Genome-enabled inference of functional genetic variants in the face, brain and behavior

Chinar Patil[1], Jonathan B. Sylvester[1,2], Kawther Abdilleh[1], Michael W. Norsworthy[3,4], Karen Pottin[5], Milan Malinsky[6,7], Ryan F. Bloomquist[1,8], Zachary V. Johnson[1], Patrick T. McGrath[1], Jeffrey T. Streelman[1]

Affiliations:
[1]School of Biological Sciences and Petit Institute of Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, GA, USA.
[2]Department of Biology, Georgia State University, Atlanta, GA, USA
[3]Catalog Technologies Inc., Boston, MA, USA.
[4]Freedom of Form Foundation, Inc., Cambridge, MA, USA.
[5]Laboratoire de Biologie du Développement (IBPS-LBD, UMR7622), Sorbonne Université, CNRS, Institut de Biologie Paris Seine, Paris, France.
[6]Zoological Institute, Department of Environmental Sciences, University of Basel, Basel, Switzerland.
[7]Wellcome Trust Sanger Institute, Cambridge, United Kingdom.
[8]Augusta University, Department of Oral Biology and Diagnostic Sciences, Department of Restorative Sciences, Dental College of Georgia, Augusta, GA, USA.

Abstract

Lake Malawi cichlid fishes exhibit extensive divergence in form and function among closely related species separated by a relatively small number of genetic changes. During the past million years, hundreds of species have diversified along an ecological axis in rock vs. sand habitats. We compared the genomes of rock- and sand-dwelling species and asked which genetic variants in which genes differed among the groups. We found that 96% of differentiated variants reside in non-coding sequence but these non-coding diverged variants are evolutionarily conserved. The majority of divergent coding variants are missense and/or loss of function. Regions near differentiated variants are enriched for craniofacial, neural and behavioral functional categories. To follow up experimentally, we used rock- vs. sand- species and their hybrids to (i) clarify the push-pull roles of BMP signaling and *irx1b* in the specification of forebrain territories during gastrulation and (ii) reveal striking context-dependent brain gene expression during adult social behavior. Our results suggest compelling ties between early brain development and adult behavior and highlight the promise of evolutionary reverse genetics – the identification of functional variants from genome sequencing in natural populations.

Introduction

Our understanding of how the genome encodes natural variation in form and function is still limited. This is the case for almost any trait, from height to behavior to complex disease (Boyle, Li et al. 2017). The reasons for this are manifold, but they include an underappreciated role of non-coding genetic variants linked to differences in traits. This is apparent in our assumptions and in syntheses of data. For instance, only 25 years ago, experts thought that the human genome might contain 70,000 to over 100,000 genes to account for our complexity (Fields, Adams et al. 1994). More recently, it has been estimated that upwards of 93% of human disease related variants – traits for which we have the most data from genome wide association studies (GWAS) – reside in noncoding DNA sequence (Maurano, Humbert et al. 2012). Many of these noncoding variants are regulatory, that is, they affect the expression of genes (Degner, Pai et al. 2012). Therefore, despite a refined understanding of how single genes work in controlled cellular environments, it remains unclear how the genome is activated to produce natural phenotypes, and this may be particularly vexing for context-dependent processes like development or behavior.

Over the past two decades, systems have been developed to identify the genetic basis of traits from nature(Streelman, Peichel et al. 2007). Amongst vertebrate animals, these traits include body armor (Colosimo, Peichel et al. 2004), color (Kratochwil, Liang et al. 2018), head and jaw shape (Albertson, Streelman et al. 2005, Shapiro, Kronenberg et al. 2013, Lamichhaney, Berglund et al. 2015), parental care (Okhovat, Berrio et al. 2015, Bendesky, Kwon et al. 2017), song (Pfenning, Hara et al. 2014) and coordinated movement (Greenwood, Wark et al. 2013). The take home message from this work has been that a small number of genes from recognizable pathways explain a considerable proportion of phenotypic variance. Yet, these studies may be biased in interpretation and limited in inference space. The focus is typically on one or two species and one trait at a time, often using hybrid pedigrees founded by a small number of individuals, and candidate gene or QTL approaches. Here, we explored a different strategy in a system of many species with many divergent traits. We sought to determine the genetic differences between closely related groups of species and then to focus experiments on

leads from genome divergence. In essence, we've asked the genome which traits to follow.

The Malawi cichlid system is an apposite one for our research aims. The assemblage comprises hundreds of closely related species that have diversified in the last 500,000 to one million years (Kocher 2004), such that the genomes of individuals across species boundaries remain highly similar (Loh, Katz et al. 2008). An appreciable fraction of genetic polymorphism identified in Malawi species is shared with cichlid lineages from throughout East Africa -- suggesting that ancient genetic variation fuels diversification of the Malawi flock (Loh, Bezault et al. 2013). Set against this background of genome similarity, Malawi cichlids exhibit staggering diversity in phenotypes including pigmentation (Streelman, Albertson et al. 2003), sex determination (Roberts, Ser et al. 2009, Parnell and Streelman 2013), craniofacial and brain patterning (Albertson, Streelman et al. 2005, Sylvester, Rich et al. 2010, Sylvester, Rich et al. 2013) and social behavior (York, Patil et al. 2018, Baran and Streelman 2020, Johnson, Moore et al. 2020). Previous work has focused on the genomic and early developmental underpinnings of this diversity, in rock- *vs*. sand-dwelling species (Loh, Katz et al. 2008, Fraser, Hulsey et al. 2009, Sylvester, Rich et al. 2010, Sylvester, Rich et al. 2013).

Rock- vs. sand- species form ecologically distinct groups similar to other ecotypes in well-known adaptive radiations (marine *vs*. freshwater sticklebacks; tree *vs*. ground finches and anoles) (Streelman and Danley 2003). The main difference in this case is that each of the rock- and sand- groups contains more than 200 species. Recent divergence, rapid speciation and meta-population dynamics synergistically lead to the broad sharing of polymorphism across the rock-sand speciation continuum (Loh, Bezault et al. 2013, Malinsky, Svardal et al. 2018). Malawi rock-dwellers tend to be strongly territorial and aggressive; they breed and feed at high density in complex rock-reef habitats. Most eat algae from the substratum with strongly reinforced jaws packed with teeth. Adult rock-dweller brains exhibit enlarged anterior components, telencephala and olfactory bulbs. Sand-dwellers are less site-specific and less aggressive. They often breed on communal leks where males build sand 'bowers' to attract females (McKaye, Louda et al. 1990).

Many capture small prey using acute vision and fast-moving gracile jaws; their brains and sensory apparatus are elaborated for more posterior structures optic tecta, thalamus and eyes (SI Figure 1). We aimed to understand evolutionary divergence between rock- and sand-dwelling lineages by identifying the number, type and spectrum of genetic variants that separate these groups.

To target this primary axis of evolutionary divergence in the Lake Malawi species assemblage (Streelman and Danley 2003), we compared whole genomes of one male individual each from 8 rock-dwelling and from 14 sand-dwelling species (SI Table 1), to an average of 25X coverage per individual. Species were chosen to represent the diversity present within each of the rock- and sand- groups (Figure 1A), in terms of body size, color pattern, ecology and phylogenetically defined lineages within the sand- species group (Malinsky, Svardal et al. 2018).

Results

The genomic signature of rock-sand divergence

We compared the genomes of 8 rock dwellers and 14 sand dwellers to uncover the genomic signature of rock- versus sand- evolutionary diversification. We aligned sequence data to a reference genome of nearly 1 gigabase (Conte, Joshi et al. 2019) and identified approximately 22 million Single Nucleotide Polymorphisms (SNPs) and 200,000 Insertion-Deletions (InDels). We calculated F_{ST} per variant, and averaged across 10kb windows, to quantify divergence between rock and sand species. We found that 0.06% of SNPs and 0.44% of InDels are alternately fixed between rock- and sand- groups. When these divergent variants and genome regions (2.5% FDR) are mapped to linkage groups (chromosomes), it is apparent that the signature of rock- vs. sand- divergence is distributed relatively evenly across the chromosomes (Figure 1B). Among fixed variants, 3.5% were found in coding regions and 96.5% were predicted to be non-coding; ~17% in intergenic regions, 38% in introns, 38% in flanking regions (within 25kb up- or downstream of a gene), and 3% in annotated UTRs. Rock vs. sand fixed coding variants were more likely to be missense/loss-of-function (72.6%) than silent (27.3%).

We next generated whole-genome alignments of five published cichlid reference genomes from across East Africa (Brawand, Wagner et al. 2014) and estimated an evolutionary conservation score for each nucleotide position. Akin to phylogenetic footprinting, this approach allows inference of function for regions that are slower to change than others due to the long-term effect of purifying selection. For both coding and non-coding portions of the genome, we found that rock-sand divergence correlates positively with evolutionary conservation scores (Figure 1C), suggesting that differentiated rock-sand variants, including many non-coding variants, are enriched for function.

