protein-coding genes in the five cichlid genomes[17]. Certain annotated cichlid genes could not be included for a few reasons: 1) Lack of tissue expression data for all five species; 2) No mapped reads for selected tissues; 3) Lack of co-expression with other genes; and 4) Use of single development stage (adult). We selected the number of modules using a combination of strategies. First, we tried to identify the optimal number of multi-tissue modules automatically from the data by scoring a model based on the penalized log likelihood. This gave us the optimal *k* of 19, however, we also looked at lower values of k, for example, *k* = 7 and *k* = 15 for observation of co-expression clustering patterns in the next step. Secondly, we manually inspected the modules to see if increases of *k* yield patterns of expression that we have not seen before or generate recurring patterns. Finally, we devised a metric for the top three random initializations, based on a silhouette index, orthology overlap, and cross-species cluster mean dissimilarity; selecting the optimal *k* stable to the initialization.

Based on our strategy we found k = 10 modules to be optimal. Using a similar approach, this time for single tissues clustering, we found k = 5 modules to be optimal.

Handling ILS in Arboretum. The Arboretum algorithm internally tries to reconcile a tree that is not obeying the species tree by adding additional duplication and loss events. An alternate approach is to use a different species trees each representing the different ILS types and estimating the parameters of each such tree. However, there are many different cases of ILS, as identified previously[17] and the number of gene trees in each category varied significantly. However, estimating the conditional distributions for each branch in each ILS type would not be feasible as there are not enough example trees.

Functional and transcription factor binding site enrichment in modules

We use the FDR-corrected hypergeometric *P*-value (*q*-value) to assess enrichment of Gene Ontology (GO) terms and TFBSs (motifs) in a given gene set. We summarize the enrichment of terms/motifs with $q<0.05$ statistical significance and conservation in all extant and ancestral species. GO terms for the five cichlids were from those published previously[17]. To study *cis*-regulatory elements likely driving tissue-specific expression patterns, we defined gene promoter regions using up to 5 kb upstream of the transcription start site (TSS) of the gene. Motif enrichment in *cis*-regulatory regions was carried out using TFBSs obtained by the method below.

Transcription factor (TF) motif scanning

TFBSs of known vertebrate transcription factors (TFs) were obtained from the JASPAR vertebrate core motif (2018 release)[51]. Binding peak information from ChiP-seq experiments of various human and mouse TFs were retrieved from GTRD v17.04[12] and associated to protein coding genes within a vicinity of 10kb. Using core motif sequences available from JASPAR[51] or alternative databases like UniPROBE[52] and HOCOMOCO[53], uniform length motifs were identified within the TF binding peaks. In cases where the core motifs were not available, they were predicted *de novo* from the peaks themselves using MEME[54] with default settings. The aforementioned steps provided a list of transcription factor-target gene (TF-TG) interactions with the exact coordinates of the corresponding binding site(s). Cichlid sites were extrapolated based on 1) orthology; 2) minimum 70% sequence similarity[55,56]; and 3) functional domain overlap as derived using *Interpro scan 5*[57] to both source organisms (human, mouse). Cichlid extrapolated sites were used to construct cichlid-species specific (CS) Position Specific Scoring Matrices (PSSMs) for each TF using the *info-gibbs* script from the RSAT tool suite[58]. In cases where the number of extrapolated sites per species was less than three, we aggregated

the sites to construct generic cichlid-wide (CW) PSSMs. Using the PSSMs for each TF, we scanned the gene promoters and CNEs with FIMO[59] using either 1) an optimal calculated p-value for each TF PSSM; or 2) FIMO[59] default *p-value* (1e-4) for JASPAR[51] PSSMs and PSSMs for which an optimal *p-value* could not be determined. Statistically significant motifs were called using a *q-value* (False Discovery Rate, FDR) <0.05 and grouped in confidence levels and scores of: 1a) overlap of mouse and human to cichlid extrapolated - 0.3; 1b) mouse to cichlid extrapolated - 0.2; 1c) human to cichlid extrapolated - 0.15; 2a) FIMO[59] scans using extrapolated CS matrices - 0.125; 2b) FIMO[59] scans using extrapolated CW matrices - 0.110; and 2c) FIMO[59] scans using JASPAR[51] matrices - 0.115.

Calculating tissue-specificity index (tau)

As a measure for tissue specificity of gene expression, we calculated τ (Tau)[60] using log-transformed and normalized gene expression data (as inputted to run Arboretum):

$$\tau = \frac{\sum_{i=1}^{n}(1 - \hat{x_i})}{n-1};\ \hat{x_i} = \frac{x_i}{\max\limits_{1 \le i \le n}(x_i)}$$

The values of tau vary from 0 to 1; ubiquitous or broad expr ($\tau \le 0.5$); intermediate expr ($0.5 < \tau < 0.9$); and tissue-specific or narrow expr ($\tau \ge 0.9$) [54]. Amongst existing methods, τ has been shown to be the best for calculating tissue-specificity[61]. Testis normally express far more genes than any other tissue, generally displaying a tissue-specific pattern of expression. As tau was used to assess genome-wide expression levels across all tissues, but between species, testis expression data was included for each species to obtain a true representation of variation in transcriptional programs.

