The genomic basis of adaptation to the deep water 'twilight zone' in Lake Malawi cichlid fishes

Christoph Hahn[1,2]*, Martin J Genner[3], George F Turner[4], Domino A Joyce[1].

1 Evolutionary and Environmental Genomics Group (@EvoHull), School of Environmental Sciences, University of Hull, Hull HU5 7RX, UK.
2 Institute of Zoology, University of Graz, A-8010 Graz, Austria.
3 School of Biological Sciences, University of Bristol, Bristol Life Sciences Building, 24 Tyndall Avenue, Bristol. BS8 1TQ. UK.
4 School of Biological Sciences, Bangor University, Bangor, Gwynedd LL57 2UW, Wales, UK.

** corresponding author*

christoph.hahn@uni-graz.at

Abstract

Deep water environments are characterized by low levels of available light at increasingly narrow spectra, great hydrostatic pressure and reduced dissolved oxygen - conditions predicted to exert highly specific selection pressures. In Lake Malawi over 800 cichlid species have evolved, and this adaptive radiation extends into the "twilight zone" below 100 metres. We use population-level RAD-seq data to investigate whether four endemic deep water species (*Diplotaxodon* spp.) have experienced divergent selection within this environment. We identify candidate genes including regulators of photoreceptor function, photopigments, lens morphology and haemoglobin, many not previously implicated in cichlid adaptive radiations. Co-localization of functionally linked genes suggests co-adapted "supergene" complexes. Comparisons of *Diplotaxodon* to the broader Lake Malawi radiation using genome resequencing data revealed functional substitutions in candidate genes. Our data provide unique insights into genomic adaptation to life at depth, and suggest genome-level specialisation for deep water habitat as an important process in cichlid radiation.

Key words: Root effect, supergene, cichlid, hemoglobin, haemoglobin, sensory drive

Deep water environments pose an array of physiological challenges to the organisms that inhabit them. Below 50m, increased hydrostatic pressure, reduced levels of dissolved oxygen and a lack of ambient light, will all produce characteristic selection pressures [1,2]. Evolutionary adaptation of species to this range of ecological challenges should be possible to detect at the genomic level, and yet surprisingly few studies have addressed this. In the context of ecological speciation and adaptive radiation, divergence along depth gradients is associated with the evolution of reproductive isolation in many marine [3–6] and freshwater species groups. Specifically, fish species within adaptive radiations of the freshwater Lakes Baikal, Tanganyika, Malawi differ extensively in the depth ranges that they occupy. Thus, investigating the genomic regions involved should provide powerful insights into the rapid ecological adaptation in these species, including physiological characteristics that have been subject to divergent selection.

The increase in hydrostatic pressure associated with depth affects multiple biological processes, from the activity of macromolecular protein assemblages such as tubulin and actin, to cellular processes such as osmoregulation and actin potential transmission in nervous cells [7,8]. As a consequence, pressure can affect the nervous system, cardiac function and membrane transport systems relatively quickly, whereas other systems are more resilient to change [9]. Coupled with the increase in pressure is a decline in the spectral range and intensity of ambient light [10]. In aquatic systems light intensity decreases exponentially with water depth [11] and in clear water the long wavelength portion of the visible light spectrum (red light) is increasingly attenuated, shifting the spectral median towards short wavelength, monochromatic (blue) 'twilight' conditions [12] in deep water environments. In marine mesopelagic fishes perhaps the most recognised morphological adaptation of the visual system for life at increased depth is eye enlargement, which accommodates the reduced light intensity by increasing the chance of photon capture [13,14]. In addition, shifts in the relative abundance of rods and cones in the retina have been associated with habitat depth [15,16] ; in vertebrates, cones mediate photopic vision under bright light conditions, while rods contain specialized pigments for scotopic vision under dim-light conditions [17].

Comparative analyses of the light absorption spectra of photopigments in different species have firmly established the role of "spectral tuning" in sensory adaptations, i.e. shifts in the maximum spectral sensitivity of photopigments towards the peak wavelength of the available light [18–21]. Differential expression and alterations in amino acid sequences of opsin genes have been identified as the underlying molecular mechanisms for spectral shifts [19,21–23]. In the context of deep water adaptations, a range of genomic modifications affecting rhodopsin genes (*Rh1* and *Rh2*) which code for the rod pigments, have been implied in spectral tuning in marine mesopelagic [6,18,19], and freshwater fishes, including the Lake Baikal sculpins (genus *Cottus*) [24] and deep-water cichlids of the East African great lakes [25]. In this study we consider the adaptations of Lake Malawi's twilight zone (50-200m) cichlids in more detail. This region is inhabited by members of an endemic deep-water haplochromine cichlid

lineage, that includes approximately 20 species of *Diplotaxodon*, plus the closely related *Pallidochromis tokolosh,* all of which are zooplanktivorous or piscivorous [26]. Sympatric species in the lineage often differ in male monochromatic nuptial colour and morphological traits including eye size [27] and this divergence must have happened since Lake Malawi achieved deep water conditions, in the last 5 million years or even more recently given evidence that the lake has been dry or shallow for much of its history [28,29]. Currently, little empirical data on species specific depth distributions of *Diplotaxodon* are available, but there are indications that some species may breed in different depth zones [30]. Additionally, survey catch records indicate that members of the *D. macrops* complex are found at depths of 150-220m during the day, while members of the *D. limnothrissa* complex occupy all oxygenated depths (ie. down to ~250m), with a peak of abundance at ~60m [31].

In this study we use population-level genome-wide SNP data to infer ecologically relevant physiological differences previously undetected between these species. We use genome scans and three independent candidate outlier approaches to test the prediction that regions of the genome involved in adaptation to depth (regions involved in adaptation to visual signal difference and hydrostatic pressure change) will be more divergent than average genomic divergence between the species. We characterize genomic variants underlying the observed interspecific eye morphological differences, and identify likely candidate regions for adaptation to life at depth.

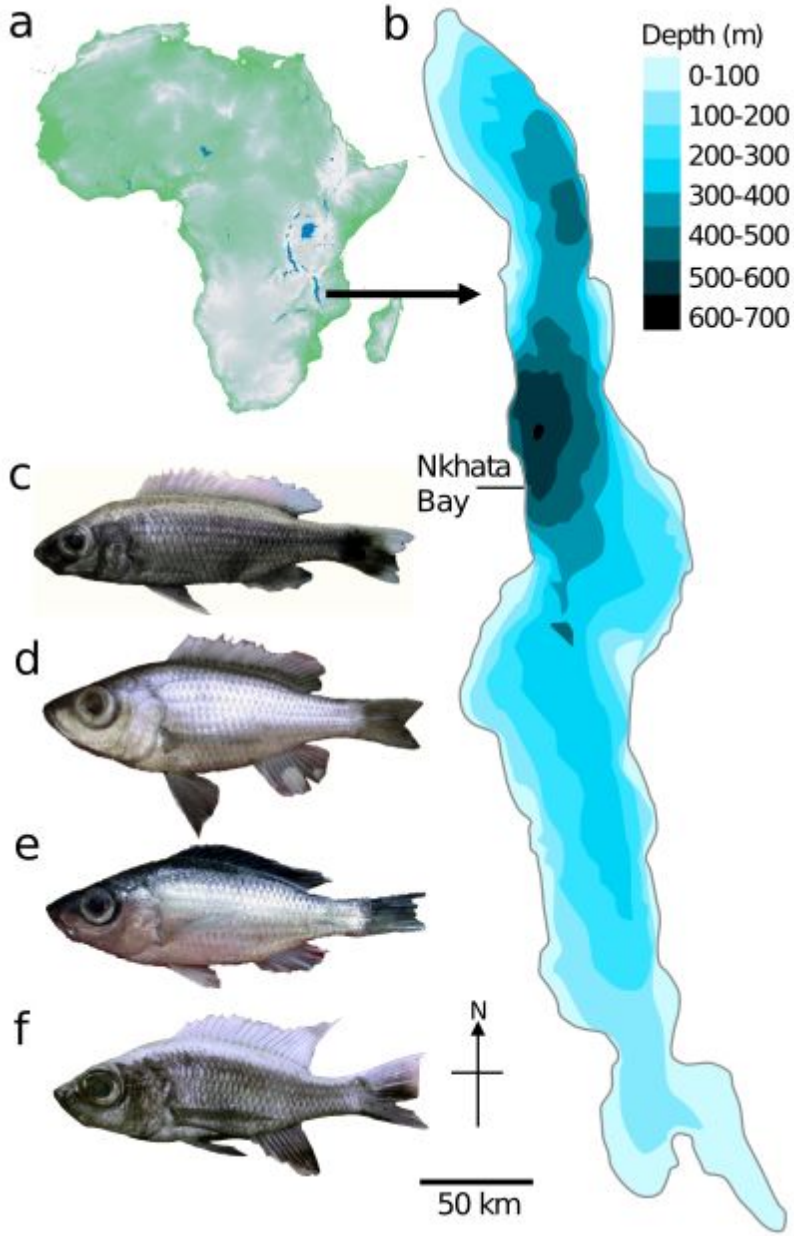

Figure 1. Maps of (a) Africa (topographic) and (b) Lake Malawi (bathymetric) indicating the sampling location Nkhata Bay. (c) *D*. 'limnothrissa black pelvic'; (d) *D*. 'macrops offshore'; (e) *D*. 'macrops black dorsal'; (f) *D*. 'macrops ngulube'.

Results

Population structure and genome-wide interspecific divergence

After stringent filtering the RAD data comprised 11,786 RAD tags that were each represented in at least 80% of individuals (minimum of 5) in each of the four populations (see Figure 1, Table S1). Three methods confirmed that these are indeed four different species: Maximum-likelihood inference based on a concatenated alignment of these RAD tags (total alignment length 1,053,675 bp) (Figure 2a), with branches separating species consistently receive high statistical support (bootstrap > 95%). *D.* 'macrops black dorsal', *D.*

'limnothrissa black pelvic', and *D.* 'macrops ngulube' were each reciprocally monophyletic, while *D.* 'macrops offshore' was paraphyletic with respect to *D.* 'limnothrissa black pelvic'. Principal component analysis (PCA) based on 11,786 SNPs (using only one single SNP per RAD tag) confirmed strong population structure, with putative conspecific individuals grouping into distinct, non-overlapping clusters along the first two principal components (Figure S1). Discriminant analysis of principal components (DAPC) consistently assigned individuals with putative conspecifics (cluster assignment probability for all individuals 100%, k=4, Figure S2). Interspecific genome-wide divergence (see Table S2) ranged from F_{ST} = 0.05 (*D.* 'limnothrissa black pelvic' vs. *D.* 'macrops offshore') to F_{ST} = 0.09 (*D.* 'limnothrissa black pelvic' vs. *D.* 'macrops ngulube'). Figure 2b illustrates genome wide patterns of F_{ST} divergence between species.