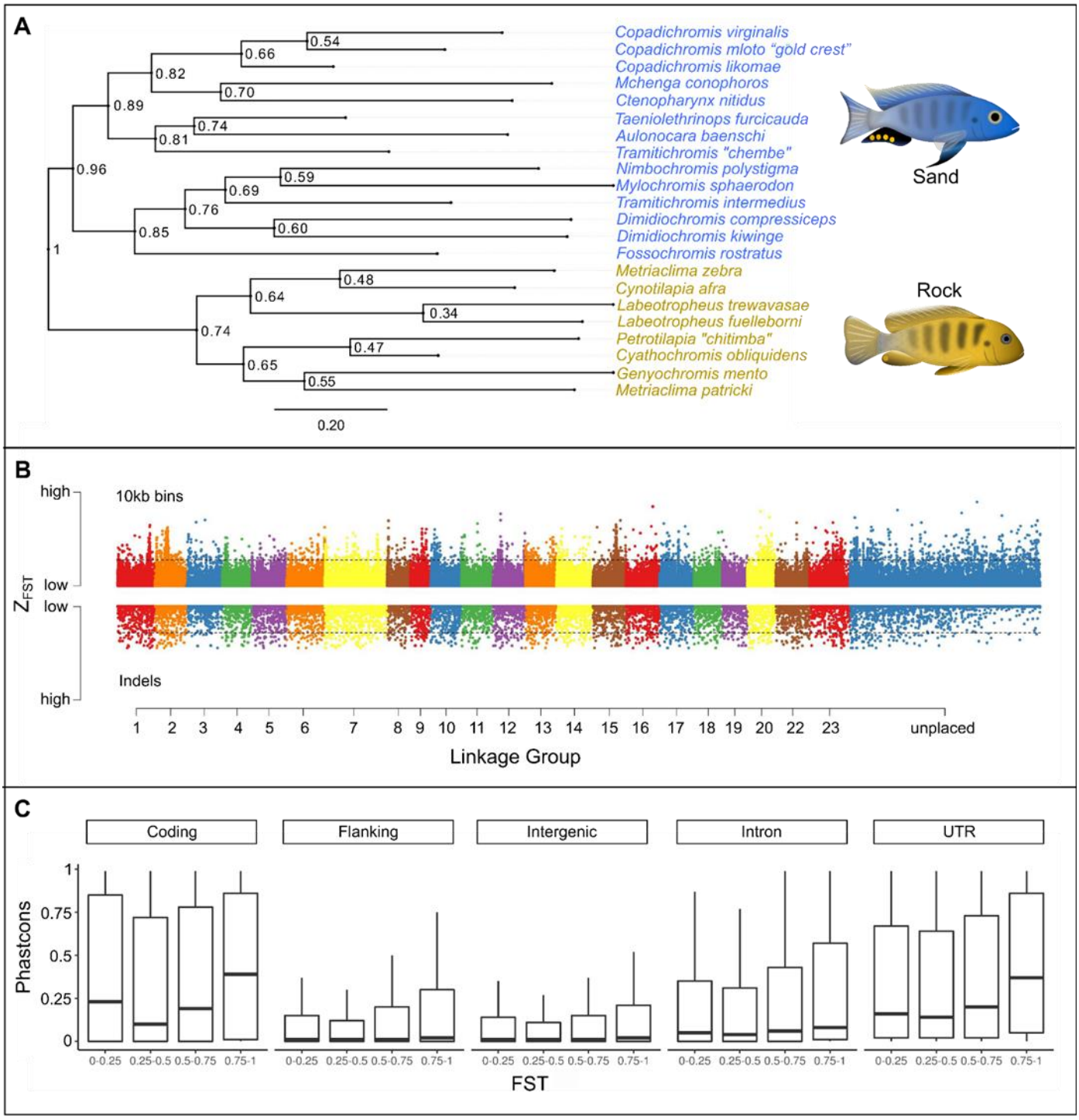


Figure 1: The genomic substrate for rock vs. sand evolution | (A) A maximum likelihood phylogeny of eight rock- and fourteen sand-dwelling species, based on variable sites (informative SNP and InDels) identified throughout the genome. “Sand” species contain representatives of the lineages “shallow benthic, deep benthic and utaka” from (Malinsky,

147 Svardal et al. 2018), while the “rock” species correspond to the “mbuna” lineage. (B) A
148 plot of Z-F_{ST} (F_{ST} normalized using Fisher’s Z-transformation) across the genome, plotting
149 genomic divergence between rock- vs. sand-dwelling groups. Single nucleotide
150 polymorphisms (SNPs) summed over 10kb bins and insertion-deletion mutations (InDels)
151 are shown on the same scale. Numbers along the x-axis refer to linkage groups (i.e.,
152 chromosomes) and threshold lines indicate 2.5% FDR. (C) Evolutionary conservation
153 (PhastCons) scores were calculated for each nucleotide across the genome, subdivided
154 by genome annotation and plotted by bins of increasing F_{ST}. The PhastCons score for
155 each genome category is significantly higher for increasing bins of F_{ST} (Wilcoxon rank
156 sum p value < 2e-16).
157

A total of 4,484 genes lie within 25 kb of either an alternately fixed variant or a highly divergent 10kb window (2.5%FDR). Pathway enrichment analysis (Ben-Ari Fuchs, Lieder et al. 2016) of human homologs/analogs for these genes reveals categories spanning early embryonic development, craniofacial morphogenesis, brain development, synaptic transmission and neuronal function (SI Table 2). In particular, rock-sand divergent genes are enriched for GO Biological Process terms 'telencephalon development' ($p < 1.7e-18$), 'adult behavior' ($p < 2e-14$), 'synaptic plasticity' ($p < 1.4e-12$), 'odontogenesis' ($p < 3.7e-11$), 'response to BMP' ($p < 3.2e-09$), 'gastrulation' ($p < 5.6e-06$), 'face morphogenesis' ($p < 8.9e-08$), 'neural crest cell differentiation' ($p < 4.3e-13$), and 'eye development' ($p < 1.3e-15$). Over-represented gene families included nuclear hormone receptors ($p < 3.0e-08$), HOXL subclass homeoboxes ($p < 1.4e-07$), TALE class homeoboxes ($p < 4.2e-04$) and Forkhead boxes ($p < 9.33-04$; for novel expression domains in cichlid *foxp2* see SI Figure 2). We observed enrichment for the mouse phenotypes 'abnormal cognition' ($p < 3.3e-15$), 'abnormal learning and memory' ($p < 2.6e-15$), 'abnormal craniofacial morphology' ($p < 4.9e-11$) and 'abnormal social/conspecific interaction' ($p < 7.1e-14$). We used the list of differentiated genes to query an Allen Brain Atlas dataset that reports gene expression in hundreds of brain regions (Hawrylycz, Lein et al. 2012). Rock- sand-divergent genes were enriched for the basomedial nucleus of the amygdala, a sub-region of the telencephalon ($p_{adj} = 0.001$) that regulates fear, anxiety, and physiological responses to territorial intruders in rodents (Adhikari, Lerner et al. 2015, Mesquita, Abreu et al. 2016), and has been linked to Social Anxiety Disorder in humans (Carvalho, Nóbrega et al. 2020). Finally, we queried databases of genes involved in human disease. Genes near divergent variants are significantly enriched for factors implicated in neurological disease like Autism Spectrum Disorder (SFARI (Abrahams, Arking et al. 2013), Fisher's exact test p value < 2e-16) and disorders related to the neural crest (Piñero, Bravo et al. 2017), (Fisher's exact test p value < 2e-16).

Given the prevalence of evolutionarily conserved, non-coding, divergent rock-sand variants and genome-wide enrichment for craniofacial and neural crest biology, we examined overlap with published datasets of mammalian neural crest and craniofacial enhancers (Rada-Iglesias, Bajpai et al. 2012, Attanasio, Nord et al. 2013). These data

allow us to identify craniofacial and cranial neural crest cell (CNCC) enhancers conserved between mammals and cichlids and fixed variants between rock and sand species within these conserved regulatory elements. A total of 275 craniofacial enhancer elements and 234 human CNCC enhancers are evolutionarily conserved between mammals and cichlids. We found divergent rock-sand mutations within the enhancer elements of key genes integral to CNCC specification and migration (SI Table 2). Notably, from both datasets, fixed rock-sand variants were found within the enhancer region of the gene *nr2f2*, a nuclear receptor and master neural crest regulator (Simoes-Costa and Bronner 2015). Rock-sand divergent variants were similarly located within craniofacial enhancers of three genes (*yap1*, *fat4*, *rere*) that function in the Hippo signaling pathway, as well as within enhancers of *irx3* and *axin2*. These data linking rock-sand fixed SNP/InDels to evolutionarily conserved, experimentally verified enhancers further underscore the importance of non-coding variation in the craniofacial evolution of rock- and sand-lineages (Roberts, Hu et al. 2011).

Genome-wide divergence between rock vs. sand Malawi cichlids involves a relatively small percentage of genetic variants. Divergent variants are (a) predominantly non-coding, (b) in long-term evolutionarily conserved loci (c) enriched for genes and pathways involved in embryonic development, brain development, brain function and behavior, and craniofacial morphogenesis. Given these strong patterns of enrichment, we used the experimental power of the Malawi cichlid system to interrogate features of early development and adult behavior that differ between rock- and sand- groups.

A gastrula-stage map of forebrain diversification

Rock- vs. sand-dwelling Malawi cichlids exhibit divergence in or near genes enriched for BMP signalling, gastrulation, eye and telencephalon development, as well as the TALE (Irx) gene family. To explore the developmental consequences of this differentiation, we investigated early forebrain specification in rock- and sand- embryos, building upon our previous studies and interest in *irx1b* and early brain development (Sylvester, Rich et al. 2010, Sylvester, Rich et al. 2013). During development, the complexity of the vertebrate

brain is first laid out in the neural plate, a single-cell thick sheet of cells that forms between non-neural ectoderm and the germ ring at gastrulation. Irx genes act as transcriptional repressors of BMP signal in gastrulation, and function to specify the neural plate (Cavodeassi, Modolell et al. 2001). BMPs, in turn, are protective of the anterior-most region of the neural plate, which will ultimately give rise to the telencephalon, and suppress the eye field (Bielen and Houart 2012). Given alternatively fixed variants in the *irx1b* gene, expected interactions between Irx and BMP signaling in the early embryo and known telencephalon vs. eye size differences between rock- vs. sand- species (Sylvester, Rich et al. 2010, Sylvester, Rich et al. 2013), we examined and quantified the early activity of *irx1b* and BMP in rock- vs. sand- embryos.

We used a custom device to orient and image cichlid embryos *in toto* at gastrula and neurula stages (White, Sylvester et al. 2015). In early gastrula (EG), *irx1b* (red) and BMP signal (green, PSMAD) delineate complementary dorsal and ventral domains of the embryo (Figure 2A). By mid-gastrula (MG), *irx1b* shows two expression domains, one in the posterior portion of the developing neural plate (np) and the second co-expressed with PSMAD activity around its anterior border (white arrowheads). By late gastrula (LG), the domains of *irx1b* expression and PSMAD activity sharpen around the leading edge of the neural plate but remain overlapping around the periphery. Notably, *irx1b* expression is expanded in the anterior domain of sand-dwellers (S) compared to rock-dwellers at EG and MG, and then defines the boundary of the neural plate earlier in sand-dwellers (S) in LG (arrowheads). As a consequence, BMP signal should have a longer-lasting influence on the neural plate in rock-dwelling species. Based upon manipulative experiments in zebrafish (Bielen and Houart 2012), this is predicted to result in a relatively larger presumptive telencephalon and smaller eye field.