Variation and evolutionary rate at coding and non-coding regions

We noticed several anomalous start site annotations of genes in *M. zebra, P. nyererei, A. burtoni* and *N. brichardi* when compared to *O. niloticus.* Owing to these anomalies, we re-defined gene start sites to extract putative promoter regions. For each gene, we used the 1st exon (+/- 100bp) of the longest protein-coding sequence in *O. niloticus* to identify, via BLAT[62], corresponding orthologous start sites in the other four cichlid genomes. We filtered the output based on coherent overlap with original annotations[17] and orthogrouping in cichlid gene trees. We re-annotated gene start sites (*M. zebra* – 10654/21673; *P. nyererei* – 10030/20611; *A. burtoni* – 10050/23436; *N. brichardi* – 8464/20119) based on BLAT orthology and end sites based on original annotations[17], which was otherwise used for annotating the remaining genes. Based on new

annotations, for all 1:1 orthologs where gene expression data is available and there is no overlap of gene bodies, we extracted putative promoter regions, taken as up to 5kb upstream of the transcription start site (TSS). Using *mafft-7.271*[63], we aligned 1:1 orthologous promoter, cds and protein sequences based on orthogrouping in gene trees (see Methods: Construction of cichlid gene trees). We estimated the number of nonsynonymous substitutions per nonsynonymous site (d*N*) and synonymous substitutions per synonymous site (d*S*) in the 1:1 protein alignments using the codeml program in the PAML package[64] for each branch and ancestral node in the species tree. Otherwise, we estimated evolutionary rate for each branch and ancestral node in the species tree at promoter regions and fourfold degenerate sites, using 1:1 promoter and cds alignments in baseml and codeml programs in the PAML package[64], requiring that at least 10% of the alignment contains nucleotides and that at least 100 nucleotides are present for each species.

Reconstructing regulatory networks

To infer essential drivers of tissue-specific expression in cichlids, we constructed regulatory and functional interaction/association networks through the integration of several datasets and approaches (Fig. S-M1). All genes and their interaction/association networks were, in most cases, constrained by Arboretum module assignments to correlate to their respective patterns of tissue-specific expression and co-expression module analysis. This maintains a structured and connected network approach.

We first used species- and module- specific gene expression levels to infer an expression-based network. For this, we merged the cichlid gene expression data into a single 30 (Five species, six tissues) dimensional dataset to learn cichlid-specific TF-Target gene relationships using the Per Gene Greedy (PGG) method, a prior expression-based network inference method[65]. We projected the network into species-specific networks by taking into account edges that would not be present due to gene loss. We then integrated predicted TF-TG interactions (see Methods: Transcription Factor motif scanning) based on TFBS prediction in module gene promoter regions.

Functional landscape of reconstructed regulatory networks

We use the FDR-corrected hypergeometric *P*-value to assess enrichment of GO terms for genes in reconstructed networks. We used GO terms for the published five cichlids[17] and carried out enrichment analysis as previously done for Arboretum module genes (see *Methods*: Functional and transcription factor binding site enrichment in modules).

Transcription Factor (TF) - Target Gene (TG) network comparisons between species

To assess degree distribution of nodes in networks, we used the *igraph* package[66] in R. Network connectivity of orthologous gene networks was assessed by creating a gene-by-gene adjacency matrix for the five species TF-TG regulatory networks. To assess conservation and variation across the whole phylogeny and/or considering any species/clade -specific novelty, three sets of nodes (genes) were considered: a) node edges of 1-to-1 orthologs (presence in all five species networks); b) node edges of 1-to-1 orthologs in species pairwise comparisons, and; c) a superset of all node edges present in networks of any of the five species. For each gene-specific network investigated, we sum up the adjacency matrices for the intersection and superset of genes. Nodes are ranked according to a score where the sum of all columns is divided by the number of occurrences (entries which are non-zero), ultimately giving a measure of conservation.

TF-TG gain and loss rates

Gain and loss rate analyses was similar to that performed previously[8]. This approach uses a continuous-time Markov process parameterized by TF-TG edge gain and loss rates, and uses an expectation-maximization (EM) based algorithm to estimate the rates[67,68]. The input network comprised target genes of 798 individual regulator genes mapped across the five cichlids species based on the gene orthology. Each species regulator required a minimum of 25 edges as having <25 edges greatly hinders statistical analysis in this context. This resulted in a total of 345 regulators with 25 to 23,935 edges, with an average of 2,609. Gain and loss rate was estimated for each regulator using the EM-based algorithm on the edge gain and loss pattern across the phylogeny of the five cichlids species. Rates were inferred using previously published five cichlid branch lengths[17] that described neutral sequence evolution across the species. Stability analysis of rate estimations were carried out by: 1) Gain and loss rate input values were scanned from 0-400 in intervals of 5 for each regulator matrix; and 2) From each scan, rates with the greatest likelihood were chosen as the recommended gain and loss rate for that regulator, defining a final set of inferred rates for the 345 regulators.