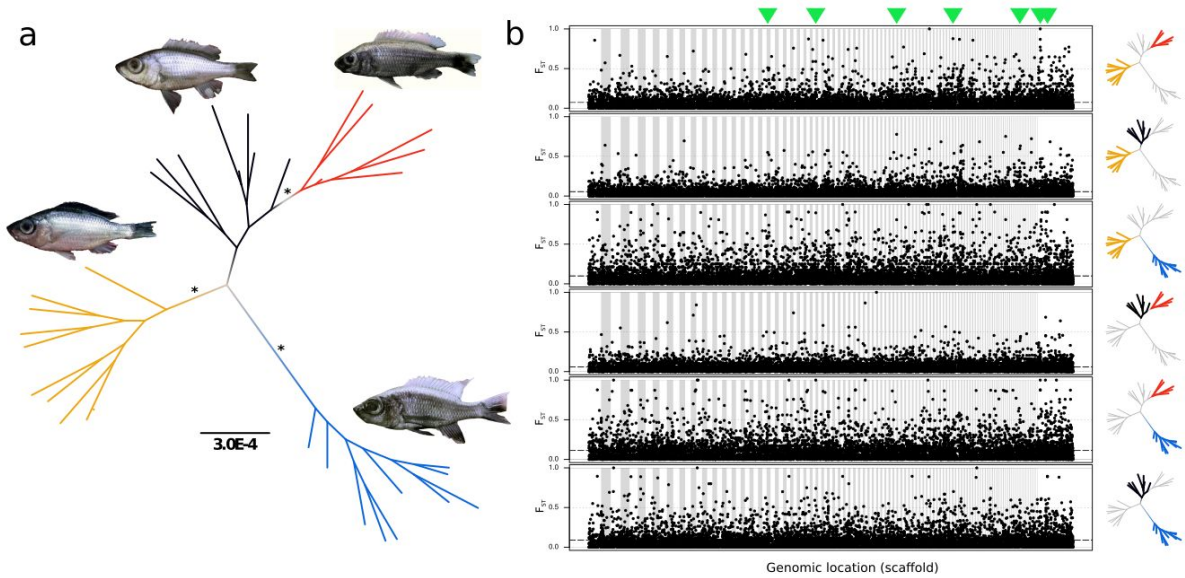

Figure 2. (a) Maximum Likelihood tree (unrooted) of phylogenomic structure among *Diplotaxodon* species. Yellow - *D.* 'macrops black dorsal'; red - *D.* 'limnothrissa black pelvic'; black - *D.* 'macrops offshore'; blue - *D.* 'macrops ngulube'. Asterisks indicate > 95% bootstrap branch support. Scale bar indicates genetic divergence (nucleotide divergence per site). (b) Genome wide pattern of pairwise F_{ST} divergence between populations. Scaffolds with minimum length of 100kb containing a minimum of 10 SNPs are displayed. Individual scaffold boundaries are indicated by alternate white/grey background. Dashed lines indicate the global pairwise F_{ST} average. Highlighted groups in the phylogenetic trees on the right hand side indicate the population pairs. Green arrow heads on top of the figure indicate locations of candidate regions supported by all three candidate outlier approaches (see Figure S3).

Candidate genomic regions under selection

We applied three independent outlier detection approaches, which highlighted 242 loci (2.1 % of total), distributed across 103 genomic scaffolds of the *Maylandia* (*Metriaclima*) *zebra* reference genome v1.1 [32]. The number of loci highlighted by individual methods ranged from 125 (1.1 % of total, *Stacks*) to 96 (0.8 % of total, *Bayenv*). Of the total candidate loci, 189 (77.8 %), 41 (16.9 %) and 13 (5.3 %), were highlighted independently by one, two, or all three methods, (see Figure S3) and these were assigned to 134, 26 and 8 candidate regions on 103, 26 and 8 genomic scaffolds, respectively. Table 1 summarizes the location and

characterizes the gene complements of the 26 genomic regions supported by at least two independent outlier identification approaches (Table S3 lists gene models associated with the candidate regions on the 103 genomic scaffolds).

Gene ontology enrichment analyses of genes in the candidate regions under divergent selection highlighted GO terms associated with sensory perception (e.g. axon extension, neuron development) and photoreceptor development (dynactin complex), embryonic development and morphogenesis (fibroblast growth factor receptor signalling pathway, positive regulation of cell proliferation, developmental growth involved in morphogenesis), and oxygen binding/transport (haemoglobin complex, oxygen transport/binding, gas transport, haem binding). Specific candidate genes under divergent selection associated with visual perception, and signal transduction include *ACTR1B* (Beta centractin), *Rh1* (Rhodopsin), *PPIase* (Peptidyl-prolyl cis-trans isomerase), *SGPP1* (Sphingosine-phosphate 1 phosphatase 1), and *DENND4B* (see Tables 1 and S5 for details). The phosphatase coded by *SGPP1* catalyses the degradation of sphingosine-1-phosphate, a key regulator in photoreceptor development [33,34]. With respect to putative physiological adaptations to deep-water environments, a genomic region centred around two haemoglobins, HBA (haemoglobin subunit alpha) and HBB1 (haemoglobin subunit beta-1), was highlighted by all three candidate outlier approaches.

Further candidates under putative selection include genes central for craniofacial and eye morphogenesis, such as *ALX3* (Aristaless-like 3), *PXDN* (Peroxidasin), *FOX* (Forkhead box transcription factor, 2x), *ZMIZ* (Zinc finger miz domain containing protein, 2x), *RPGRIP1L* (retinitis pigmentosa GTPase regulator interacting protein 1, synonym Fantom), *MEIS2* (homeobox meis2 protein), *FGFR1* (fibroblast growth factor receptor 1), *SOX* (Sry box transcription factor, 2x), *DKK1* (Dickkopf1), *NEUCRIN* (Draxin) and *TCF7L1* (transcription factor 7-like 1, formerly known as *TCF3*). *RPGRIP1L* has been shown to interact biochemically with *RPGR* (Retinitis Pigmentosa GTPase Regulator) [35], which plays a central role in controlling access of both membrane and soluble proteins to the photoreceptor outer segment. Loss and mutations in *RPGR* have been associated with a range of retinal diseases in human patients, including a variant of cone dystrophy, characterized by progressive dysfunction of photopic (cone-based) day vision with preservation of scotopic (rod-based) night vision [36]. *TCF7L1* is involved in the regulation of early embryonic craniofacial development via the Wnt/β-catenin signalling pathway and is expressed during human embryonic eye development [37]. Generally the Tcf/Lef family of molecules mediate canonical Wnt signaling by regulating downstream target gene expression [38]. *TCF7L1* has been demonstrated to directly repress *SOX4* [39] which is expressed in early zebrafish eye development [40] and also plays an active role in Wnt signalling by stabilizing β-catenin via a complex feedback loop [41]. Knockdown of *SOX4* in zebrafish resulted in structural malformations of the eye [40]. *SOX2*, also affects Wnt signalling via feedback inhibition, and has been shown to crucially regulate retina formation in *Xenopus* [42] and mice [43]. Mutations in the underlying gene have been associated with recessively inherited frontonasal malformation in humans [44]. Within the candidate regions identified by our analyses we found cases of co-localization of genes potentially functionally relevant for deep-water adaptation, such as the close proximity of *ZMIZ*, *DKK1* and *PPIase* (scaffold 197, Figures 3c and S3), and *ACTR1B*, *FGFR1* and *TCF7L1* (scaffold 45, see figure S3). *ACTR1B* is involved in regulating photoreceptor cell differentiation [45] and survival [46]. *ZMIZ* has been previously

shown to directly interact *in vitro* with *Msx2*, an important regulatory element involved in skull- [47] and specifically in eye morphogenesis [48]. *DKK1* is an antagonistic inhibitor of the Wnt/β-catenin signalling pathway, repeatedly implicated as a central mediator of craniofacial development in vertebrates [49,50], including cichlids [51,52]. In Lake Malawi cichlids an amino acid substitution in β-catenin has been found to be associated with alternate jaw morphologies. *DKK1* inhibits the stabilization of β-catenin by binding to *LRP6*. Colocalized with these two factors is a peptidyl-prolyl cis-trans isomerase (*PPIase*). Generally, *PPIases* are ubiquitous proteins, but in the context of this study it is worth noting that members of a *PPIase* subgroup (cyclophilins) have been shown to play a critical role in opsin biogenesis in *Drosophila* [53] and cattle [54]. For some of the above mentioned candidate genes our outlier approaches have highlighted more than one paralog on separate genomic scaffolds. *ZMIZ*, *FOX* and *SOX* transcription factors were present in two separate candidate regions, each (see tables 1 and S3).

Genomic regions associated with interspecific eye size variation

Both vertical (Figure 3a) and horizontal (Figure S5) eye diameter differed significantly among the four *Diplotaxodon* species (ANOVA: vertical eye diameter $F_{3,171}$ = 104.6, P < 0.001; horizontal eye diameter $F_{3,171}$ = 127.7, P < 0.001;), with *D. limnothrissa* eye diameters being consistently significantly smaller in all pairwise comparisons (TukeyHSD: P < 0.001, Table S4). Enlargement of the eyes is often associated with adaptation to deep-water environments [13,14]. We applied a Bayesian linear model approach using the observed eye size differences to identify genomic regions particularly associated with this phenotypic trait. These analyses highlighted the population allele frequencies of 42 loci (0.4 % of total, Figure 3b), clustering into 37 genomic windows, as being highly correlated with interspecific eye diameter differences, indicated by their average relative rank (ARR >= 0.95) across 20 independent Bayenv runs. Eight SNPs (0.07 % of total) were found to be highly significantly associated with eye diameter using our most stringent filtering criteria (ARR >= 0.95 and smoothed ARR p < 0.001). This set of candidate SNPs were in six genomic regions (Figure 3c). Three of these six genomic regions inferred as most significantly correlated with interspecific eye diameter variation, are also highlighted as candidate regions under selection by at least two of three independent outlier detection approaches (Figure 3c).
GO enrichment analyses of the gene complements identified by the eye diameter-informed analyses indicates significant overrepresentation of GO terms associated with e.g. ‘regulation of response to external stimulus’ and ‘intermediate filament organization’. The corresponding genomic regions contain genes encoding for the photopigments, *Rh1*, *OPN4* (melanopsin) and *OPN5* (neuropsin), as well *SGPP1 and PPIase*, a structural eye lens protein (*BFSP2*), genes expressed during normal eye development (*TMX3, XFIN*), and central transcription factors regulating embryonic craniofacial development (*ZMIZ, DKK1, ALX3*). Particularly *BFSP2*, *OPN4*, *TMX3* and *XFIN* were found in close proximity (scaffold 215). Tables 2 and S5 detail the gene complements in the genomic regions most significantly associated with eye size differences. Candidate genes contained in regions highlighted by both the eye morphology informed analysis and the candidate outlier detection approaches include *Rh1*, *SGPP1*, *ZMIZ*, *DKK1*, *PPIase*, *ALX3* and *RPGRIP1L* (Table S5).