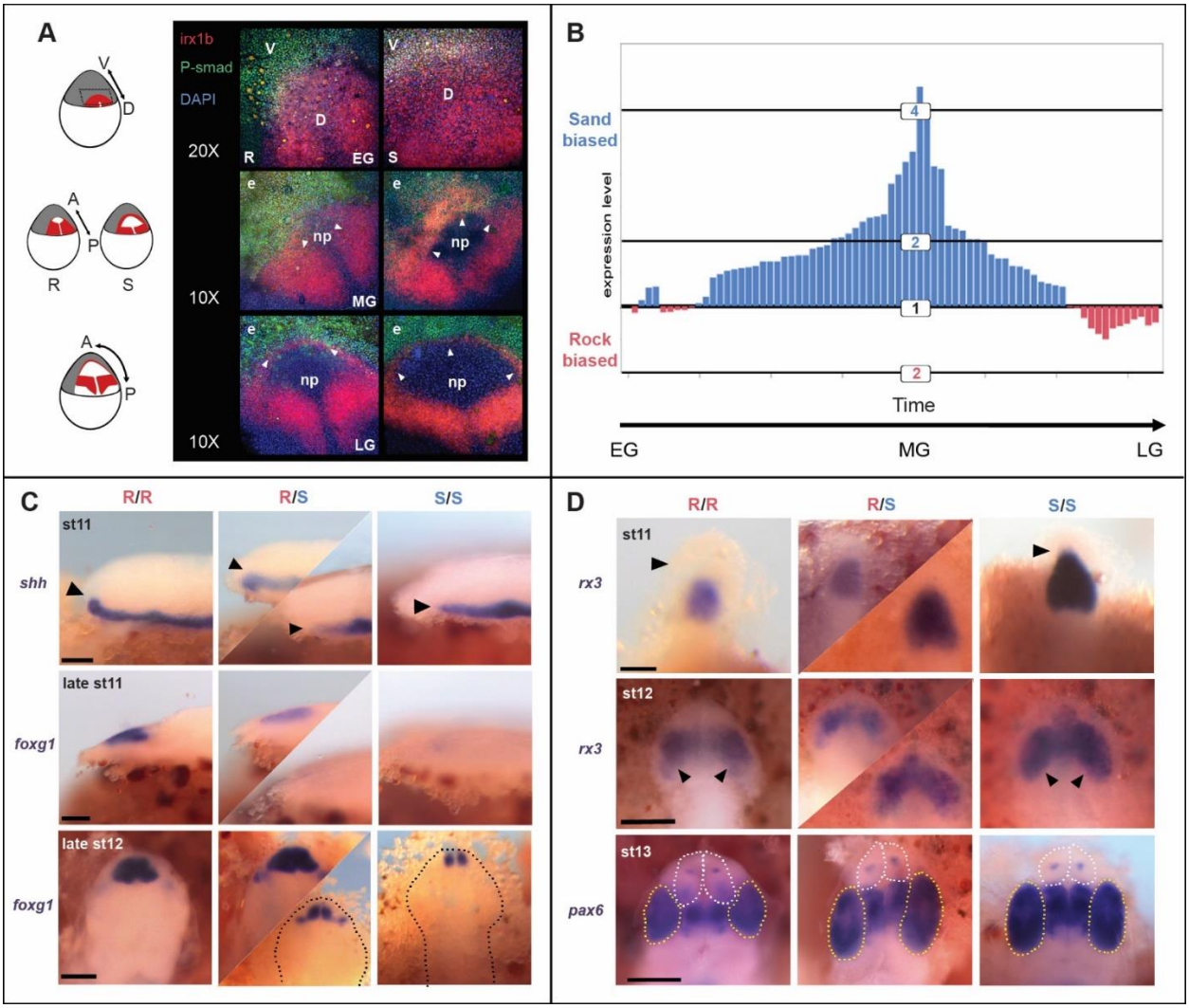


Figure 2: A gastrula-stage map of forebrain diversification | (A) Double in situ hybridization – immunohistochemistry to visualize *irx1b* expression (red) and BMP [PSMAD] activity (green) across the stages of cichlid gastrulation; DAPI in blue. Three rows represent early, mid and late gastrulation (EG, MG, LG) in embryos of rock- (R) and sand-dwelling (S) species. Sand-dwellers show expanded *irx1* expression in the dorsal portion of the embryo at EG, expanded *irx1b* expression in the anterior domain at MG (arrowheads) and clear PSMAD activity from the developing neural plate (np) earlier in LG (arrowheads). e=epidermis. Schematics at left show *irx1b* expression domains in red, on cartoons of cichlid embryos. (B) Relative expression of rock- (red) and sand- (blue) *irx1b* alleles, sampled from 74 heterozygous rock X sand F_2 embryos, across the stages of gastrulation. F_2 embryos were sampled at stage 9 (gastrulation). Because Malawi cichlid

species are maternal mouthbrooders and eggs are fertilized in batches per brood, each brood's embryos vary in timing of fertilization by up to 4h. Embryos within broods can therefore be sub-staged in gastrulation (see methods). Each bar on the plot represents the relative allelic expression of sand- and rock- *irx1b* in a heterozygous F_2 individual. Quantification of allele-specific expression (ASE) shows that levels are sand-biased, and that this effect is strongest in MG. (C) in situ hybridization of *shh* and *foxg1*, during neurula stages, showing development of the telencephalon in rock- X sand- F_2 embryos, indexed for *irx1b* genotype. F_2 individuals homozygous for rock- *irx1b* alleles (R/R) show a more dorsal progression of *shh* expression (black arrowheads), an earlier and a larger expression domain of the telencephalon marker *foxg1*. The top two rows are lateral views; bottom row is a dorsal view. Dotted lines demarcate the outline of the embryo in dorsal view. Heterozygous individuals exhibit greater variation in expression domains (middle columns), indicating that genetic factors other than variants in *irx1b* contribute to this phenotype. (D) in situ hybridization for *rx3* and *pax6*, during neurula stages, chart the development of the eye field in rock- X sand- F_2 embryos indexed for *irx1b* genotypes. F_2 individuals homozygous for sand- *irx1b* alleles (S/S) show larger domains of *rx3* (black arrowheads) and larger eyes (*pax6*, also marked by yellow dotted line), but smaller telencephala (white dotted line). All panels are dorsal views. Heterozygous individuals exhibit greater variation in expression domains (middle columns), indicating that genetic factors other than variants in *irx1b* contribute to this phenotype.

We developed a panel of rock- x sand- hybrid crosses to formally evaluate the role of *irx1b* in forebrain diversification. First, we used quantitative RT-PCR to measure allele-specific expression (ASE) in heterozygous rock- x sand- F_2 hybrids, across the stages of gastrulation. We observed that the sand- *irx1b* allele was expressed at significantly higher levels (average of 2.5-fold; p = 4.5e-13; Student's t-test) and that this difference was largely confined to MG (Figure 2B). Next, we used hybrid embryos to chart the development of the telencephalon and the eye field. Rock- x sand- F_2 hybrids, indexed for *irx1b* genotype, were raised to neurula and somitogenesis stages and we examined the expression of *shh* (which induces *foxg1* and the ventral forebrain), *foxg1* (a marker of the telencephalon), and *rx3* (a marker of the eye field), by in situ hybridization. F_2 individuals homozygous for rock- *irx1b* alleles exhibited a larger and more rostral domain of *shh* expression, an earlier and larger domain of *foxg1* and a smaller *rx3* domain (Figure 2C, D). These differences between rock- vs. sand- *irx1b* genotypes match expression divergence observed amongst rock- vs. sand- species (Sylvester, Rich et al. 2010, Sylvester, Rich et al. 2013). Finally, when we compared the relative size of the telencephalon among *irx1b* genotypes, individuals homozygous for rock- alleles exhibited larger telencephala (SI Figure 3). We conclude that genetic variants in and around the *irx1b* gene contribute to divergent specification of the Malawi cichlid forebrain, likely via spatial, temporal and quantitative variation in the expression of *irx1b* itself.

Our genome sequencing revealed a near-fixed InDel in the 3' UTR of the Malawi cichlid *irx1b* gene (SI Figure 4). Rock- species possess an 85bp insertion, compared to cichlid species from outside of the Malawi lineage. Sand-dwellers largely lack the insertion and exhibit a 6bp deletion compared to outgroups. The insertion shows strong genetic similarity to a fragment from the Rex1 family of non-LTR retrotransposons (Volff, Korting et al. 2000). Given the likelihood that *Astatotilapia calliptera* populations surrounding Lake Malawi may have seeded the Malawi evolutionary radiation and contributed to rock- and sand-dwelling lineages (Loh, Bezault et al. 2013, Malinsky, Svardal et al. 2018), we explored the presence/absence of this InDel in *Astatotilapia* samples. We found that most *Astatotilapia* individuals and populations had the rock- *irx1b* allele (the insertion), but that an individual from Chizumulu Island was fixed for the sand- allele and two individuals

sampled from Itupi were heterozygous. Because the 85bp insertion in rock- species is a partial Rex1 fragment, and sand- species carry a 6bp deletion compared to outgroups, we speculate that the current rock- and sand- divergent alleles were generated by at least two imperfect excision events of an element that invaded the genome of the Malawi + *Astatotilapia* ancestor. Rex1/Babar retrotransposons have been active in African cichlid genomes, and are known to influence gene expression when inserted in 5' and 3' UTRs (Brawand, Wagner et al. 2014). Future experiments will determine whether this Rex1 insertion causes the differences in *irx1b* gene expression and forebrain specification we observed.

Genomics of divergent social challenge and opportunity

Rock- and sand-dwelling Malawi cichlids live in strikingly different social and physical environments. Rock-dwelling males tend to be more aggressive than sand-dwellers (Baran and Streelman 2020) and defend territories year-round as sites for feeding and breeding. By contrast, sand-dwellers are more exploratory than rock-dwellers (Johnson, Moore et al. 2020) and only breeding males tend to be territorial, often building sand bowers to attract females and mitigate male-male aggression. Given these observations and genome-wide enrichment for categories related to adult behavior and social interaction, we designed an experimental paradigm to investigate brain gene expression profiles associated with divergent rock- vs. sand- social behaviors.

We evaluated social challenge and opportunity amongst males using a large tank with a 'rock' habitat at one end and 'sand' at the other, separated by glass bottom (Figure 3A). When parental rock- species are placed in this tank paradigm, males court females on the rock side of the tank. Males of sand- species court females over sand and construct species-appropriate bowers. When single hybrid rock- x sand- F_1 males are placed in this arena with hybrid F_1 females, males invariably court females over the 'rock' habitat. However, when two rock- x sand- hybrid F_1 males (brothers) were allowed to compete for gravid hybrid F_1 females in this tank paradigm, we observed something different. One male, typically the larger, courted females over the rock habitat, and the other

simultaneously constructed bowers to court females over the sand. We found no difference in gonadal-somatic index (GSI), an established biological metric of reproductive status and maturity, between F_1 males behaving as 'socially rock' vs. 'socially sand.' (SI Figure 5). Our observation of divergent behavior between F_1 brothers in the same tank suggests an interaction between the genome and the social environment.