Regulatory rewiring analysis of gene sets

The DyNet package[22] as implemented in Cytoscape v3.2.1[69] was used for network visualization and calculation of dynamic rewiring scores of Transcription Factor (TF) – Target Gene (TG) interactions as predicted by Transcription Factor motif scanning and TF-TG co-expression derived relationships (PGG method[65]). To ensure rewiring of TFs are correctly compared between species, and not based on gene loss/poor annotation,

we only included edges for analysis where the TF had a 1-to-1 orthologous relationship in species where the TF-TG interaction/relationship exists. Also, we filtered out any TGs and their TF interaction/relationships if, based on orthologous gene *tblastx*[70], whether the gene was present in the genome but not annotated. Of the 18,799 orthogroups used for generating modules of co-expressed genes and network interactions, 4,209 orthogroups had non-1-to-1 genes actually present in the genome of at least one of the five species. These 4,209 orthogroups were filtered out, retaining 907,418/1,131,812 predicted TF-TG interactions/relationships across the five species. These edges, were used for network rewiring analysis.

Identification of segregating sites in TFBSs

Species pairwise SNPs were identified based on an *M. zebra* v1.1 assembly centered 8-way teleost multiz alignment[17]. Pairwise SNPs were then overlapped with TFBS positions as determined by TF motif scanning using *bedtools* intersect[71]. Pairwise SNPs of *M.zebra* were overlapped with SNPs in Lake Malawi species[18] using *bedtools* intersect[71]. Both sets of pairwise SNPs overlapping motifs and lake species SNPs were then filtered based on the presence of the same pairwise SNP in orthologous promoter alignments. This ensured concordance of whole-genome alignment derived SNPs with SNPs in orthologous promoter alignments and predicted motifs. At each step, reference and alternative allele complementation was accounted for to ensure correct overlap.

Expression of protein DNA-binding domains (DBDs)

DNA-binding domains (DBDs) of selected cichlid proteins (NR2C2 and RXRB) were predicted based on alignment and conservation to annotated human and mouse orthologs. RNA was extracted from brain, liver and testis tissues of adult *M. zebra* and *N. brichardi* using the RNeasy Plus Mini Kit (Qiagen), achieving RNA integrity (RIN) in the range of 8-10 (Agilent Bioanalyzer Total RNA Pico Assay). First strand cDNA synthesis of DBD-specific regions was carried out using RevertAid H Minus Reverse Transcriptase (Thermo Scientific) and DBDs amplified (2-step RT-PCR) using Platinum Taq DNA Polymerase (Invitrogen) and the primers listed in Table S-M1a. Resulting cDNA was concentrated using Minelute PCR purification (Qiagen) and 700 ng used for *in vitro* transcription/translation using TnT T7 Quick for PCR DNA (Promega) and the Fluorotect GreenLys tRNA (Promega) labelling system. DBD expression was resolved by SDS-PAGE and detection using the fluorescein filter in the ChemiDoc Touch (Bio-Rad) system.

Electrophoretic mobility shift assay (EMSA) validation of predicted TF-TG interactions

EMSA was carried out using double-stranded Cy5 fluorophore 5'-modified (IDT) DNA probes, *in vitro* expressed DBDs (see above) and the Gel Shift Assay Core System (Promega). Double-stranded DNA probes were generated by annealing sense and antisense oligonucleotides (see Table S-M1a) in annealing buffer (10 mM Tris pH 7.5, 1 mM EDTA, 50 mM NaCl) for 3 mins at 96°C, 1 min at 90°C, 1 min at 85°C, 3 mins at 72°C, 1 min at 65°C, 1 min at 57°C, 1 min at 50°C, 3 mins at 42°C and 3 mins at 25°C in a PCR thermocycler. Binding reactions were carried out in a final volume of 9 µl composed of Gel Shift Binding 5x Buffer (20% glycerol, 5mM $MgCl_2$, 2.5mM EDTA, 2.5mM DTT, 250mM NaCl, 50mM Tris-HCl (pH 7.5), 0.25mg/ml poly(dI-dC)•poly(dI-dC)); 0.01µM of Cy5-dsDNA probe covering the motif and flanking region (28nt); and either 23ng (RXRB, 10.42kDa) or 27ng (NR2C2, 10.73kDa) of expressed DBD. For EMSA validation with increasing Nr2c2 DBD concentrations, 1× = 27ng. For kit controls, 0.01µM of human SP1 DNA probe was combined with 10,000ng HeLa nuclear extract. Binding reactions were incubated at room temperature for 20 mins. Protein-DNA complexes were resolved on 1mm NuPAGE 4-12% Bis-Tris polyacrylamide gels (Invitrogen) in 0.5× TBE at 100V for 60 mins. Protein-DNA complexes were detected using the Cy5 filter on the ChemiDoc MP (Bio-Rad) system. Exposure settings were adjusted in Image Lab v6.0.1_build34 (Bio-Rad) with same high (5608), low (1152) and gamma (1.0) values set for all associated images.