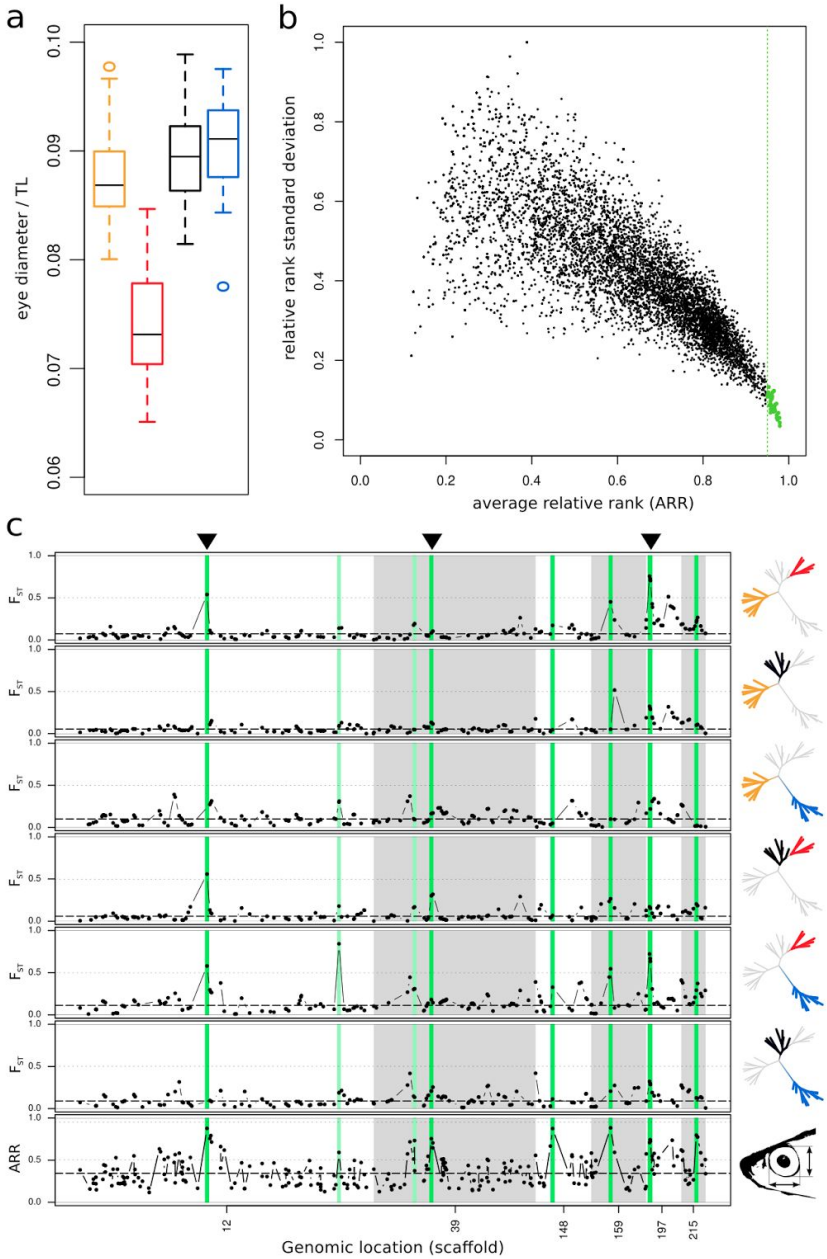

Figure 3. (a) *Diplotaxodon* eye diameter (normalized by total length of the fish, TL) across the four species. Yellow - *D*. 'macrops black dorsal'; red - *D*. 'limnothrissa black pelvic'; black - *D*. 'macrops offshore'; blue - *D*. 'macrops ngulube'. (b) Pope plot, summarizing correlations of population allele frequency with vertical eye diameter, inferred from 20 independent Bayenv runs. Dots illustrate per locus average relative rank (ARR) versus relative rank standard deviation. The green vertical line delimits the 95th ARR percentile. The 42 loci most consistently correlated with eye size (ARR >= 0.95) are shown in green. (c) Pairwise F_{ST} divergence (top six panels) and global allele frequency correlations with interspecific vertical eye diameter differences (bottom panel). Displayed are the six scaffolds containing the most significantly correlated loci. Corresponding regions are highlighted in shades of green (lightgreen - ARR >= 0.95; darkgreen - ARR >= 0.95 and smoothed ARR $P < 0.001$). Dots represent the kernel smoothed averages across 50kb windows. Dashed lines indicate the genome wide average of F_{ST}/ARR. Population pairs are indicated by the highlighted populations in the phylogenetic trees on the right hand side. Black arrowheads on top indicate regions that were also supported as candidate outlier

loci by at least two of three independent outlier detection approaches. Tables 2 and S5 summarizes the gene complements in highlighted regions.

Functional adaptations in candidate genes

Using genome-level resequencing data obtained via the Malawi cichlid diversity sequencing project (NCBI Bioproject PRJEB1254), we examined nucleotide polymorphisms to identify potentially functional substitutions in a number of genes highlighted by the population level analyses described above, specifically the genes coding for rhodopsin, phakinin, melanopsin, as well as haemoglobin subunits alpha and beta. The resequencing data include 156 individuals from 78 haplochromine species from in and around Lake Malawi, including six *Diplotaxodon* species. These analyses revealed a number of non-synonymous substitutions, as well as indel-, 5'- and 3'-UTR polymorphisms, potentially relevant for visual and physiological adaptation to deep-water conditions (Table S6). Across *Diplotaxodon* species, we identified a total of eight non-synonymous and three 3'-UTR polymorphisms in the RH1 gene coding for rhodopsin. Three of the non-synonymous polymorphisms involve variants so far restricted (private) to *Diplotaxodon*, i.e. were not observed in any of the other samples from the greater Lake Malawi species flock. One further non-synonymous variant appears private to the *Diplotaxodon-Pallidochromis* lineage. Three of the amino acid residues affected by non-synonymous substitutions (amino acid positions 83, 133 and 189) were previously considered likely to be involved in spectral tuning (see discussion for further details). Variants private to *Diplotaxodon*, were also detected in the genes coding for phakinin and melanopsin (Table S6). We detected a total of 13 amino acids containing non-synonymous polymorphisms in the haemoglobin subunit beta gene. One of the variants appeared private to *Diplotaxodon*, while several others were observed in high frequency in *Diplotaxodon* and in additional taxa of a distinct deep-water lineage of benthic Lake Malawi cichlids, including *Alticorpus*, *Lethrinops* and *Aulonocara*. Several of the affected residues are associated with changes in O_2 affinity in the human haemoglobin subunit beta homologue.

Discussion

We identified genomic regions showing signals of strong differentiation between four species of deep water *Dipotaxodon* species in Lake Malawi. In shallow water cichlids, species tend to segregate in habitat, diet and depth distributions. Thus, we predicted such resource partitioning is likely to be taking part in deep water cichlids, and that genomic regions associated with adaptation should appear overrepresented in outlier loci. Gene ontology enrichment analyses highlighted GO terms associated with sensory perception and photoreceptor development, embryonic development and morphogenesis, and oxygen binding/transport. We first used RADseq data, and assumed regions within 50kb of the highlighted SNP loci were candidates for selection among *Diplotaxodon* species. We subsequently used genome-level resequencing data and found that patterns of SNP diversity in coding (and regulatory) regions may reflect ecological differences between *Diplotaxodon* and the rest of the species flock.

Our analyses indicate a role for key genomic regions containing genes which could be associated with adaptation to depth. A region centred around two haemoglobin genes shows a signal consistent with selection, and this region is relevant for physiological adaptation to deep-water conditions in two ways. The first is through enhancing the "Root effect", a pH dependent decrease in oxygen-carrying capacity of some fish haemoglobins. This facilitates O_2 secretion into the swim bladder, the specialized organ found in most teleost fishes used to achieve neutral buoyancy in open water by regulating partial gas pressure in response to the ambient hydrostatic pressure. At the molecular level, such 'Root haemoglobins' are characterized by a range of amino acid replacements in the globin α- and β-chains, compared to 'normal' haemoglobins [55–57] and may have evolved independently a number of times [58]. The two affected *Diplotaxodon* haemoglobin subunit genes fulfil the minimal structural requirements for Root haemoglobins as previously defined [59]. The second way selection could act on this region is via haemoglobin O_2 binding affinity [60]. Given teleost fishes can differ in their tolerance towards low levels of dissolved oxygen, and the amount of dissolved oxygen in water typically decreases with depth, it is plausible that divergent selection is operating on oxygen tolerance in Lake Malawi [61]. Meromictic freshwater lakes exhibit complete oxygen depletion below a defined oxic-anoxic boundary layer, which in Lake Malawi has been identified at ~230 m water depth [62]. Selection for high O_2 affinity haemoglobin alleles was previously demonstrated in response to altitude related hypoxia in birds [63–65]. A number of the residues affected by non-synonymous variants in *Diplotaxodon* (Table S6) have been associated with changes in oxygen affinity in the human haemoglobin subunit beta-1. While both the Root effect and haemoglobin oxygen binding affinity have previously been predicted to be likely targets of natural selection [2,61], the current study is, to our knowledge, the first to find evidence for selection associated with haemoglobin genes in fish. The exact effect of the observed non-synonymous changes on the Root effect and/or the O_2 affinity of *Diplotaxodon* haemoglobins needs to be determined experimentally, but the presence of the same changes in other unrelated deep-water Lake Malawi genera such as deep-benthic *Lethrinops* and *Alticorpus* spp. is consistent with adaptation to depth. Whether these changes have arisen multiple times independently, or been acquired through introgression warrants further investigation. There is evidence of hybridization as a driver in adaptive radiation in the Lake Malawi species flock [66–68] and these loci could allow a test of whether these deep water adaptations have been retained after a hybridization event, and allowed subsequent radiation into this challenging habitat [67].

Previous work has shown that the four *Diplotaxodon* species we studied differ significantly with respect to head- and overall body morphology [27], and craniofacial variation in cichlids is frequently associated with trophic adaptations. Strongly differentiated cranial dimensions include eye size differences and light conditions at different water depths are a likely selective driving force. We identified candidate genomic regions under selection that are highly enriched for genes involved in the regulation of craniofacial development. Our F_{ST} outlier detection approaches highlight a number of *Wnt* factors, further supporting the central role of *Wnt* signalling for regulating cichlid craniofacial gene expression [69]. One genomic region in particular (scaffold 197, Figure 3b) received strong support in all analyses, and contains a group of colocalized genes; namely *ZMIZ* (skull and eye morphogenesis), DKK1 (vertebrate craniofacial development) and a PPIase (opsin biogenesis). The close proximity

of *ZMIZ* and *DKK1* in particular may indicate they are inherited together as a craniofacial "supergene".

The larger eye of species in the *Diplotaxodon macrops* "complex" relative to those in the *Diplotaxodon limnothrissa* "complex" is consistent with their presumed depth distributions (Figures 3a and S5). The environment at depths of 50 - 200 m [70] is depauperate of long wavelength light and dominated by shorter wavelength blue light as depth increases [12]. Our analyses consistently highlight a genomic region centred around the *Rh1* gene (scaffold 12, Figure 3b), which codes for rhodopsin, the principal photopigment of retinal rod photoreceptors, central for scotopic vision under dim light conditions. Changes in both the coding sequence as well as in gene expression of opsins [71] may mediate visual performance across light environments, and the pivotal role of rhodopsin for visual adaptation in deep water light environments has been confirmed by a number of studies [6,19,25,72,73]. Whether the observed non-synonymous variants in the rhodopsin gene (Table S6) result in functionally important shifts of the spectral sensitivity of the photopigment between *Diplotaxodon* species needs further investigation. However, it is worth noting that three of the observed 10 amino acid residues found affected by non-synonymous polymorphisms identified in the current study (amino acid positions 83, 133 and 189) have previously been considered likely to be involved in spectral tuning, because of their close proximity to the chromophore or the chromophore-binding pocket [25]. Amino acid replacements at position 83, specifically, have been demonstrated experimentally to cause spectral shifts towards blue [25]. A further four affected residues (166, 169, 297, 298) have also been implicated in a recent study of spectral shifts via mutations in the rhodopsin gene in an isolated East African crater lake [21].

Our eye-size informed analyses suggest that *BFSP2*, the gene coding for phakinin, has been diverging among *Diplotaxodon* species. The lens of the vertebrate eye is composed of specialized epithelial lens fibre cells containing beaded filaments, specific cytoskeletal structures unique to the lens [74]. Phakinin is one of the two principal proteins forming the beaded filaments, so our results also suggest a role for eye lens structure in adaptation to dim, short wave-length light environments, a concept that has so far attracted very little attention. The role of beaded filaments in lens biology is not fully understood, but they appear essential in maintaining optical clarity and transparency of the lens [75,76]. Mutations in *BFSP2* are associated with cataract formation, i.e. a clouding of the eye lens in humans [77,78]. Cataracts reduce the intensity and alter the chromaticity of light traveling through the lens [79], with potentially great effect on visual perception [80]. After cataract surgery patients usually report a change in color appearance associated with additional short-wavelength light reaching the retina [79,80]. The wider implication is that *BFSP2* may specifically regulate lens transparency for blue light and could be a key- and previously unrecognised mediator for adaptation in dim, blue-dominated light environments.