We used RNA-seq to investigate gene expression profiles associated with behavior of rock- x sand- F_1 hybrid males that were actively courting females over rock vs. sand. Whole brains of F_1 males tested singly (n=2 lone) as well as F_1 brothers assayed in dyads (n=4 dyads) were collected during courtship, and interrogated by RNA-seq. Strikingly, gene expression profiles clustered not by fraternal relatedness, but rather by behavioral context (Figure 3B). Males from dyads that courted females over rocks had expression profiles similar to single males (who also courted over rocks) but distinct from their brothers that built bowers and courted females over sand in the same tank. Genes were considered significantly differentially expressed between 'social rock' and 'social sand' brains if they exhibited both a fold change ≥ 2 and crossed the threshold of $p_{adj} < 0.05$. Based on this criterion, we found 832 genes differentially expressed between rock- vs. sand-behaving males (Figure 3B, SI Table 4). Among differentially expressed genes, we observed significant functional enrichment for GO Biological Process categories 'synaptic signaling' ($p < 2.3e-21$), 'synaptic plasticity' ($p < 3.6e-09$), 'visual behavior' ($p < 2.09e-06$); mouse phenotypes 'abnormal learning/memory/conditioning' ($p < 5.9e-07$), 'abnormal telencephalon morphology' ($p < 3.95e-07$), 'abnormal spatial learning' ($p < 9.9e-07$) and pathways 'axon guidance' ($p < 3.3e-05$), 'oxytocin signaling' ($p < 5.2e-05$) and 'estrogen signaling' ($p < 1.8e-04$) (SI Table 4). Matches against the Allen Brain Atlas database of gene expression yielded enrichment for exclusively sub-regions of the telencephalon: CA3, hippocampus ($p_{adj} = 4.6e-6$), CA2, hippocampus ($p_{adj} = 7.3e-5$), CA4, hippocampus ($p_{adj} = 0.001$), claustrum ($p_{adj} = 0.001$), subiculum ($p_{adj} = 0.003$), dentate gyrus ($p_{adj} = 0.03$) and the basomedial nucleus of the amygdala ($p_{adj} = 0.03$). The hippocampus encodes episodic memory and spatial representations of the environment (Olton, Becker et al. 1979), and more recently its subregions have been shown to play critical roles in anxiety,

371 social interaction, and social memory formation (Hitti and Siegelbaum 2014, Zou, Chen
372 et al. 2016, Chiang, Huang et al. 2018). Roughly 38% of differentially expressed genes
373 also contained genetically differentiated SNP/InDels between rock- and sand- species (p-
374 value < 2e-6, Fisher's exact test), implying considerable cis-acting genetic variation.
375 Enrichment of categories related to brain function and synaptic plasticity showed greater
376 overlap than expected (Figure 3C). These context-dependent differences suggest rapid
377 and concerted changes in brain gene expression as males experienced and responded
378 to different social challenges and opportunities (O'Connell and Hofmann 2012, York, Patil
379 et al. 2018).
380

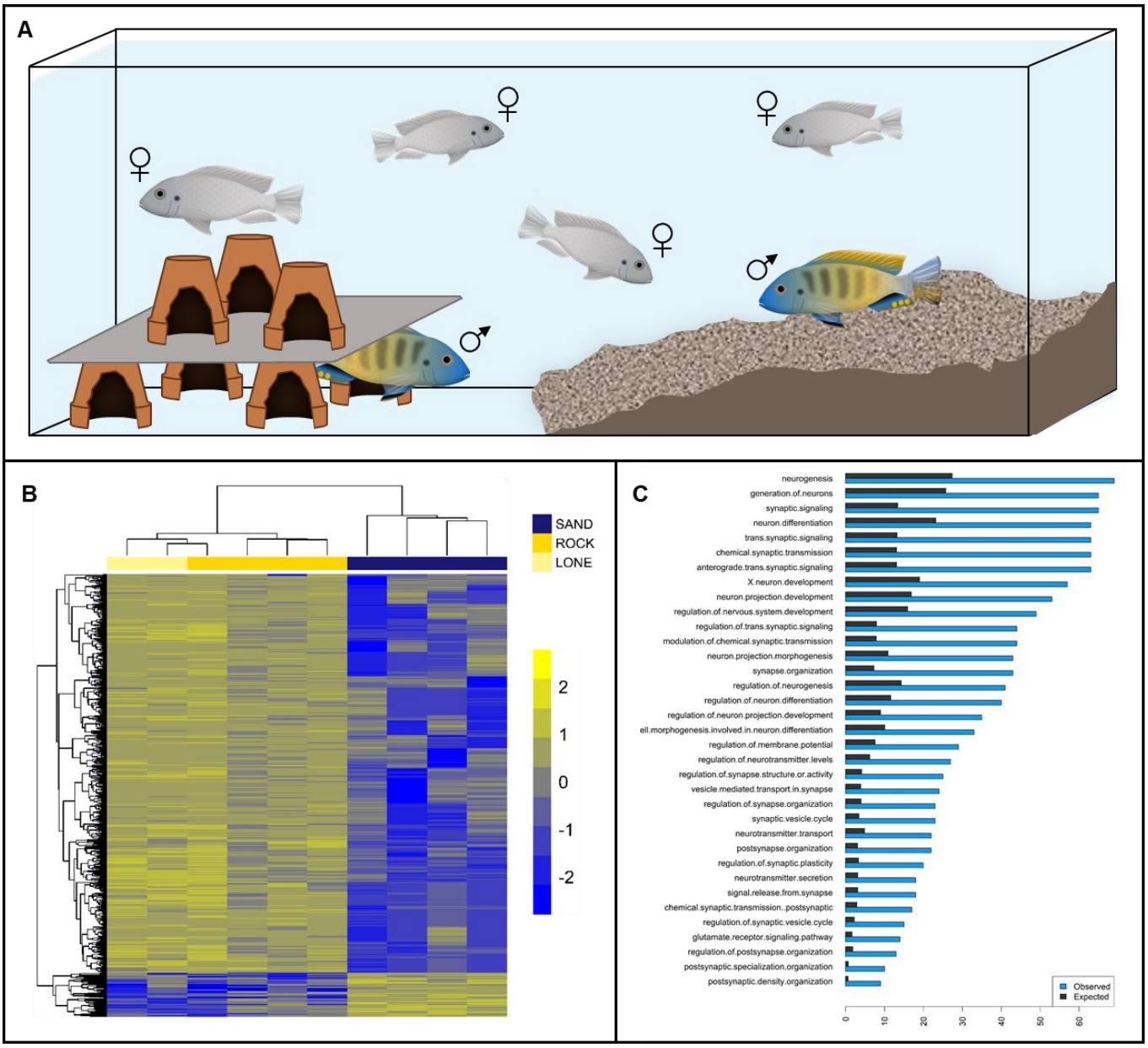


Figure 3: Genomics of divergent social context | (A) Schematic of the social context behavioral paradigm, in which rock-, sand- and rock- X sand- F_1 hybrids were evaluated. (B) Heatmap of genes differentially expressed in the brains of F_1 males behaving in either rock or sand social contexts. Each row in the heatmap is a gene, each column is an individual. Two F_1 males were not paired with other males, and courted females over the rock habitat (lone). All other F_1 males (n=8) were introduced to the testing arena in dyads. Notably, male brain gene expression clusters by social context and not fraternal relationships. (C) Gene Ontology (GO) Biological Process terms that show greater than

390 expected overlap between (i) genes differentially expressed in the brains of social rock-
391 vs. social sand- males and (ii) genes that are differentiated in the genomes of rock- vs
392 sand- species groups.

393

Discussion

Genome-enabled inference of evolutionary change in morphology and behaviour

A fundamental problem in evolutionary biology is understanding the cellular, developmental and genetic basis of how traits change. This is a challenge because we lack sufficient information about how genes work in outbred genomes from nature and we do not fully comprehend the causal role of noncoding variation in specifying form and function. This problem is especially difficult for traits that are only observed in particular contexts, like development and behaviour. To make progress, we and others have focused on study systems exhibiting abundant phenotypic diversity built from a relatively small number of genetic changes. Here we identify and characterize the genetic variants that demarcate one of the deepest evolutionary splits amongst Lake Malawi cichlid groups, that between rock- and sand-dwelling species thought to have diverged in the last one million years. We found a small percentage (less than 0.1%) of genetic variants to be differentiated between rock- and sand- groups, and that the majority of differentiated variants (>96%) were noncoding. Differentiated non-coding variants were more likely to be in an evolutionarily conserved locus as a function of genetic differentiation, suggesting that divergent rock- vs. sand- noncoding changes are functional. To support this idea, we identified alternately fixed rock- vs. sand- noncoding variants within experimentally verified, vertebrate-conserved craniofacial and cranial neural crest cell enhancers. The latter observation is similar in type to the discovery of human-specific deletions within mammal-conserved regulatory sequence (McLean, Reno et al. 2011).

Recently we surveyed genome-wide divergence between sand-dweller sub-groups that construct pit vs. castle bowers, sand-made structures to attract females for mating (York, Patil et al. 2018). Mapping those variants to the same genome reference, we expected distinct patterns of diversification because rock-sand and pit-castle divergence likely occurred at different times, along different trait axes, under the control of different evolutionary forces (Streelman and Danley 2003). Consistent with expectation, there is clear clustering of genome divergence on chromosome 11 for the pit-castle comparison (SI Figure 6), while all chromosomes carry the signature of rock-sand diversification

(Figure 1B). However, contrary to our expectation, rock- vs. sand- and pit- vs. castle-radiations have diverged in similar gene sets. Out of 3070 genes identified near 10kb high F_{ST} regions in the rock- vs. sand- comparison, 483 overlap with 1090 genes identified near high F_{ST} regions in the pit- vs. castle- comparison (p-value < 2e-9, Fisher's exact test, SI Table 3). This result may imply that evolutionary diversification in Lake Malawi is limited, or constrained, by chromosomal location.

Overall, genes in proximity to rock-sand divergent variants were enriched for functional categories related to early forebrain and craniofacial development, neuronal function and social behavior. This list of variants, coupled with consistent patterns of functional and pathway enrichment, motivated follow up experiments focused on early brain development and adult social behavior (Figure 4). It is apparent from our work here and previously (Sylvester, Rich et al. 2010, Sylvester, Rich et al. 2013), that Malawi cichlid brains and nervous systems begin to differ during gastrulation in pathways that can be predicted from divergent genome sequences. This is interesting for at least two reasons. First, this observation runs counter to the 'late equals large' textbook example (Finlay and Darlington 1995) of how brains evolve differences in relative proportions of their parts (Sylvester, Rich et al. 2010). Similarly, such early variation in development is not thought to be a driving force in evolution, precisely because early changes can have global and ramifying effects. Collectively, our findings provide a partial description of the conditions wherein variation during the earliest stages of development can contribute to evolutionary diversification. In each case we have examined, variation in gene expression is quantitative, heterochronic and limited to a precise stage or time period.

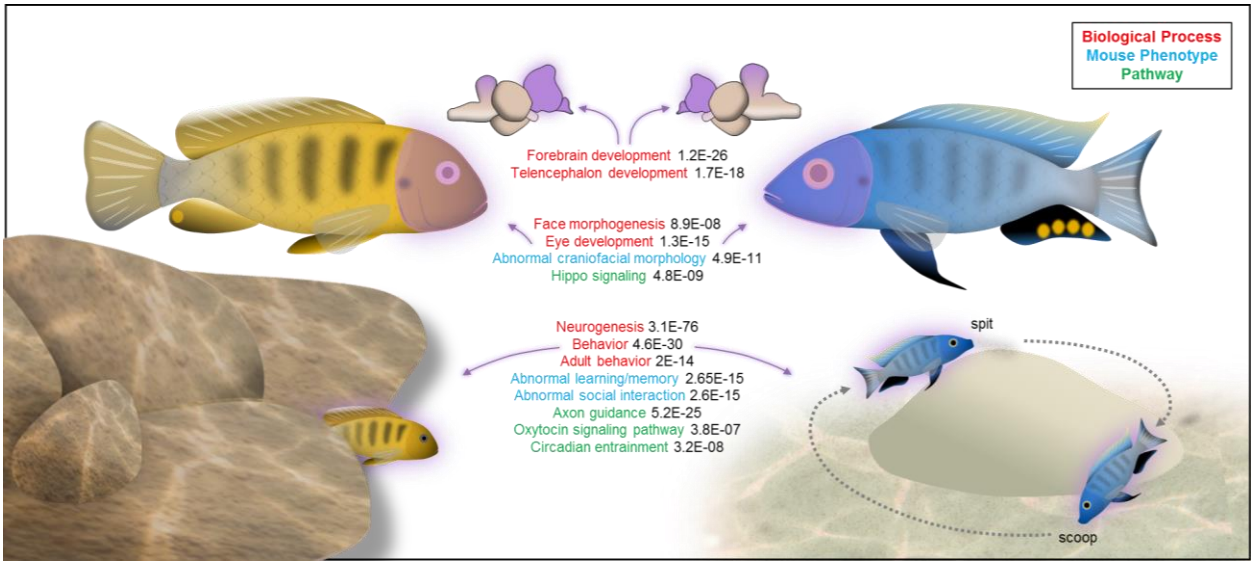


Figure 4: Genome-enabled discovery of evolutionary change in morphology and behavior. | Summary cartoon synthesizing significant enrichment categories that differentiate the genomes of rock- vs. sand-dwelling Malawi cichlids. Strong and consistent enrichment of craniofacial, neural and behavioral categories motivated follow on experiments in early brain development (Figure 2) and adult social behavior (Figure 3).

Sydney Brenner recognized the relationship between the genetic specification of nervous systems and the behavioral output of the brain (Brenner 1974). However, because these events take place so far apart in the lifespan of a vertebrate, they are rarely studied simultaneously. Here, the genome connects the two phenomena: rock- vs. sand-divergent gene sets indicating that both brain development and social behavior have been under divergent selection during the evolutionary diversification of these groups. To evaluate social behavior in rock- vs. sand-dwelling Malawi cichlids, we constructed a social context arena. The presence of sand and simulated rocky caves was sufficient to elicit species-appropriate male behavior when rock- or sand- males were tested with rock- or sand- gravid females. When rock- x sand- F_1 hybrid males were tested, one per tank, with F_1 hybrid females, males courted females in the rock quadrant of the tank. Notably, when dyads of F_1 brothers were tested in this tank paradigm with gravid F_1 females, we observed simultaneous 'social rock' and 'social sand' behavior. Brain gene expression profiles from behaving males clustered by social context (social rock vs social sand), and not by fraternal relationships. Differentially expressed genes were enriched for brain regions and pathways implicated in social interaction and overlapped significantly with rock- vs. sand- divergent genetic variants.

Social context is known to influence the brain. For instance, our clustering results are similar to those of Whitfield and colleagues (Whitfield et al. 2003) who showed that brain gene expression in honey bees was predictive of behavior. Likewise, changes in brain morphology and gene expression predictably accompany the ascent to dominance in the cichlid fish *Astatotilapia burtoni* (Fernald and Maruska 2012). Our data seem not to fit the model of dominant-subordinate however. In our experiments, the gonado-somatic index (GSI) did not differ between social rock vs. social sand brothers within dyads. Both males exhibited nuptial coloration, courted females and in cases with multiple gravid females, both brothers reproduced. Body size was associated with divergent social rock vs. social sand behavior of F_1 males; the social rock brother was always larger (mean mass was 26.96g ± 3.4 [SE] compared to 19.45g ± 2.2).

Our experiments demonstrate that F_1 hybrid male brains can express both social rock- and social sand- behavioral programs, and that social context determines which program is executed. This observation is similar to, but also different than, pit-digging x castle-building F_1 male Malawi cichlids who carry out parental bower behaviors in a specific sequence (York, Patil et al. 2018). Notably, in both cases, the hybrid males exhibit one of the two different parental behaviors at any one time – there is no intermediate behavior. In the pit- vs. castle- case, we think that the bower structure itself and/or a threshold signal from females might lock the hybrid male brain into a behavioral state. In the rock- vs sand-case here, it appears that other social cues (i.e., the presence and size of a rival male) lock the hybrid male brain into a behavioral state. These context-dependent behaviors, accompanied by changes in brain gene expression, are compelling examples of interaction between the genome and the social environment. The cellular and genetic basis of these behaviors and their plasticity deserves further attention. Our comparative genomic and brain gene expression data, combined with enrichment testing and experimental approaches, highlight that the Malawi cichlid telencephalon will be central to this future work.

Acknowledgements: This work was supported by grants from the NIH (R01GM101095, 2R01DE019637-10 to JTS and F32GM128346-01A1 to ZVJ) and the Human Frontiers Science Program (RGP0052/2019 to JTS). We would like to thank Shweta Biliya and the Genomics Core at Georgia Tech for invaluable assistance with the NGS sequencing. We also thank the members of the Streelman lab for comments on this manuscript.

Competing interests: The authors declare no competing interests.

Methods

Genome sequencing

We extracted genomic DNA, from fin clips (Qiagen DNeasy, Cat #69504), from 8 rock dwelling and 14 sand dwelling Lake Malawi species (SI Table 2). We made libraries using the Illumina Nextera Library prep kit and performed paired-end sequencing on the Illumina Hi-Seq 2500 at Georgia Tech. The *Metriaclima zebra* reference genome version MZ_UMD2a (Conte, Joshi et al. 2019) was used for genome alignment, variant discovery and annotation using standard BWA (Li and Durbin 2009) and GATK practices (Van der Auwera, Carneiro et al. 2013). The maximum likelihood tree in Figure 1A was constructed using SNPhylo (Lee, Guo et al. 2014), from variant data.

Genetically Divergent Regions

Vcftools (Danecek, Auton et al. 2011) was used to calculate F_{ST} (--weir-fst-pop) between the 8 rock and 14 sand species. Variants with F_{ST} = 1 were noted to be alternately fixed between rock and sand lineages in our dataset. F_{ST} was also measured across 10kb windows (--fst-window-size). Significance thresholds were marked using the fdrtool package in R. All variants were annotated using Snpeff 4.3i (Cingolani, Platts et al. 2012). We tested the genes within 25 kb of significantly differentiated variants for enrichment of functional categories. The cichlid gene names were converted to human analogs using Treefam based mapping (Ramakrishnan Varadarajan, Mopuri et al. 2018) and functional enrichment was determined using the TOPPFUN web-browser interface (Chen, Bardes et al. 2009).

PhastCons analysis

Pairwise alignments were generated using lastz v1.02(Harris 2007), with the following parameters: “B=2 C=0 E=150 H=0 K=4500 L=3000 M=254 O=600 Q=human_chimp.v2.q T=2 Y=15000”. This was followed by using USCS genome utilities (https://genome.ucsc.edu/util.html,

538 https://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/FOOTER) axtChain tool with -
539 minScore=5000. Additional tools with default parameters were then used following the
540 UCSC whole-genome alignment paradigm
541 (http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto) in order to
542 obtain a contiguous pairwise alignment. Multiple alignments were generated from
543 pairwise alignments with the multiz v11.2 (Blanchette, Kent et al. 2004) program, using
544 default parameters and the following pre-determined phylogenetic tree: ((((*M. zebra*, *P.*
545 *nyererei*), *A. burtoni*), *N. brichardi*), *O. niloticus*) in agreement with Brawand et al.
546 (Brawand, Wagner et al. 2014). Sequence conservation scores were then obtained using
547 PhastCons (Siepel, Bejerano et al. 2005) with a phylogenetic model estimated by the
548 phyloFit (Siepel and Haussler 2004) program, both from the PHAST software package
549 (v.1.3). The model fitting was done using default parameters. PhastCons was run in two
550 iterations, first to obtain the free parameters of the model (--estimate-trees and --no-post-
551 probs) and then using the output from this we ran PhastCons again to attain the
552 conservation scores with --target-coverage 0.3 --expected-length 100.
553

554 *Vertebrate-conserved enhancer elements*

555 A comparative genomic approach was used to identify putative craniofacial and neural
556 crest CNEs in mammals that segregate SNPs between rock-sand cichlid species.
557 Experimentally verified and published genome-wide craniofacial and neural crest
558 enhancers active during early embryonic stages that play a role in shaping the
559 development of neural crest and craniofacial structures in mammals were identified from
560 published literature (Rada-Iglesias, Bajpai et al. 2012, Attanasio, Nord et al. 2013). We
561 used the liftOver tool (Kent, Sugnet et al. 2002), which maps orthologous genomic regions
562 between species to convert genomic coordinates from one species to another. Using a
563 Human to *Oreochromis niloticus* to *Metriaclima zebra* mapping and a Mouse to
564 *Oreochromis niloticus* to *Metriaclima zebra* mapping, we identified the orthologous
565 genomic locations of the published craniofacial and neural crest enhancers in cichlids.
566 We designated any alternately fixed variant (variant with F_{ST} = 1) that was also within an
567 orthologous CNE as putatively involved in the rock-sand divergence (SI Table 2).