BFSP2 is co-localized with *OPN4*, *TMX3* and *XFIN*. The role of the latter genes in visual adaptation is unknown. However, *XFIN* is expressed during the formation of the retina in *Xenopus* [81] and deletion of the *TMX3* gene in humans has been linked to a genetic disease associated with retarded growth of the eye [82]. *OPN4* codes for the opsin based photopigment melanopsin, which is central to a distinct photoreceptor class, the melanopsin retinal ganglion cells (mRGCs), that was discovered only relatively recently [83]. Initially

mRGCs were shown to mediate so-called non-image forming visual responses [84], but more recent evidence suggests that mRGCs may contribute significantly to assessing brightness and play a more general role in supporting vision in mammals [85] and it may well play a role in deep-water cichlid vision. In Lake Malawi cichlids the associated genomic region is clearly highly enriched for vision related genes and might represent another coadapted gene complex. Exploration of whole genome resequencing data obtained for the larger Malawi flock revealed non-synonymous mutations private to the deep water lineage in both the *BFSP2* and the *OPN4* gene, as well as intron and 5'-UTR variation in the respective genes between *Diplotaxodon* species. Further investigation is required to fully understand how this genomic region may be involved in deep-water visual adaptation. The potential role of lens structure in adaptation is further confirmed by the highlighting of the *FGFR1*, which is considered essential for lens fibre differentiation [86] and *MEIS2*, a gene that directly regulates *Pax6* during vertebrate lens morphogenesis. The latter transcription factor has been demonstrated to play essential roles in lens differentiation and has previously been referred to as the 'master control gene for morphogenesis and evolution of the eye' [87,88].

Depth- and habitat segregation may be caused by a number of factors including competition for resources, breeding territories or enemy-free space [89]. The four *Diplotaxodon* also differ in male nuptial coloration [27] which strongly implies an important role of visually informed mate choice, despite the twilight conditions they experience in their natural environments. While many pelagic fish use counter-shading in background matching for camouflage [90], *D. '*macrops ngulube' and *D. '*macrops black dorsal' males have an opposite nuptial coloration which may allow them to be more visible. The sensory drive hypothesis[91] predicts that visual systems (along with signals and signalling behaviour) will differentiate if local environments differ in their signal transmission quality. The observed interspecific differences in male nuptial colour in *Diplotaxodon* in combination with the inferred genomic footprints of sensory adaptation are consistent with the idea that reproductive isolation could arise as a consequence of sensory drive in deep water systems.

In summary, our work has shown that the selection pressures associated with deep water environments can be identified by their effect on the genome. In addition to genes previously associated with depth related spectral shifts (rhodopsin), we identify novel mechanisms of adaptation to deep water conditions worth further investigation (e.g. Root effect haemoglobins and eye lens filament proteins). Outlier tests tend to highlight large-effect loci with relatively simply genetic architecture [92] so these will be interesting possibilities with which to identify parallel evolution in other systems such as populations experiencing high altitude hypoxia, or marine deep water systems. Our results provide evidence of fixed genomic changes since deep water conditions in Lake Malawi were attained possibly as recently as 75,000 years ago [28,29], and raise the intriguing possibility that hybridization between *Diplotaxodon* and the deep-benthic clade of cichlids may have facilitated the latter's expansion into the twilight zone. Finally, we find that in candidate genomic regions under selection functionally associated genes are frequently in close proximity, such that cichlid adaptations and ecological differentiation may be facilitated by the presence of linked, coadapted gene complexes, or "supergenes".

Methods

RAD data sampling, DNA extraction, library preparation and Illumina sequencing

A total of 40 individuals from four *Diplotaxodon* species were collected from Nkhata Bay (Figure 1, Table S1), photographed, and fin clips stored in ethanol at -80°C. Genomic DNA was extracted from fin clips using the Qiagen DNAeasy Blood and Tissue Kit, according to manufacturer's instructions. DNA was quantified using PicoGreen fluorimetry (Quant-iT PicoGreen Kit, Invitrogen) and quality checked on 0.8% agarose gels. Paired-end RAD libraries were prepared by the NERC/NBAF facility at Edinburgh Genomics, following [93] and [94]. Genomic DNA was digested using *SbfI*, a barcoded RAD P1 adapter ligated, followed by sonic shearing, size selection and ligation of P2 adapters. Libraries were PCR amplified, quantified and sequenced in separate flow cells on an Illumina HiSeq 2000 platform with 100 bp, paired-end chemistry.

RAD data processing

Modules from the Stacks v.1.20 [95] program suite were used for the initial processing of the RAD data as follows: Raw reads were demultiplexed based on their in-line barcode and quality trimmed using the program *process_radtags*. Putative PCR-duplicates were subsequently removed using the program *clone_filter* from Stacks v.1.20. The remaining reads for each individual were mapped to the draft genome of *M. zebra* [32], version MetZeb1.1_prescreen (downloaded from the link provided; last accessed 23.12.2016) http://archive.broadinstitute.org/ftp/pub/assemblies/fish/M_zebra/MetZeb1.1_prescreen/M_zebra_v0.assembly.fasta; using BWA v.0.7.10-r789 [96], allowing for up to 8 mismatches per read. Any reads mapping to more than one genomic location were removed from the dataset using a custom script (*split_sam.pl*) and the remaining reads were converted to bam format using Samtools v.0.1.19-4428cd [97]. The program *pstacks* from Stacks v.1.20 was used to group the mapped reads into stacks (minimum of 5 identical reads per stack) and call SNPs. A global catalog of RAD tags was then built using *cstacks* from all individuals with at least 20,000 valid tags identified by *pstacks* (individuals with fewer valid tags were omitted from further analyses). The *populations* program from Stacks v.1.20 was used to calculate basic population genetics statistics and to produce data files for downstream analyses (e.g. in plink format, structure format) for each population individually, for each pairwise population comparison and across a global dataset of all populations. Analyses were limited to tags identified in all populations and in at least 80% of individuals per population, respectively. Maximum likelihood inference was performed on the concatenated RAD tags using RAxML v8.2.4 [98]. SNP datasets to be used in downstream analyses were limited to only a single SNP per RAD-tag (*--write_single_snp* option in *populations*) to reduce the effects of linkage. We furthermore excluded any singleton SNPs, i.e. sites with minor allele counts of one, from any global analyses. Further conversions to file formats not supported by *populations* were performed by PGDSpider v.2.0.7.3 [99] or by custom scripts. Principal component analysis and discriminant analysis of principal components [100] was performed using the Adegenet package in R [101].

Identification of putative candidate loci under selection

Three independent approaches to outlier detection were applied: (1) Pairwise Fst values were calculated by *populations*. A Fisher's exact test as applied by the sliding window algorithm implemented in *populations* (window size 50kb) was used to calculate significance levels [95] and loci with p < 5e-7 in at least one pairwise comparison were considered candidate loci under selection identified by *Stacks*. (2) The Bayesian approach to outlier loci detection implemented in *Bayescan* v.2.1 [102] was applied to pairwise, as well as a global SNP dataset, produced by *populations*. In all Bayescan analyses, prior odds for the neutral model were set to 10, with the remaining parameters set to default. A false discovery rate (FDR) threshold of 0.05 was applied and loci with α < 0.05 in at least one pairwise comparison or the global Bayescan run were considered candidate loci identified by *Bayescan*. (3) *Bayenv* v.2 [103] was used to identify candidate outlier based on the global SNP dataset obtained for all four populations. *Bayenv* applies a Bayesian linear model method that accounts for population history by incorporating a covariance matrix of population allele frequencies and accounting for differences in sample size among populations [103]. 10 independent covariance matrices were constructed for sets of 5,000 SNPs randomly selected from the global dataset as represented in a vcf file produced by *populations*, using a custom script (*vcf_2_div.py*). In brief, for each random set covariance matrices were obtained by running *Bayenv* for 100,000 iterations. Convergence of covariance matrices was assessed visually in R [104] and a final covariance matrix was obtained by averaging across the 10 independent runs. Using this matrix to account for population history *Bayenv* was then run for 20 independent replicates with 1,000,000 iterations each on a set of SNPs present in at least 5 individuals in each population. Candidate loci under selection were identified based on the population differentiation statistic *xTx* generated by *Bayenv* as follows: For each run we ranked loci based on their respective value for *xTx*, and calculated a rank statistic similar to the empirical p-value approach used by [105]. In our approach the highest ranking SNP would be assigned the highest relative rank of 1, while the lowest ranking SNP would be assigned a rank of $1/N$, with N being the total number of SNPs in the dataset. For each SNP we then calculate the average relative rank (ARR) and standard deviation across the 20 independent *Bayenv* runs. Loci yielding an ARR in the 95th percentile were considered candidate loci under selection identified by *Bayenv*. Genomic candidate regions were defined as +-50kb windows up and downstream of candidate SNPs, supported by one, two, or three outlier detection approaches, respectively. Candidate regions of consecutive candidate SNPs were merged in case an overlap between the corresponding windows was detected.

Identification of SNPs correlated with morphological differences between populations

The Bayesian linear model approach implemented in *Bayenv* [103] is frequently used to infer correlations between SNP allele frequencies and environmental variables. The method yields Bayes factors (BF), which are interpreted as the weight of evidence for a model in which an environmental factor is affecting the distribution of variants relative to a model in which environmental factors have no effect on the distribution of the variant [105]. The four *Diplotaxodon* populations analysed in the current study have been previously identified to differ significantly in head morphological traits [27], which in turn are known to correlate with environmental variables in cichlids [106]. In particular, *Diplotaxodon* 'limnothrissa black pelvic'

has a smaller eye than the other species. We applied the *Bayenv* approach, using vertical and horizontal eye diameter (normalized by total fish length) from [27], to identify SNPs correlated with eye morphological differences between populations and to characterize the genomic regions involved in regulating the observed morphological differences. The full *Bayenv* workflow and the subsequent analyses are made available as Jupyter notebooks in a dedicated Github repository. Prior to *Bayenv* analyses population averages were standardized by subtracting the global mean and dividing the result by the global standard deviation. Standardized population averages were then used as 'environmental variables' in 20 independent runs of *Bayenv*, each using 1,000,000 iterations. Transformed rank statistics by means of average relative ranks (ARR) of Bayes factors were calculated across the 20 runs as described above. We then applied a Gaussian weighting function to generate a kernel-smoothed moving average of the transformed rank statistic for each polymorphic site (based on 100 kb sliding windows centered at each SNP). To test for statistical significance of windows we applied a bootstrap resampling procedure (10,000 permutations). In each permutation new values for ARR were sampled with replacement from across the dataset and the smoothed statistic was calculated for each replicate set using the coordinates of the original SNP for the weighing function. For each SNP, the data obtained via bootstrap resampling was used as empirical null distribution of the test statistic against which the original smoothed average ARR was compared to determine a P-value. The approach we have implemented is similar to the method used by the *populations* program of the *stacks* software suite [95]. Windows significant at the $p<0.001$ level were considered candidate regions associated with eye morphological differences if they also contained at least one individual SNP locus yielding an ARR in the 95th percentile of the original distribution.

Identification the function of genomic regions under selection

Gene models for *M. zebra* [32] were obtained from the Broad Institute (downloaded from ftp://ftp.broadinstitute.org/pub/vgb/cichlids/Annotation/Protein_coding/; last accessed 23.12.2016). Peptide sequences were subjected to a similarity search against a custom build of the NCBI's nr protein database (restricted to Metazoan proteins) using BLAST [107] and screened for known domains using InterProScan 5.8-49.0 [108]. Results from these analyses were reconciled in Blast2GO v.3 [109] to obtain putative functional annotation including Gene Ontology (GO) terms [110] for *M. zebra* gene models, where possible. GO term enrichment analyses using Fisher's exact tests with multiple testing correction of FDR [111], as implemented in Blast2GO v.3 [109], were performed for gene complements in genomic candidate regions supported by one, two, or three candidate outlier approaches, respectively.