568

569 *Brain region enrichment analysis*

570 We identified 10,391 cichlid genes with human homologues and generated an expression
571 matrix for each gene across 250 human brain structures spanning telencephalon,
572 diencephalon, mesencephalon, and metencephalon using adult human brain microarray
573 data collected by the Allen Brain Institute (Hawrylycz, Lein et al. 2012). Cortical regions
574 and gyri for which fish do not have putative homologues were excluded from the analysis
575 (100/350, leaving 250 regions for subsequent analysis). The expression matrix was
576 generated using the get_expression function in the ABAEnrichment Bioconductor
577 package in R (Grote, Prufer et al. 2016). We then calculated the specificity of expression
578 for each gene in each of these brain regions using the specificity.index function in the pSI
579 package for R. This function calculates a matrix of gene expression specificity indices,
580 and corresponding p-values, as described previously (Dougherty, Schmidt et al. 2010,
581 Xu, Wells et al. 2014). We then tested whether 1) genes within 25kb of rock vs. sand
582 significantly differentiated variants (described above under “Genetically Divergent
583 Regions”), and 2) genes that were differentially expressed between rock- vs. sand-
584 behaving F1 hybrid males, were enriched for transcriptional markers of specific brain
585 regions using the *fisher.iteration* function with Benjamini-Hochberg correction, again
586 using the pSI package for R. For enrichment testing of differentially expressed genes, we
587 restricted analysis to genes that met the following criteria: 1) transcripts for the gene were
588 detected in all eight paired behaving males, and 2) at least 6 transcripts were detected in
589 each subject.

590

591 *Staging during gastrulation*

592 Cichlid gastrulation was split into three sub stages within the gastrula stage 9 (Murata,
593 Tamura et al. 2010). Gastrulation lasts 8 to 12 hours, depending on the species, and is
594 defined as after the shield (as described in zebrafish) stage until the presence of the first
595 somite at the beginning of neurula (stage 10). Embryos were classified as early gastrula
596 (EG) by an asymmetry in epiboly after shield stage until the formation of a ridge that is

597 analogous to the anterior neural ridge (ANR) in chick and mouse and the anterior neural
598 border (ANB) in zebrafish. At that point embryos were classified as mid gastrula (MG).
599 MG lasts until the formation of the dorsal-ventral axis, defined by further lengthening of
600 one side of the embryo, which begins to thicken as epiboly progresses. This is the dorsal
601 side of the embryo, and the side opposite the ANR is classified the ventral side of the
602 embryo. At this point the embryos are defined as late gastrula (LG). LG ends with the
603 specification of the neural plate, which appears as a portion of the dorsal embryo that is
604 raised relative to ventral side, usually in line with the ANR.

605

606 *Immunohistochemical staining*

607 Embryos were harvested at 24 hours post fertilization (hpf) from each of the rock -dwelling
608 cichlids *Metriaclima patricki* and *Metriaclima zebra* and the sand-dwelling cichlid
609 *Copadichromis borleyi* and *Tramitichromis intermedius*. The embryos were cultured until
610 they reached gastrula stage, approximately 36 to 40 hpf, then fixed at intervals throughout
611 gastrula until neurula. The embryos were then treated with auto-fluorescence reducer
612 (1.55mL 5M NaCl, 250ul Tris-HCl, pH 7.5, and 95mg NaBH4) overnight, and 10% 2-
613 mercaptoethanol for 1 hour. Next, whole mount *in situ* hybridization was done, using a
614 modification methods we published previously (Fraser, Bloomquist et al. 2008). *irx1b* was
615 visualized using Fast Red (naphthol chromogen, Roche Diagnostics), which fluoresces
616 at near red wavelengths (500-650 nm). After *in situ* hybridization, embryos were
617 immunostained for pSMAD 1,5,8 protein, using published protocols (Tucker, Mintzer et
618 al. 2008). Embryos were then bathed in Vectashield (Vector Labs) containing DAPI and
619 placed in a specially built mold(White, Sylvester et al. 2015) that accommodates the large
620 yolk and holds the embryo upright. Embryos were then scanned using a Zeiss LSM 700-
621 405 confocal microscope and processed using LSM 700 software and Image J.
622

623 *Rock-Sand hybridization and genotyping*

624 Two rock-sand crosses, one between *Copadichromis borleyi* (CB, sand-dweller sire) and
625 *Metriaclima zebra* (MZ, rock-dweller dam) and another between *Mchenga conophoros*

(MC, sand- sire) and *Petrotilapia sp.* 'thick bar' (PT, rock- dam), were artificially generated by taking the eggs from the dam just prior to spawning and mixing with sperm from the sire. The resultant F_1 were grown in tanks and allowed to spawn normally to generate F_2. Several F_2 broods were taken from multiple F_1 females for each cross, a total of 355 individuals for the CB x MZ cross and 608 for the MC x PT cross. The embryos were fixed at every stage starting at gastrula (stage 9) until early pharyngula (stage 14). The F_2 embryos were RNA-extracted at stage 9. DNA extraction was performed by fixing the embryos (stage 11-14) in 70% ethanol, then removing the tail from each individual and extracting the DNA using an extraction kit (Qiagen). Following extraction, the F_2 embryos were genotyped using custom probes (CAAATCTCCC[C/T]CCGCGGC, Taqman custom probes, Invitrogen) designed to identify a SNP in *irx1b* using RT-PCR. A subset of the embryos was also sequenced at a 900 bp interval around the *irx1b* SNP to verify the custom probes.

Quantitative F_2 Analysis

We quantified *irx1b* in F_2 at stage 9 and separated by genotypic class. The 74 heterozygous rock X sand F_2 embryos were dissected to remove most of the yolk and the total RNA was extracted from each individual using an RNA Extraction Kit (Qiagen). The amount of mRNA specific to each allele of *irx1b* was quantified by using the RNA-to-Ct kit (Invitrogen) and the custom probes. The delta Ct for each heterozygote was generated with the equation, 2^(allele from dam – allele from sire). We tested the data with an ANOVA, followed by a Tukey's multiple comparison test to determine significance between genotype classes.

Forebrain and eye measurements

The forebrain and eyes were measured by integrating the area of transverse sections in embryos of rock- and sand-dweller cichlid species, using previously published methods (Sylvester, Rich et al. 2013). The rock-dweller species included *Cynotilapia afra* (CA, planktivore), *Labeotropheus fuelleborni* (LF, algivore) and *Metriaclima zebra* (MZ,

generalist); sand-dweller species included *Aulonocara jacobfreibergi* (AJ, 'sonar' hunter), *Copadichromis borleyi* (CB, planktivore) and *Mchenga conophoros* (MC, insectivore/generalist). Embryos from each species, as well as the F_2 individuals, were measured starting from the earliest the telencephalon can be differentiated from the forebrain (mid-somitogenesis, stage 12) and at each subsequent stage until the forebrain has defined prosomeres (early pharyngula, stage 14) (Sylvester, Rich et al. 2010). To keep measurements standardized across stages, all measurements were defined by forebrain morphology at the earliest timepoint (stage 12). The 'eye' measurement remains consistent at all stages, the 'anterior' measurement includes the telencephalon and presumptive olfactory bulb, and the 'posterior' measurement includes the diencephalon and each of its constitutive prosomeres (dorsal and ventral thalamus and hypothalamus). To facilitate measurements, we used gene expression of *rx3* (for stage 12 embryos) and *pax6* (stage 13 and 14) to identify the different structures of the forebrain and eye.

RNA Extraction and Sequencing, Adult Social Behavior

Adult F_1 hybrid males (Supplementary table 4) were introduced to an assay tank containing females of the same cross and simulated rock habitat on one side and simulated sand habitat on the other side separated by empty tank space (Figure 3A). Male brains were harvested within 20 minutes of exhibiting territoriality and displays for females by rapid decapitation and whole brains were immediately stored in RNAlater (Thermo Fisher Cat# AM7020).

Tissues were frozen in liquid nitrogen, homogenized using a mortar and pestle and placed in trizol. Following standard chloroform extraction, RNeasy mini columns (Qiagen Cat No./ID: 74104) were utilized to purify RNA for sequencing. Total RNA was quantified using Qubit (Molecular Probes) and quality analyzed using the Agilent 2100 Bioanalyzer System for RNA library preparation. RNA input was normalized to 1µg and libraries were prepared using the TruSeq Stranded mRNA Sample Prep Kit (Illumina- Kit A). Libraries were again quantified, quality assessed, and normalized for sequencing on the HiSeq 2500 Illumina Sequencing System (Georgia Tech Genomics Core, standard practices).

Experimental design and raw files can be accessed on the NCBI Gene Expression Omnibus database under the accession number GSE122500.

Differential Gene Expression Analysis

Raw sequence reads from whole brain transcriptomes were quality controlled using the NGS QC Toolkit (Patel and Jain 2012). Raw reads with an average PHRED quality score below 20 were filtered out. Filtered reads were also trimmed of low-quality bases at the 3' end. High quality sequence reads were aligned to the *M.zebra* reference genome MZ_UMD2a(Conte, Joshi et al. 2019) using TopHat v2.0.9 (Trapnell, Pachter et al. 2009). On average, across all samples, over 95% of reads mapped to the reference genome. The resulting TopHat2 output bam files were sorted and converted to sam files using samtools v0.19 (Li, Handsaker et al. 2009). Sorted sam files were used as input for the HTSeq-count v0.6.1 program to obtain fragment counts for each locus (Anders, Pyl et al. 2015). Fragment counts were scale-normalized across all samples using the calcNormFactors function in the edgeR package v3.6.8 (Robinson, McCarthy et al. 2010). Relative consistency among replicates and samples was determined via the Multidimensional scaling (MDS) feature within the edgeR package in R. The native R function *hclust(dist)* used to cluster samples. Scale-normalized fragment counts were converted into $\log_2$ counts per million reads mapped (cpm) with precision weights using voom and fit to a linear model using limma v3.20.9 (Ritchie, Phipson et al. 2015). Pairwise contrasts were constructed between socially rock and socially sand samples. After correcting for multiple comparisons using the Benjamini-Hochberg method (Hochberg and Benjamini 1990), genes were considered differentially expressed between socially rock and socially sand samples if they exhibited both a fold change ≥ 2 and $P_{adj} < 0.05$. Using Treefam based mapping (Ramakrishnan Varadarajan, Mopuri et al. 2018) the cichlid gene names were converted to human analogs and functional enrichment was determined using the TOPPFUN web-browser (Chen, Bardes et al. 2009).