Whole genome resequencing data

Following up on our initial results, we selected the genes coding for rhodopsin, phakinin and melanopsin, as candidate genes for visual adaptations, as well as the genes coding for haemoglobin subunits alpha and beta-1, potentially involved in physiological adaptations to deep water conditions. We downloaded data mapping to scaffolds 12, 81, and 215 (where the candidate genes are located) from the Cichlid Diversity Sequencing project (SRA accession: PRJEB1254). Variant calling was carried out as in [21]. We used these data to

examine a number of candidate genes highlighted by our analyses in more detail with respect to putatively functionally relevant nucleotide polymorphisms in *Diplotaxodon* as well as the greater Lake Malawi cichlid flock.

Reproducibility statement

To ensure reproducibility of our analyses we have deposited detailed descriptions of the bioinformatics steps, custom scripts and further supplementary files (e.g. detailed morphological measurements, functional gene annotation results, GO enrichment tables) in a Github repository (https://github.com/HullUni-bioinformatics/Diplotaxodon_twilight_RAD; DOI: 10.5281/zenodo.259371). The raw RAD sequencing data is deposited with NCBI (Bioproject: PRJNA347810; SRA accessions: SRX2269491-SRX2269504). Whole genome resequencing data for the greater Lake Malawi flock is deposited on Genbank as part of the Cichlid diversity sequencing project (Bioproject accession: PRJEB1254).

Author contributions

The paper was written by all authors. CH, MJG and DAJ conceived it, MJG collected and shared samples. Data analysis was carried out by CH.

Acknowledgements

The RAD data was produced by Edinburgh Genomics, as a result of a NERC funded grant awarded to DAJ (NE/K000829/1). We would like to extend our thanks to Eric Miska, Richard Durbin and Milan Malinsky for permission to use the whole-genome data, assistance with variant calling, and comments on the manuscript.

Table legends

Table 1: Genomic location (scaffold id), putative functional annotation of genes (if available) and RADtag ids localized in candidate regions under selection. Numbers in parenthesis after the RADtag id describe the number of independent outlier detection approaches supporting the respective tag, followed by the same tag's ARR and smoothed ARR level of significance (NS: > 0.05; *: 0.05-0.005; **: 0.005-0.001; ***: <0.001), inferred by the eye morphology informed analysis. Displayed are only gene complements for regions supported by at least two independent outlier detection approaches. Full list of highlighted regions including gene ids can be found in supplementary table S3.

scaffold id	functional gene annotation	tag id
12	membrane-associated guanylate ww and pdz domain-containing protein 1-like isoform x4, rod	6728(2|0.9683,***)

	opsin	
29	pol polyprotein	25871(3|0.7601,NA)
39	retrovirus polyprotein, serine threonine-protein phosphatase 2a 56 kda regulatory subunit epsilon isoform-like isoform x2, sphingosine-1-phosphate phosphatase 1-like, synaptotagmin-16-like, wd repeat-containing protein 89-like	31651(2|0.9505,***), 31655(2|0.7980,NA)
45	beta-centractin-like isoform x1, btb poz domain-containing protein kctd9-like, e3 ubiquitin-protein ligase march5-like, fibroblast growth factor receptor 1-a-like isoform x1	53279(3|0.8330,NA)
48	e3 ubiquitin-protein ligase siah2-like, neuronal acetylcholine receptor subunit alpha-10-like, olfactory receptor 2k2-like	35820(2|0.9254,NA)
55	aminoacyl trna synthase complex-interacting multifunctional protein 2-like isoform x2, parvalbumin beta-like, PREDICTED: uncharacterized protein LOC101481854, serine threonine-protein kinase lmtk2-like isoform x2, thymic cpv3-like, voltage-dependent calcium channel gamma-1 subunit-like, voltage-dependent calcium channel gamma-4 subunit-like	38241(2|0.7076,NA)
81	aquaporin-8-like, delphilin-like isoform x1, forkhead box protein j1-a-like, hemoglobin subunit alpha-d-like isoform x1, hemoglobin subunit beta-1-like, inactive rhomboid protein 1-like isoform x1, leucine carboxyl methyltransferase 1-like, rho gtpase-activating protein 17-like isoform x1	45603(3|0.7580,NA), 45610(3|0.9135,NA)
111	forkhead box protein b1-like, transcription initiation factor iia subunit 2-like, zinc finger protein 395-like, zinc finger protein 395-like isoform x1	3666(2|0.8285,NA), 3667(2|0.8252,NA)
114	low-density lipoprotein receptor-related protein 1b-like	4122(3|0.8057,NA), 4123(2|0.7773,NA)
133	polymeric immunoglobulin receptor-like isoform x2	7497(2|0.8258,NA), 7498(2|0.8489,NA)
136	sec14 domain and spectrin repeat-containing protein 1, solute carrier family 22 member 5-like, zinc finger protein 385b-like isoform x1	7844(2|0.8799,NA)
162	myelin transcription factor 1-like protein, peroxidasin homolog isoform x1	11894(2|0.9226,NA)
174	2-aminoethanethiol dioxygenase-like, cd48 antigen, cd48 antigen-like, kinesin-like protein kif11-like, solute carrier family 22 member 15-like, tbc1 domain family member 12-like	13433(3|0.6997,NA)

	isoform x1	
190	alpha- -mannosylglycoprotein 6-beta-n-acetylglucosaminyltransferase b-like, cytosolic fe-s cluster assembly factor nubp2-like, probable crossover junction endonuclease eme2-like isoform x1, spry domain-containing socs box protein 3-like, tbc1 domain family member 24-like isoform x1, vacuolar atp synthase 16 kda proteolipid subunit	15577(2|0.8195,NA)
197	2-oxoglutarate mitochondrial isoform x1, 2-oxoglutarate mitochondrial-like, af355375_1 reverse transcriptase, chromaffin granule amine transporter-like, complement c5, dickkopf-related protein 1-like, disintegrin and metalloproteinase domain-containing protein 9-like, peptidyl-prolyl cis-trans isomerase-like, pr domain zinc finger protein 12-like, telo2-interacting protein 2-like, unnamed protein product, zinc finger miz domain-containing protein 2-like	16065(3|0.9620,***), 16067(3|0.9670,***), 16069(3|0.9605,***), 16078(2|0.6278,NA)
203	titin isoform x7, titin-like, trichohyalin-like isoform x5	17952(2|0.8735,NA), 17967(2|0.8692,NA), 17970(2|0.8610,NA)
212	neuropeptide y receptor type 2-like	19042(2|0.9391,NA)
215	25-hydroxycholesterol 7-alpha-hydroxylase-like, armadillo repeat-containing protein 1-like, histone-lysine n-methyltransferase prdm9-like	19195(2|0.8150,NA), 19205(2|0.8706,NA), 50449(2|0.6460,NA)
219	e3 ubiquitin-protein ligase rnf168-like, neuronal tyrosine-phosphorylated phosphoinositide-3-kinase adapter 2-like isoform x1, papilin-like, sodium potassium-transporting atpase subunit beta-3-like, transmembrane 4 l6 family member 5-like, zinc finger protein 665-like	19444(2|0.8291,NA), 19447(3|0.8121,NA), 19543(3|0.7766,NA)
227	c-1-tetrahydrofolate cytoplasmic-like, zinc finger and btb domain-containing protein 1-like isoform x1, zinc finger and btb domain-containing protein 25-like isoform x1	20380(2|0.9414,NA)
242	gtp cyclohydrolase 1-like, low quality protein: wd repeat and hmg-box dna-binding protein 1-like, protein smaug homolog 1-like isoform x1	21791(3|0.7555,NA)
344	protein heg homolog 1-like	29332(2|0.7725,NA)

Table 2: Genomic location (scaffold id), putative functional annotation of genes (if available) and RADtag ids localized in candidate regions most significantly associated with interspecific eye size variation. Numbers in parenthesis after the RADtag id describe the tag's ARR and smoothed ARR level of significance (NS: >0.05; *: 0.05-0.005; **: 0.005-0.001; ***: <0.001). Displayed are only gene complements for regions supported at significance level p<0.001.

Full list of highlighted regions supported by ARR >= 0.95 including gene ids can be found in supplementary table S5.

scaffold id	functional gene annotation	tag id
12	membrane-associated guanylate ww and pdz domain-containing protein 1-like isoform x4, rod opsin	6728(0.9683,***)
39	a-kinase anchor protein 6, neuronal pas domain-containing protein 3-like, serine threonine-protein phosphatase 2a 56 kda regulatory subunit epsilon isoform-like isoform x2, sphingosine-1-phosphate phosphatase 1-like, synaptotagmin-16-like, wd repeat-containing protein 89-like	31602(0.9562,*), 31651(0.9505,***)
159	dynein heavy chain axonemal-like	11119(0.9666,***)
197	2-oxoglutarate mitochondrial isoform x1, 2-oxoglutarate mitochondrial-like, af355375_1 reverse transcriptase, chromaffin granule amine transporter-like, dickkopf-related protein 1-like, disintegrin and metalloproteinase domain-containing protein 9-like, peptidyl-prolyl cis-trans isomerase-like, zinc finger miz domain-containing protein 2-like	16065(0.9620,***), 16067(0.9670,***), 16069(0.9605,***)
215	afg3-like protein 2-like, c-c chemokine receptor type 9-like, chromodomain y-like isoform x2, fas-activated serine threonine kinase, kelch-like protein 18-like isoform x1, melanopsin-like, nucleolar transcription factor 1-like, phakinin-like, poly -binding-splicing factor puf60-like isoform x1, protein disulfide-isomerase tmx3-like, zinc finger protein xfin-like isoform x1	19268(0.9664,***)

References

1. Morita, T. High-pressure adaptation of muscle proteins from deep-sea fishes, Coryphaenoides yaquinae and C. armatus. *Ann. N. Y. Acad. Sci.* **1189,** 91–94 (2010).

2. Marquis, R. E. *et al. Effects of high pressure on biological systems*. **17,** (Springer Science & Business Media, 2012).

3. Jennings, R. M., Etter, R. J. & Ficarra, L. Population differentiation and species

formation in the deep sea: the potential role of environmental gradients and depth. *PLoS One* **8,** e77594 (2013).

4. Brown, A. & Thatje, S. Explaining bathymetric diversity patterns in marine benthic invertebrates and demersal fishes: physiological contributions to adaptation of life at depth. *Biol. Rev. Camb. Philos. Soc.* **89,** 406–426 (2014).
5. Wilson, G. & Hessler, R. R. Speciation in the deep sea. *Annu. Rev. Ecol. Syst.* **18,** 185–207 (1987).
6. Shum, P., Pampoulie, C., Sacchi, C. & Mariani, S. Divergence by depth in an oceanic fish. *PeerJ* **2,** e525 (2014).
7. Pradillon, F. & Gaill, F. Pressure and life: some biological strategies. *Rev. Environ. Sci. Biotechnol.* **6,** 181–195 (2006).
8. Somero, G. N. Adaptations to high hydrostatic pressure. *Annu. Rev. Physiol.* **54,** 557–577 (1992).
9. Brauer, R. W. & Torok, Z. Hydrostatic pressure effects on the central nervous system: perspectives and outlook [and discussion]. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **304,** 17–30 (1984).
10. Bowmaker, J. K. & Hunt, D. M. Evolution of vertebrate visual pigments. *Curr. Biol.* **16,** R484–9 (2006).
11. Tyler, P. A. *Ecosystems of the deep oceans*. (Elsevier, 2003).
12. Von der Emde, G., Mogdans, J. & Kapoor, B. G. *The senses of fish: adaptations for the reception of natural stimuli*. (Springer Science & Business Media, 2012).
13. Marshall, N. B. & Marshall, O. *Aspects of deep sea biology*. (Hutchinson London, 1954).
14. de Busserolles, F., Fitzpatrick, J. L., Paxton, J. R., Marshall, N. J. & Collin, S. P. Eye-size variability in deep-sea lanternfishes (Myctophidae): an ecological and phylogenetic study. *PLoS One* **8,** e58519 (2013).
15. Hunt, D. E., Rawlinson, N. J. F., Thomas, G. A. & Cobcroft, J. M. Investigating

photoreceptor densities, potential visual acuity, and cone mosaics of shallow water, temperate fish species. *Vision Res.* **111,** 13–21 (2015).