References

Abrahams, B. S., D. E. Arking, D. B. Campbell, H. C. Mefford, E. M. Morrow, L. A. Weiss, I. Menashe, T. Wadkins, S. Banerjee-Basu and A. Packer (2013). "SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs)." Mol Autism **4**(1): 36.

Adhikari, A., T. N. Lerner, J. Finkelstein, S. Pak, J. H. Jennings, T. J. Davidson, E. Ferenczi, L. A. Gunaydin, J. J. Mirzabekov, L. Ye, S.-Y. Kim, A. Lei and K. Deisseroth (2015). "Basomedial amygdala mediates top-down control of anxiety and fear." Nature **527**(7577): 179-185.

Albertson, R. C., J. T. Streelman, T. D. Kocher and P. C. Yelick (2005). "Integration and evolution of the cichlid mandible: The molecular basis of alternate feeding strategies." Proceedings of the National Academy of Sciences of the United States of America **102**(45): 16287-16292.

Anders, S., P. T. Pyl and W. Huber (2015). "HTSeq--a Python framework to work with high-throughput sequencing data." Bioinformatics **31**(2): 166-169.

Attanasio, C., A. S. Nord, Y. W. Zhu, M. J. Blow, Z. R. Li, D. K. Liberton, H. Morrison, I. Plajzer-Frick, A. Holt, R. Hosseini, S. Phouanenavong, J. A. Akiyama, M. Shoukry, V. Afzal, E. M. Rubin, D. R. FitzPatrick, B. Ren, B. Hallgrimsson, L. A. Pennacchio and A. Visel (2013). "Fine Tuning of Craniofacial Morphology by Distant-Acting Enhancers." Science **342**(6157).

Baran, N. M. and J. T. Streelman (2020). "Ecotype differences in aggression, neural activity and behaviorally relevant gene expression in cichlid fish." Genes Brain Behav **19**(6): e12657.

Ben-Ari Fuchs, S., I. Lieder, G. Stelzer, Y. Mazor, E. Buzhor, S. Kaplan, Y. Bogoch, I. Plaschkes, A. Shitrit, N. Rappaport, A. Kohn, R. Edgar, L. Shenhav, M. Safran, D. Lancet, Y. Guan-Golan, D. Warshawsky and R. Shtrichman (2016). "GeneAnalytics: An Integrative Gene Set Analysis Tool for Next Generation Sequencing, RNAseq and Microarray Data." Omics-a Journal of Integrative Biology **20**(3): 139-151.

Bendesky, A., Y.-M. Kwon, J.-M. Lassance, C. L. Lewarch, S. Yao, B. K. Peterson, M. X. He, C. Dulac and H. E. Hoekstra (2017). "The genetic basis of parental care evolution in monogamous mice." Nature **544**(7651): 434.

Bielen, H. and C. Houart (2012). "BMP signaling protects telencephalic fate by repressing eye identity and its Cxcr4-dependent morphogenesis." Developmental cell **23**(4): 812-822.

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson and E. D. Green (2004). "Aligning multiple genomic sequences with the threaded blockset aligner." Genome research **14**(4): 708-715.

744 Bloomquist, R. F., T. E. Fowler, J. B. Sylvester, R. J. Miro and J. T. Streelman (2017). "A
745 compendium of developmental gene expression in Lake Malawi cichlid fishes." BMC Dev Biol
746 **17**(1): 3.

747 Bonkowsky, J. L., X. Wang, E. Fujimoto, J. E. Lee, C. B. Chien and R. I. Dorsky (2008). "Domain-
748 specific regulation of foxP2 CNS expression by lef1." BMC Dev Biol **8**: 103.

749 Boyle, E. A., Y. I. Li and J. K. Pritchard (2017). "An Expanded View of Complex Traits: From
750 Polygenic to Omnigenic." Cell **169**(7): 1177-1186.

751 Brawand, D., C. E. Wagner, Y. I. Li, M. Malinsky, I. Keller, S. Fan, O. Simakov, A. Y. Ng, Z. W.
752 Lim, E. Bezault, J. Turner-Maier, J. Johnson, R. Alcazar, H. J. Noh, P. Russell, B. Aken, J. Alfoldi,
753 C. Amemiya, N. Azzouzi, J. F. Baroiller, F. Barloy-Hubler, A. Berlin, R. Bloomquist, K. L.
754 Carleton, M. A. Conte, H. D'Cotta, O. Eshel, L. Gaffney, F. Galibert, H. F. Gante, S. Gnerre, L.
755 Greuter, R. Guyon, N. S. Haddad, W. Haerty, R. M. Harris, H. A. Hofmann, T. Hourlier, G. Hulata,
756 D. B. Jaffe, M. Lara, A. P. Lee, I. MacCallum, S. Mwaiko, M. Nikaido, H. Nishihara, C. Ozouf-
757 Costaz, D. J. Penman, D. Przybylski, M. Rakotomanga, S. C. Renn, F. J. Ribeiro, M. Ron, W.
758 Salzburger, L. Sanchez-Pulido, M. E. Santos, S. Searle, T. Sharpe, R. Swofford, F. J. Tan, L.
759 Williams, S. Young, S. Yin, N. Okada, T. D. Kocher, E. A. Miska, E. S. Lander, B. Venkatesh, R.
760 D. Fernald, A. Meyer, C. P. Ponting, J. T. Streelman, K. Lindblad-Toh, O. Seehausen and F. Di
761 Palma (2014). "The genomic substrate for adaptive radiation in African cichlid fish." Nature
762 **513**(7518): 375-381.

763 Brenner, S. (1974). "Genetics of Caenorhabditis-Elegans." Genetics **77**(1): 71-94.

764 Carvalho, F. R., C. D. R. Nóbrega and A. T. Martins (2020). "Mapping gene expression in social
765 anxiety reveals the main brain structures involved in this disorder." Behavioural Brain Research
766 **394**: 112808.

767 Cavodeassi, F., J. Modolell and J. L. Gómez-Skarmeta (2001). "The Iroquois family of genes: from
768 body building to neural patterning." Development **128**(15): 2847-2855.

769 Chen, J., E. E. Bardes, B. J. Aronow and A. G. Jegga (2009). "ToppGene Suite for gene list
770 enrichment analysis and candidate gene prioritization." Nucleic Acids Res **37**(2): W305-W311.

771 Chiang, M.-C., A. J. Huang, M. E. Wintzer, T. Ohshima and T. J. McHugh (2018). "A role for
772 CA3 in social recognition memory." Behavioural brain research **354**: 22-30.

773 Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Y. Lu and D.
774 M. Ruden (2012). "A program for annotating and predicting the effects of single nucleotide
775 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2;
776 iso-3." Fly **6**(2): 80-92.

777 Colosimo, P. F., C. L. Peichel, K. Nereng, B. K. Blackman, M. D. Shapiro, D. Schluter and D. M.
778 Kingsley (2004). "The genetic architecture of parallel armor plate reduction in threespine
779 sticklebacks." PLoS Biol **2**(5): E109.

780 Conte, M. A., R. Joshi, E. C. Moore, S. P. Nandamuri, W. J. Gammerdinger, R. B. Roberts, K. L.
781 Carleton, S. Lien and T. D. Kocher (2019). "Chromosome-scale assemblies reveal the structural
782 evolution of African cichlid genomes." GigaScience **8**(4).

783 Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G.
784 Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin and G. P. A. Grp (2011). "The variant
785 call format and VCFtools." Bioinformatics **27**(15): 2156-2158.

786 Degner, J. F., A. A. Pai, R. Pique-Regi, J.-B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon,
787 K. Michelini, N. Lewellen and G. E. Crawford (2012). "DNase I sensitivity QTLs are a major
788 determinant of human expression variation." Nature **482**(7385): 390.

789 Dougherty, J. D., E. F. Schmidt, M. Nakajima and N. Heintz (2010). "Analytical approaches to
790 RNA profiling data for the identification of genes enriched in specific cells." Nucleic Acids Res
791 **38**(13): 4218-4230.

792 Fernald, R. D. and K. P. Maruska (2012). "Social information changes the brain." Proc Natl Acad
793 Sci U S A **109 Suppl 2**: 17194-17199.

794 Fields, C., M. D. Adams, O. White and J. C. Venter (1994). "How many genes in the human
795 genome?" Nature Genetics **7**(3): 345-346.

796 Finlay, B. L. and R. B. Darlington (1995). "Linked regularities in the development and evolution
797 of mammalian brains." Science **268**(5217): 1578-1584.

798 Fraser, G. J., R. F. Bloomquist and J. T. Streelman (2008). "A periodic pattern generator for dental
799 diversity." Bmc Biology **6**(1).

800 Fraser, G. J., C. D. Hulsey, R. F. Bloomquist, K. Uyesugi, N. R. Manley and J. T. Streelman
801 (2009). "An ancient gene network is co-opted for teeth on old and new jaws." PLoS biology **7**(2):
802 e1000031.

803 Greenwood, A. K., A. R. Wark, K. Yoshida and C. L. Peichel (2013). "Genetic and neural
804 modularity underlie the evolution of schooling behavior in threespine sticklebacks." Curr Biol
805 **23**(19): 1884-1888.

806 Grote, S., K. Prufer, J. Kelso and M. Dannemann (2016). "ABAEnrichment: an R package to test
807 for gene set expression enrichment in the adult and developing human brain." Bioinformatics
808 **32**(20): 3201-3203.

809 Harris, R. S. (2007). Improved Pairwise Alignment of Genomic DNA. Doctor of Philosophy, The
810 Pennsylvania State University.

811 Hawrylycz, M. J., E. S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller, L. N. van
812 de Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley, C. Abajian, C. F. Beckmann, A. Bernard, D.
813 Bertagnolli, A. F. Boe, P. M. Cartagena, M. M. Chakravarty, M. Chapin, J. Chong, R. A. Dalley,
814 B. David Daly, C. Dang, S. Datta, N. Dee, T. A. Dolbeare, V. Faber, D. Feng, D. R. Fowler, J.
815 Goldy, B. W. Gregor, Z. Haradon, D. R. Haynor, J. G. Hohmann, S. Horvath, R. E. Howard, A.

816 Jeromin, J. M. Jochim, M. Kinnunen, C. Lau, E. T. Lazarz, C. Lee, T. A. Lemon, L. Li, Y. Li, J.
817 A. Morris, C. C. Overly, P. D. Parker, S. E. Parry, M. Reding, J. J. Royall, J. Schulkin, P. A.
818 Sequeira, C. R. Slaughterbeck, S. C. Smith, A. J. Sodt, S. M. Sunkin, B. E. Swanson, M. P. Vawter,
819 D. Williams, P. Wohnoutka, H. R. Zielke, D. H. Geschwind, P. R. Hof, S. M. Smith, C. Koch, S.
820 G. N. Grant and A. R. Jones (2012). "An anatomically comprehensive atlas of the adult human
821 brain transcriptome." Nature **489**(7416): 391-399.