16. Landgren, E., Fritsches, K., Brill, R. & Warrant, E. The visual ecology of a deep-sea fish, the escolar Lepidocybium flavobrunneum (Smith, 1843). *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369,** 20130039 (2014).
17. Fernald, R. D. in *Sensory Biology of Aquatic Animals* 435–466 (1988).
18. Wang, F. Y., Tang, M. Y. & Yan, H. Y. A comparative study on the visual adaptations of four species of moray eel. *Vision Res.* **51,** 1099–1108 (2011).
19. Nakamura, Y. *et al.* Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 11061–11066 (2013).
20. Hunt, D. M., Hankins, M. W., Collin, S. P. & Justin Marshall, N. *Evolution of Visual and Non-visual Pigments:* (Springer US, 2014).
21. Malinsky, M. *et al.* Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* **350,** 1493–1498 (2015).
22. Carleton, K. Cichlid fish visual systems: mechanisms of spectral tuning. *Integr. Zool.* **4,** 75–86 (2009).
23. Cortesi, F. *et al.* Ancestral duplications and highly dynamic opsin gene evolution in percomorph fishes. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 1493–1498 (2015).
24. Hunt, D. M., Fitzgibbon, J., Slobodyanyuk, S. J., Bowmaker, J. K. & Dulai, K. S. Molecular evolution of the cottoid fish endemic to Lake Baikal deduced from nuclear DNA evidence. *Mol. Phylogenet. Evol.* **8,** 415–422 (1997).
25. Sugawara, T. *et al.* Parallelism of amino acid changes at the RH1 affecting spectral sensitivity among deep-water cichlids from Lakes Tanganyika and Malawi. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 5448–5453 (2005).
26. Turner, G. F., Robinson, R. L., Shaw, P. W. & Carvalho, G. R. in *The cichlid diversity of*

Lake Malawi/Nyasa/Niassa: identification, distribution and taxonomy (ed. Snoeks, J.) 198–251 (Cichlid Press, 2004).

27. Genner, M. J. *et al.* Reproductive isolation among deep-water cichlid fishes of Lake Malawi differing in monochromatic male breeding dress. *Mol. Ecol.* **16,** 651–662 (2007).
28. Delvaux, D. Age of Lake Malawi (Nyasa) and water level fluctuations. *Muse´e royal de l'Afrique Centrale (Tervuren), De´partement de Ge´ologie et Mine´ralogie. rapport annuel.* **1993(1994),** 99–108 (1993).
29. Ivory, S. J. *et al.* Environmental change explains cichlid adaptive radiation at Lake Malawi over the past 1.2 million years. *Proc. Natl. Acad. Sci. U. S. A.* (2016). doi:10.1073/pnas.1611028113
30. Genner, M. J. *et al.* Population structure on breeding grounds of Lake Malawi's 'twilight zone' cichlid fishes. *J. Biogeogr.* **37,** 258–269 (2010).
31. Thompson, A. B. & Allison, E. H. Distribution and breeding biology of offshore cichlids in Lake Malawi/ Niassa. *Environ. Biol. Fishes* **41,** 235–254 (1996).
32. Brawand, D. *et al.* The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513,** 375–381 (2014).
33. Miranda, G. E., Abrahan, C. E., Politi, L. E. & Rotstein, N. P. Sphingosine-1-phosphate is a key regulator of proliferation and differentiation in retina photoreceptors. *Invest. Ophthalmol. Vis. Sci.* **50,** 4416–4428 (2009).
34. Rotstein, N. P., Miranda, G. E., Abrahan, C. E. & German, O. L. Regulating survival and development in the retina: key roles for simple sphingolipids. *J. Lipid Res.* **51,** 1247–1262 (2010).
35. Khanna, H. *et al.* A common allele in RPGRIP1L is a modifier of retinal degeneration in ciliopathies. *Nat. Genet.* **41,** 739–745 (2009).
36. Yang, Z. *et al.* Mutations in the RPGR gene cause X-linked cone dystrophy. *Hum. Mol. Genet.* **11,** 605–611 (2002).

37. Gaston-Massuet, C. *et al.* Transcription factor 7-like 1 is involved in hypothalamo-pituitary axis development in mice and humans. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E548–57 (2016).

38. Behrens, J. *et al.* Functional interaction of beta-catenin with the transcription factor LEF-1. *Nature* **382,** 638–642 (1996).

39. Gribble, S. L., Kim, H.-S., Bonner, J., Wang, X. & Dorsky, R. I. Tcf3 inhibits spinal cord neurogenesis by regulating sox4a expression. *Development* **136,** 781–789 (2009).

40. Wen, W., Pillai-Kastoori, L., Wilson, S. G. & Morris, A. C. Sox4 regulates choroid fissure closure by limiting Hedgehog signaling during ocular morphogenesis. *Dev. Biol.* **399,** 139–153 (2015).

41. Bhattaram, P. *et al.* SOXC proteins amplify canonical WNT signaling to secure nonchondrocytic fates in skeletogenesis. *J. Cell Biol.* **207,** 657–671 (2014).

42. Agathocleous, M. *et al.* A directional Wnt/beta-catenin-Sox2-proneural pathway regulates the transition from proliferation to differentiation in the Xenopus retina. *Development* **136,** 3289–3299 (2009).

43. Heavner, W. E., Andoniadou, C. L. & Pevny, L. H. Establishment of the neurogenic boundary of the mouse retina requires cooperation of SOX2 and WNT signaling. *Neural Dev.* **9,** 27 (2014).

44. Twigg, S. R. F. *et al.* Frontorhiny, a distinctive presentation of frontonasal dysplasia caused by recessive mutations in the ALX3 homeobox gene. *Am. J. Hum. Genet.* **84,** 698–705 (2009).

45. Whited, J. L., Cassell, A., Brouillette, M. & Garrity, P. A. Dynactin is required to maintain nuclear position within postmitotic Drosophila photoreceptor neurons. *Development* **131,** 4677–4686 (2004).

46. Tsujikawa, M., Omori, Y., Biyanwila, J. & Malicki, J. Mechanism of positioning the cell nucleus in vertebrate photoreceptors. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 14819–14824

(2007).

47. Wu, L. *et al.* Miz1, a novel zinc finger transcription factor that interacts with Msx2 and enhances its affinity for DNA. *Mech. Dev.* **65,** 3–17 (1997).
48. Foerst-Potts, L. & Sadler, T. W. Disruption of Msx-1 and Msx-2 reveals roles for these genes in craniofacial, eye, and axial development. *Dev. Dyn.* **209,** 70–84 (1997).
49. Liu, B., Rooker, S. M. & Helms, J. A. Molecular control of facial morphology. *Semin. Cell Dev. Biol.* **21,** 309–313 (2010).
50. Brugmann, S. A. *et al.* Comparative gene expression analysis of avian embryonic facial structures reveals new candidates for human craniofacial disorders. *Hum. Mol. Genet.* **19,** 920–930 (2010).
51. Loh, Y.-H. E. *et al.* Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids. *Genome Biol.* **9,** R113 (2008).
52. Parsons, K. J., Trent Taylor, A., Powder, K. E. & Albertson, R. C. Wnt signalling underlies the evolution of new phenotypes and craniofacial variability in Lake Malawi cichlids. *Nat. Commun.* **5,** 3629 (2014).
53. Stamnes, M. A., Shieh, B. H., Chuman, L., Harris, G. L. & Zuker, C. S. The cyclophilin homolog ninaA is a tissue-specific integral membrane protein required for the proper synthesis of a subset of Drosophila rhodopsins. *Cell* **65,** 219–227 (1991).
54. Ferreira, P. A., Nakayama, T. A. & Travis, G. H. Interconversion of red opsin isoforms by the cyclophilin-related chaperone protein Ran-binding protein 2. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 1556–1561 (1997).
55. Perutz, M. F. & Brunori, M. Stereochemistry of cooperative effects in fish an amphibian haemoglobins. *Nature* **299,** 421–426 (1982).
56. Fago, A., Bendixen, E., Malte, H. & Weber, R. E. The anodic hemoglobin of Anguilla anguilla. Molecular basis for allosteric effects in a root-effect hemoglobin. *J. Biol. Chem.* **272,** 15628–15635 (1997).

57. Pelster, B. The generation of hyperbaric oxygen tensions in fish. *News Physiol. Sci.* **16,** 287–291 (2001).

58. Mazzarella, L. *et al.* Crystal structure of Trematomus newnesi haemoglobin re-opens the root effect question. *J. Mol. Biol.* **287,** 897–906 (1999).

59. Mazzarella, L. *et al.* Minimal structural requirements for root effect: crystal structure of the cathodic hemoglobin isolated from the antarctic fish Trematomus newnesi. *Proteins* **62,** 316–321 (2006).

60. Thom, C. S., Dickson, C. F., Gell, D. A. & Weiss, M. J. Hemoglobin variants: biochemical properties and clinical correlates. *Cold Spring Harb. Perspect. Med.* **3,** a011858 (2013).

61. Mandic, M., Todgham, A. E. & Richards, J. G. Mechanisms and evolution of hypoxia tolerance in fish. *Proc. Biol. Sci.* **276,** 735–744 (2009).

62. Darwall, W. R. T., Allison, E. H., Turner, G. F. & Irvine, K. Lake of flies, or lake of fish? A trophic model of Lake Malawi. *Ecol. Modell.* **221,** 713–727 (2010).

63. Gou, X. *et al.* Hypoxia adaptation and hemoglobin mutation in Tibetan chick embryo. *Sci. China C Life Sci.* **48,** 616–623 (2005).

64. Natarajan, C. *et al.* Convergent evolution of hemoglobin function in high-altitude Andean waterfowl involves limited parallelism at the molecular sequence level. *PLoS Genetics* **11,** e1005681 (2015).

65. Natarajan, C. *et al.* Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. *Science* **354,** 336–339 (2016).

66. Joyce, D. A. *et al.* Repeated colonization and hybridization in Lake Malawi cichlids. *Curr. Biol.* **21,** R108–9 (2011).

67. Genner, M. J. & Turner, G. F. Ancient hybridization and phenotypic novelty within Lake Malawi's cichlid fish radiation. *Mol. Biol. Evol.* **29,** 195–206 (2012).

68. Nichols, P. *et al.* Secondary contact seeds phenotypic novelty in cichlid fishes. *Proc.*

Biol. Sci. **282,** 20142272 (2015).

69. Parsons, K. J., Trent Taylor, A., Powder, K. E. & Albertson, R. C. Wnt signalling underlies the evolution of new phenotypes and craniofacial variability in Lake Malawi cichlids. *Nat. Commun.* **5,** 3629 (2014).
70. Snoeks, J. & Konings, A. *The Cichlid Diversity of Lake Malawi/Nyasa/Niassa: Identification, Distribution and Taxonomy*. (Cichlid Press, 2004).
71. Hofmann, C. M. *et al.* The eyes have it: regulatory and structural changes both underlie cichlid visual pigment diversity. *PLoS Biol.* **7,** e1000266 (2009).
72. Wang, F. Y., Tang, M. Y. & Yan, H. Y. A comparative study on the visual adaptations of four species of moray eel. *Vision Res.* **51,** 1099–1108 (2011).
73. Sivasundar, A. & Palumbi, S. R. Parallel amino acid replacements in the rhodopsins of the rockfishes (Sebastes spp.) associated with shifts in habitat depth. *J. Evol. Biol.* **23,** 1159–1169 (2010).
74. Ramachandran, R. D., Perumalsamy, V. & Hejtmancik, J. F. Autosomal recessive juvenile onset cataract associated with mutation in BFSP1. *Hum. Genet.* **121,** 475–482 (2007).
75. Blankenship, T. N., Hess, J. F. & FitzGerald, P. G. Development-and differentiation-dependent reorganization of intermediate filaments in fiber cells. *Invest. Ophthalmol. Vis. Sci.* **42,** 735–742 (2001).
76. Oka, M., Kudo, H., Sugama, N., Asami, Y. & Takehana, M. The function of filensin and phakinin in lens transparency. *Mol. Vis.* **14,** 815–822 (2008).
77. Jakobs, P. M. *et al.* Autosomal-dominant congenital cataract associated with a deletion mutation in the human beaded filament protein gene BFSP2. *Am. J. Hum. Genet.* **66,** 1432–1436 (2000).
78. Conley, Y. P. *et al.* A juvenile-onset, progressive cataract locus on chromosome 3q21-q22 is associated with a missense mutation in the beaded filament structural

protein-2. *Am. J. Hum. Genet.* **66,** 1426–1431 (2000).

79. Delahunt, P. B., Webster, M. A., Ma, L. & Werner, J. S. Long-term renormalization of chromatic mechanisms following cataract surgery. *Vis. Neurosci.* **21,** 301–307 (2004).
80. Marmor, M. F. Ophthalmology and art: simulation of Monet's cataracts and Degas' retinal disease. *Arch. Ophthalmol.* **124,** 1764–1769 (2006).
81. Rijli, F. M., De Lucchini, S., Ciliberto, G. & Barsacchi, G. A Zn-finger protein, Xfin, is expressed during cone differentiation in the retina of the frog Xenopus laevis. *Int. J. Dev. Biol.* **37,** 311–317 (1993).
82. Chao, R. *et al.* A male with unilateral microphthalmia reveals a role for TMX3 in eye development. *PLoS One* **5,** e10565 (2010).
83. Provencio, I. *et al.* A novel human opsin in the inner retina. *J. Neurosci.* **20,** 600–605 (2000).
84. Panda, S. *et al.* Melanopsin is required for non-image-forming photic responses in blind mice. *Science* **301,** 525–527 (2003).
85. Brown, T. M. *et al.* Melanopsin contributions to irradiance coding in the thalamo-cortical visual system. *PLoS Biol.* **8,** e1000558 (2010).
86. Zhao, H. *et al.* Fibroblast growth factor receptor signaling is essential for lens fiber cell differentiation. *Dev. Biol.* **318,** 276–288 (2008).
87. Gehring, W. J. The master control gene for morphogenesis and evolution of the eye. *Genes Cells* **1,** 11–15 (1996).
88. Gehring, W. J. & Ikeo, K. Pax 6: mastering eye morphogenesis and eye evolution. *Trends Genet.* **15,** 371–377 (1999).
89. Schluter, D. *The Ecology of Adaptive Radiation*. (OUP Oxford, 2000).
90. Ruxton, G. D., Speed, M. P. & Kelly, D. J. What, if anything, is the adaptive function of countershading? *Anim. Behav.* **68,** 445–451 (2004).
91. Endler, J. A. Signals, signal conditions, and the direction of evolution. *American*

Naturalist **139,** S125–S153 (1992).

92. Wolf, J. B. W. & Ellegren, H. Making sense of genomic islands of differentiation in light of speciation. *Nat. Rev. Genet.* (2016). doi:10.1038/nrg.2016.133
93. Baird, N. a. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3,** e3376 (2008).
94. Ogden, R. *et al.* Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Mol. Ecol.* **22,** 3112–3123 (2013).
95. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22,** 3124–3140 (2013).
96. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
97. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
98. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
99. Lischer, H. E. L. & Excoffier, L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28,** 298–299 (2012).
100. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11,** 94 (2010).
101. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27,** 3070–3071 (2011).
102. Foll, M. & Gaggiotti, O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180,** 977–993 (2008).

103. Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J. K. Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185,** 1411–1423 (2010).

104. Team, R. C. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. (2014).

105. Hancock, A. M. *et al.* Adaptations to climate-mediated selective pressures in humans. *PLoS Genet.* **7,** e1001375 (2011).

106. Bouton, N., Visser, J. D. & Barel, C. D. N. Correlating head shape with ecological variables in rock-dwelling haplochromines (Teleostei: Cichlidae) from Lake Victoria. *Biol. J. Linn. Soc. Lond.* **76,** 39–48 (2002).

107. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).

108. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33,** W116–20 (2005).

109. Conesa, A. & Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* **2008,** 619832 (2008).

110. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

111. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57,** 289–300 (1995).

Supplementary

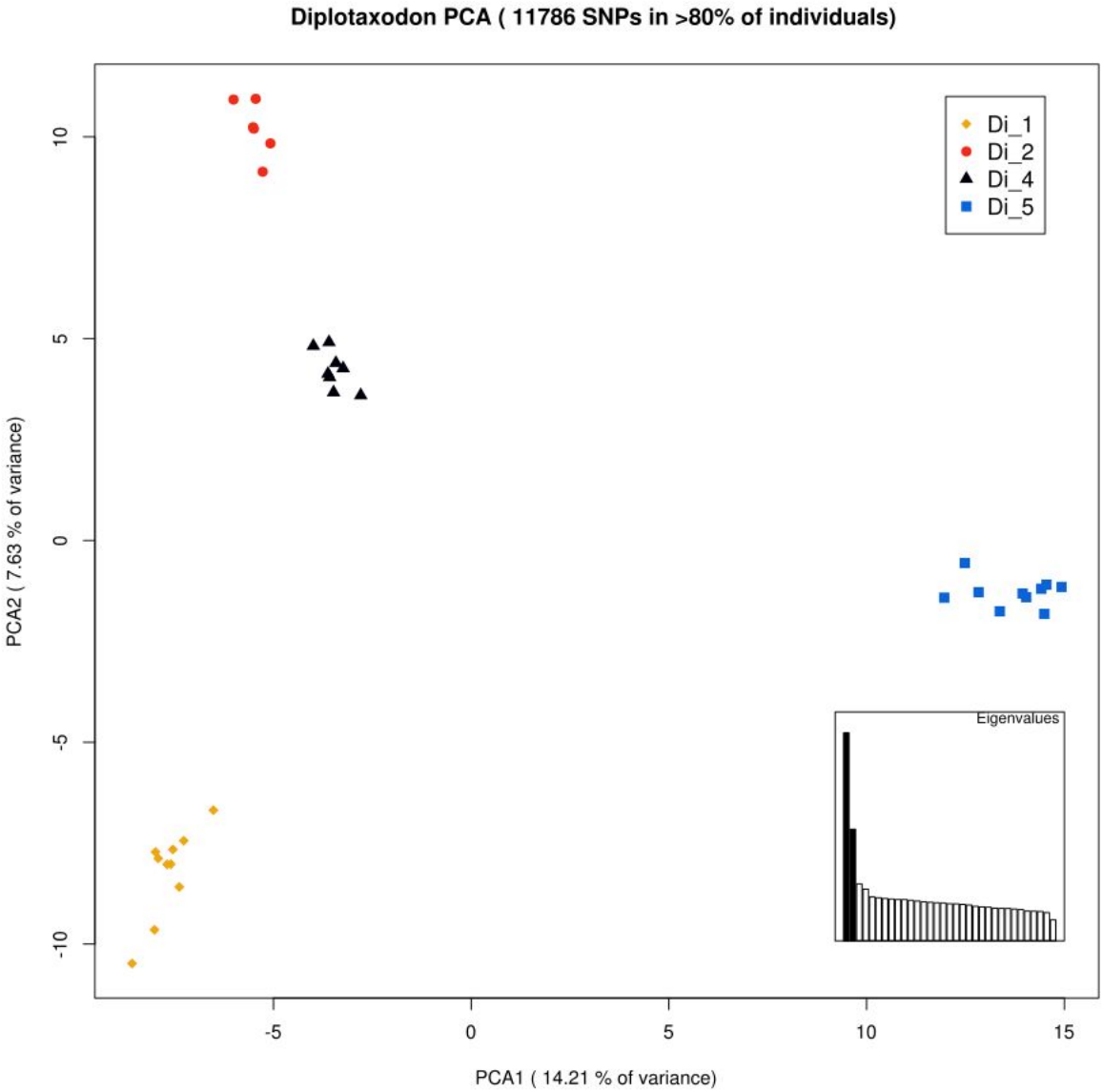

Figure S1. Principal component analysis based on 11,786 SNPs, *Diplotaxodon* species. Yellow - *D*. 'macrops black dorsal'; red - *D*. 'limnothrissa black pelvic'; black - *D*. 'macrops offshore'; blue - *D*. 'macrops ngulube'.

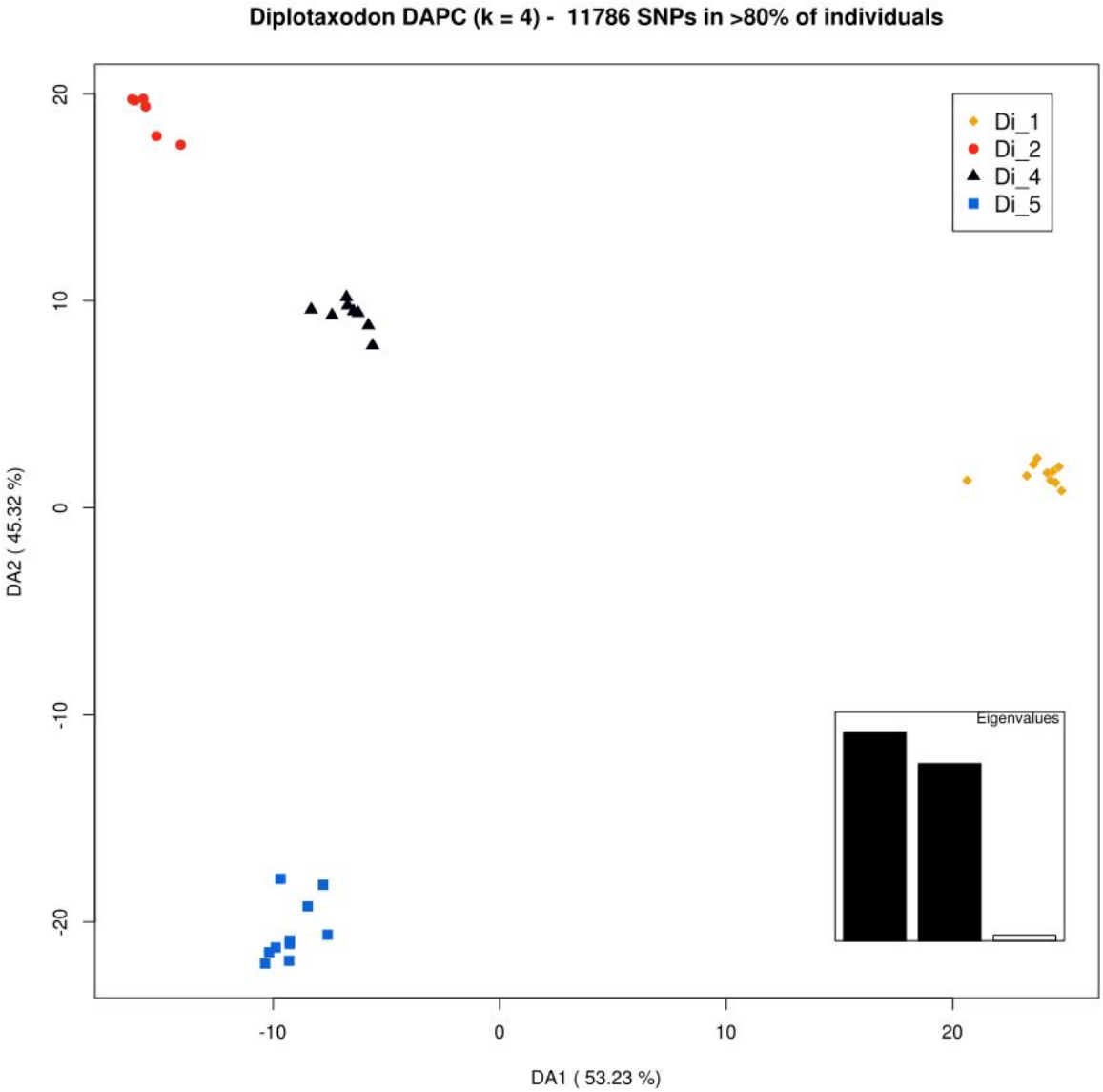

Figure S2. DAPC based on 11,786 SNPs, *Diplotaxodon* species. Yellow - *D*. 'macrops black dorsal'; red - *D*. 'limnothrissa black pelvic'; black - *D*. 'macrops offshore'; blue - *D*. 'macrops ngulube'.

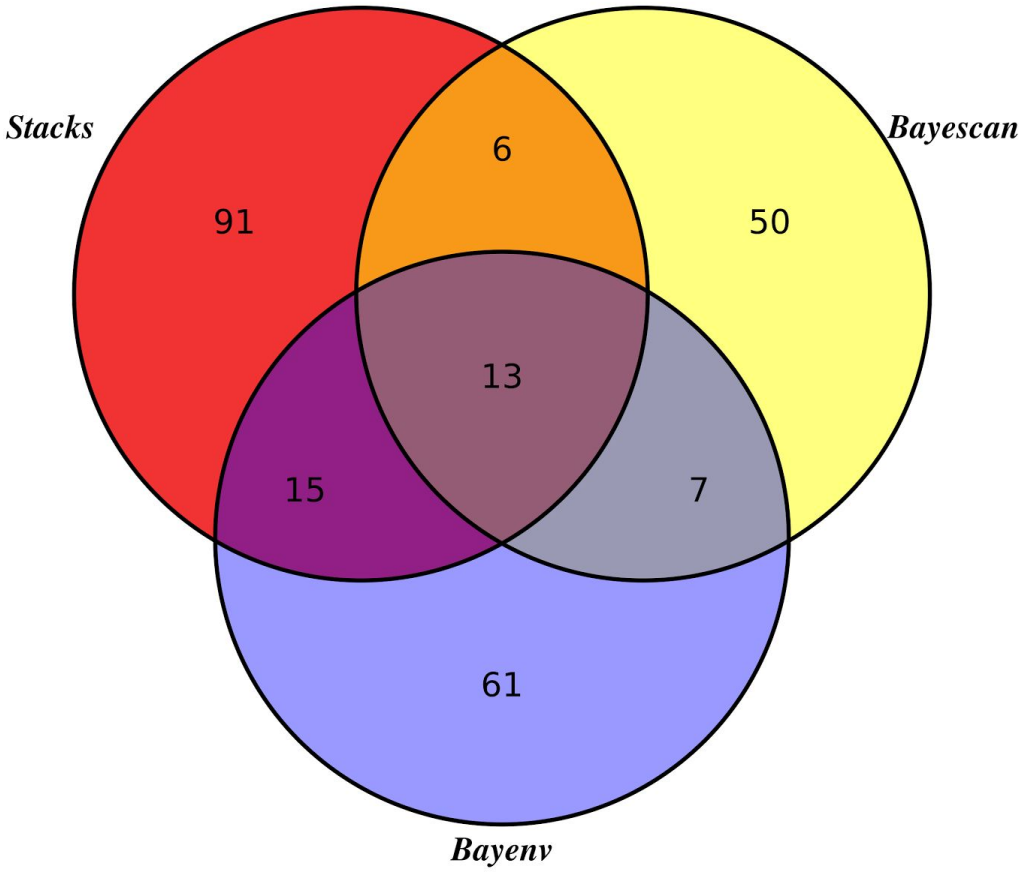

Figure S3. Number and concordance of candidate outlier loci highlighted by three independent approaches.

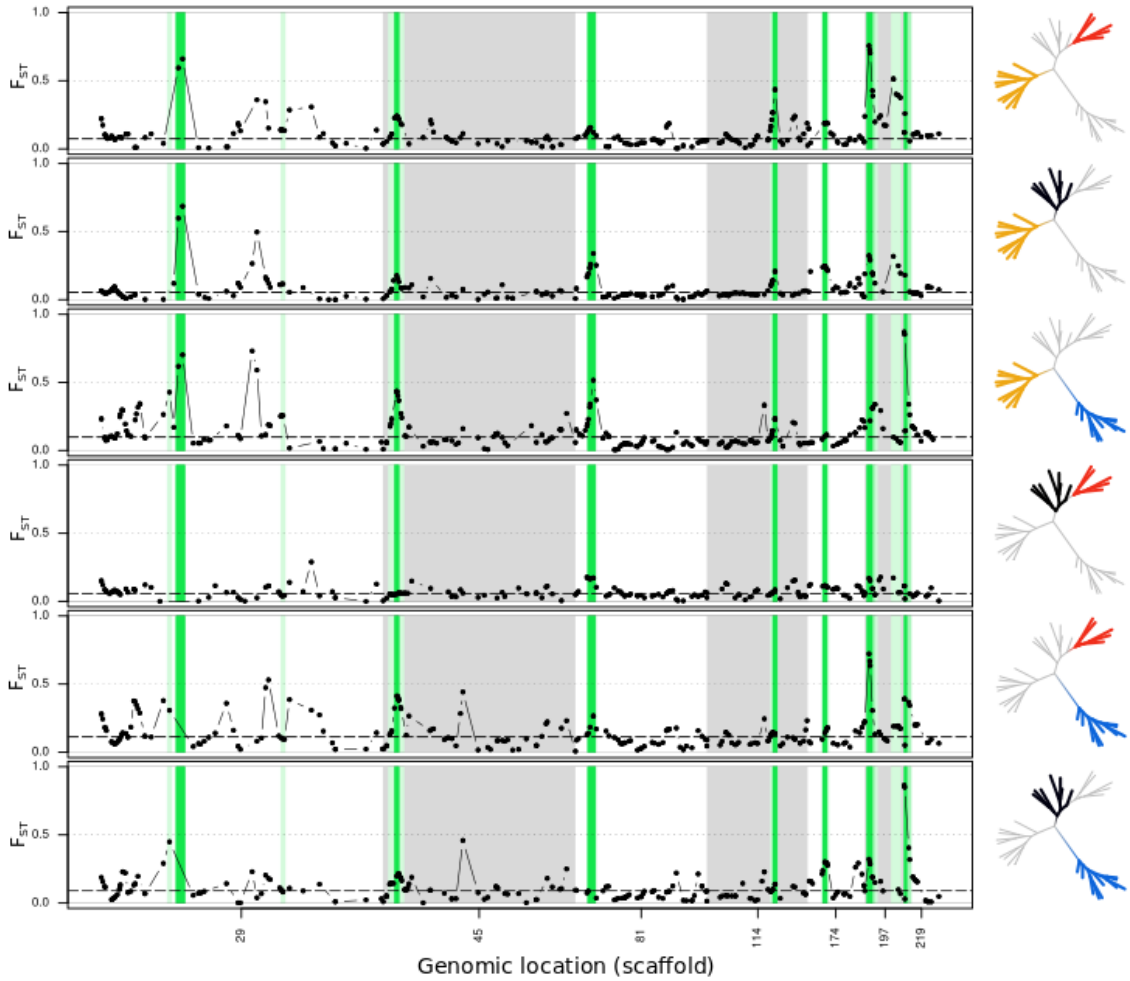

Figure S4. Pairwise Fst divergence at the six scaffolds (scaffold id on the x axis) containing loci highlighted as outliers by three independent detection approaches. Displayed are only scaffolds containing a minimum of 5 SNPs. Putative candidate regions are highlighted in shades of green. Dark green regions indicate support by three approaches (see tables 1 and S3 for summary of gene complements in highlighted regions). Dots represent the kernel smoothed averages across 50kb windows. Dashed lines indicate the genome wide Fst average. Population pairs are indicated by the highlighted regions in the phylogenetic trees on the right hand side, *Diplotaxodon* species. Yellow - *D*. 'macrops black dorsal'; red - *D*. 'limnothrissa black pelvic'; black - *D*. 'macrops offshore'; blue - *D*. 'macrops ngulube'.

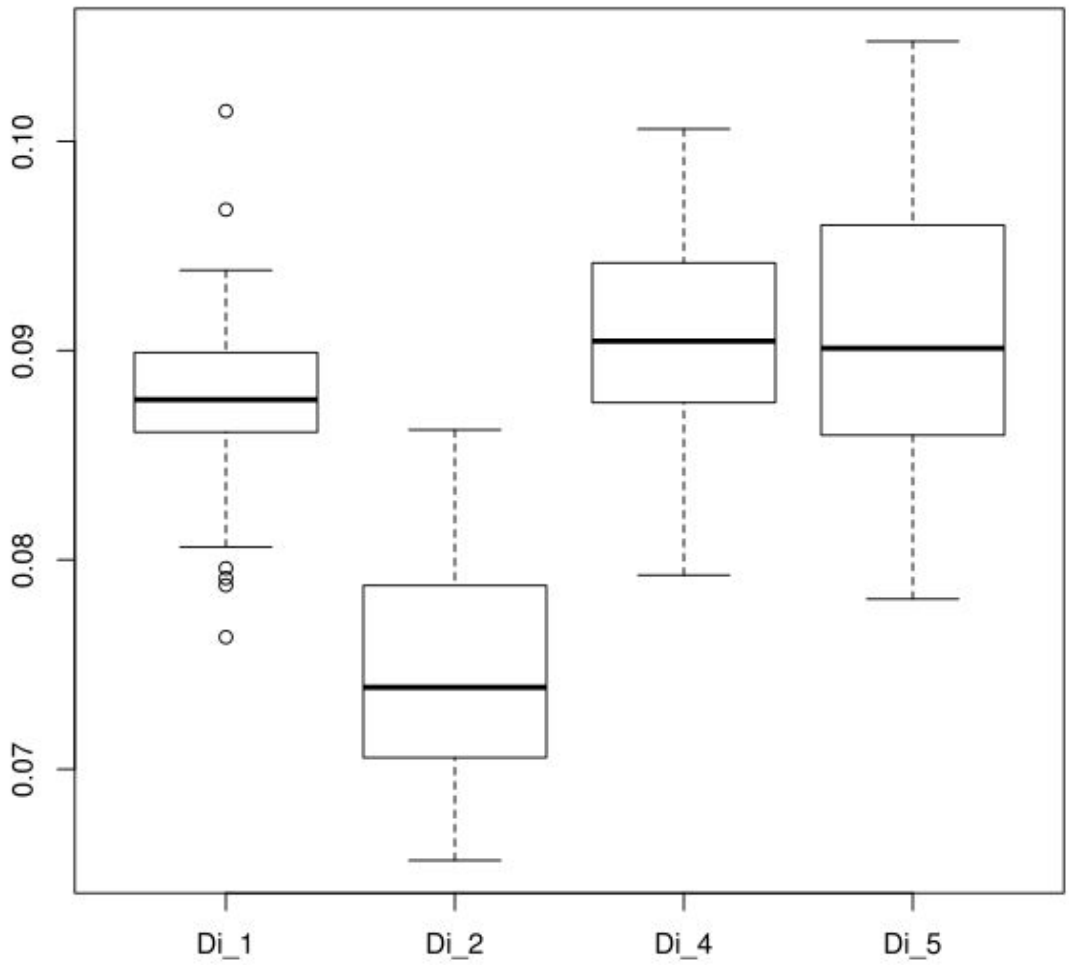

Figure S5. *Diplotaxodon* interspecific vertical eye diameter variation (normalized by total length of the fish).