822 Hitti, F. L. and S. A. Siegelbaum (2014). "The hippocampal CA2 region is essential for social
823 memory." Nature **508**(7494): 88-92.

824 Hochberg, Y. and Y. Benjamini (1990). "More powerful procedures for multiple significance
825 testing." Stat Med **9**(7): 811-818.

826 Johnson, Z. V., E. C. Moore, R. Y. Wong, J. R. Godwin, J. T. Streelman and R. B. Roberts (2020).
827 "Exploratory behaviour is associated with microhabitat and evolutionary radiation in Lake Malawi
828 cichlids." Animal Behaviour **160**: 121-134.

829 Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler and D. Haussler
830 (2002). "The human genome browser at UCSC." Genome research **12**(6): 996-1006.

831 Kocher, T. D. (2004). "Adaptive evolution and explosive speciation: The cichlid fish model."
832 Nature Reviews Genetics **5**(4): 288-298.

833 Kratochwil, C. F., Y. Liang, J. Gerwin, J. M. Woltering, S. Urban, F. Henning, G. Machado-
834 Schiaffino, C. D. Hulsey and A. Meyer (2018). "Agouti-related peptide 2 facilitates convergent
835 evolution of stripe patterns across cichlid fish radiations." Science **362**(6413): 457-460.

836 Lamichhaney, S., J. Berglund, M. S. Almen, K. Maqbool, M. Grabherr, A. Martinez-Barrio, M.
837 Promerova, C. J. Rubin, C. Wang, N. Zamani, B. R. Grant, P. R. Grant, M. T. Webster and L.
838 Andersson (2015). "Evolution of Darwin's finches and their beaks revealed by genome
839 sequencing." Nature **518**(7539): 371-375.

840 Lee, T. H., H. Guo, X. Y. Wang, C. Kim and A. H. Paterson (2014). "SNPhylo: a pipeline to
841 construct a phylogenetic tree from huge SNP data." Bmc Genomics **15**.

842 Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler
843 transform." Bioinformatics **25**(14): 1754-1760.

844 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R.
845 Durbin and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format
846 and SAMtools." Bioinformatics **25**(16): 2078-2079.

847 Loh, Y. H. E., E. Bezault, F. M. Muenzel, R. B. Roberts, R. Swofford, M. Barluenga, C. E. Kidd,
848 A. E. Howe, F. Di Palma, K. Lindblad-Toh, J. Hey, O. Seehausen, W. Salzburger, T. D. Kocher
849 and J. T. Streelman (2013). "Origins of Shared Genetic Variation in African Cichlids." Molecular
850 Biology and Evolution **30**(4): 906-917.

851 Loh, Y. H. E., L. S. Katz, M. C. Mims, T. D. Kocher, S. V. Yi and J. T. Streelman (2008).
852 "Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake
853 Malawi cichlids." Genome Biology **9**(7).

854 Malinsky, M., H. Svardal, A. M. Tyers, E. A. Miska, M. J. Genner, G. F. Turner and R. Durbin
855 (2018). "Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected
856 by gene flow." Nature Ecology & Evolution **2**(12): 1940-1955.

857 Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R.
858 Sandstrom, H. Qu and J. Brody (2012). "Systematic localization of common disease-associated
859 variation in regulatory DNA." Science: 9718-9723.

860 McKaye, K. R., S. M. Louda and J. Jay R. Stauffer (1990). "Bower Size and Male Reproductive
861 Success in a Cichlid Fish Lek." The American Naturalist **135**(5): 597-613.

862 McLean, C. Y., P. L. Reno, A. A. Pollen, A. I. Bassan, T. D. Capellini, C. Guenther, V. B. Indjeian,
863 X. Lim, D. B. Menke, B. T. Schaar, A. M. Wenger, G. Bejerano and D. M. Kingsley (2011).
864 "Human-specific loss of regulatory DNA and the evolution of human-specific traits." Nature
865 **471**(7337): 216-219.

866 Mesquita, L. T., A. R. Abreu, A. R. de Abreu, A. A. de Souza, S. R. de Noronha, F. C. Silva, G.
867 S. V. Campos, D. A. Chianca and R. C. de Menezes (2016). "New insights on amygdala:
868 Basomedial amygdala regulates the physiological response to social novelty." Neuroscience **330**:
869 181-190.

870 Murata, Y., M. Tamura, Y. Aita, K. Fujimura, Y. Murakami, M. Okabe, N. Okada and M. Tanaka
871 (2010). "Allometric growth of the trunk leads to the rostral shift of the pelvic fin in teleost fishes."
872 Dev Biol **347**(1): 236-245.

873 O'Connell, L. A. and H. A. Hofmann (2012). "Evolution of a vertebrate social decision-making
874 network." Science **336**(6085): 1154-1157.

875 Okhovat, M., A. Berrio, G. Wallace, A. G. Ophir and S. M. Phelps (2015). "Sexual fidelity trade-
876 offs promote regulatory variation in the prairie vole brain." Science **350**(6266): 1371-1374.

877 Olton, D. S., J. T. Becker and G. E. Handelmann (1979). "Hippocampus, space, and memory."
878 Behavioral and Brain sciences **2**(3): 313-322.

879 Parnell, N. F. and J. T. Streelman (2013). "Genetic interactions controlling sex and color establish
880 the potential for sexual conflict in Lake Malawi cichlid fishes." Heredity **110**(3): 239-246.

881 Patel, R. K. and M. Jain (2012). "NGS QC Toolkit: A Toolkit for Quality Control of Next
882 Generation Sequencing Data." Plos One **7**(2): 7.

883 Pfenning, A. R., E. Hara, O. Whitney, M. V. Rivas, R. Wang, P. L. Roulhac, J. T. Howard, M.
884 Wirthlin, P. V. Lovell and G. Ganapathy (2014). "Convergent transcriptional specializations in the
885 brains of humans and song-learning birds." Science **346**(6215): 1256846.

Piñero, J., À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz and L. I. Furlong (2017). "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants." Nucleic Acids Research **45**(D1): D833-D839.

Rada-Iglesias, A., R. Bajpai, S. Prescott, S. A. Brugmann, T. Swigut and J. Wysocka (2012). "Epigenomic Annotation of Enhancers Predicts Transcriptional Regulators of Human Neural Crest." Cell Stem Cell **11**(5): 633-648.

Ramakrishnan Varadarajan, A., R. Mopuri, J. T. Streelman and P. T. McGrath (2018). "Genome-wide protein phylogenies for four African cichlid species." BMC Evol Biol **18**(1): 1.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Res **43**(7): e47.

Roberts, R. B., Y. Hu, R. C. Albertson and T. D. Kocher (2011). "Craniofacial divergence and ongoing adaptation via the hedgehog pathway." Proc Natl Acad Sci U S A **108**(32): 13194-13199.

Roberts, R. B., J. R. Ser and T. D. Kocher (2009). "Sexual conflict resolved by invasion of a novel sex determiner in Lake Malawi cichlid fishes." Science **326**(5955): 998-1001.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

Shapiro, M. D., Z. Kronenberg, C. Li, E. T. Domyan, H. Pan, M. Campbell, H. Tan, C. D. Huff, H. Hu, A. I. Vickrey, S. C. Nielsen, S. A. Stringham, H. Hu, E. Willerslev, M. T. Gilbert, M. Yandell, G. Zhang and J. Wang (2013). "Genomic diversity and evolution of the head crest in the rock pigeon." Science **339**(6123): 1063-1067.

Shi, Z., G. Luo, L. Fu, Z. Fang, X. Wang and X. Li (2013). "miR-9 and miR-140-5p target FoxP2 and are regulated as a function of the social context of singing behavior in zebra finches." J Neurosci **33**(42): 16510-16521.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller and D. Haussler (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." Genome Res **15**(8): 1034-1050.

Siepel, A. and D. Haussler (2004). "Phylogenetic estimation of context-dependent substitution rates by maximum likelihood." Molecular biology and evolution **21**(3): 468-488.

Simoes-Costa, M. and M. E. Bronner (2015). "Establishing neural crest identity: a gene regulatory recipe." Development **142**(2): 242-257.

Streelman, J., C. L. Peichel and D. Parichy (2007). "Developmental genetics of adaptation in fishes: the case for novelty." Annu. Rev. Ecol. Evol. Syst. **38**: 655-681.

Streelman, J. T., R. C. Albertson and T. D. Kocher (2003). "Genome mapping of the orange blotch colour pattern in cichlid fishes." Mol Ecol **12**(9): 2465-2471.

Streelman, J. T. and P. D. Danley (2003). "The stages of vertebrate evolutionary radiation." Trends in Ecology & Evolution **18**(3): 126-131.

Sylvester, J. B., C. A. Rich, Y.-H. E. Loh, M. J. van Staaden, G. J. Fraser and J. T. Streelman (2010). "Brain diversity evolves via differences in patterning." Proceedings of the National Academy of Sciences **107**: 9718-9723.

Sylvester, J. B., C. A. Rich, C. Yi, J. N. Peres, C. Houart and J. T. Streelman (2013). "Competing signals drive telencephalon diversity." Nature Communications **4**(1): 4.

Trapnell, C., L. Pachter and S. L. Salzberg (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.

Tucker, J. A., K. A. Mintzer and M. C. Mullins (2008). "The BMP signaling gradient patterns dorsoventral tissues in a temporally progressive manner along the anteroposterior axis." Dev Cell **14**(1): 108-119.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel and M. A. DePristo (2013). "From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline." Curr Protoc Bioinformatics **43**: 11.10.11-33.

Volff, J. N., C. Korting and M. Schartl (2000). "Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes." Mol Biol Evol **17**(11): 1673-1684.

White, D. E., J. B. Sylvester, T. J. Levario, H. Lu, J. T. Streelman, T. C. McDevitt and M. L. Kemp (2015). "Quantitative multivariate analysis of dynamic multicellular morphogenic trajectories." Integrative Biology **7**(7): 825-833.

Xu, X., A. B. Wells, D. R. O'Brien, A. Nehorai and J. D. Dougherty (2014). "Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders." J Neurosci **34**(4): 1420-1431.

York, R. A., C. Patil, K. Abdilleh, Z. V. Johnson, M. A. Conte, M. J. Genner, P. T. McGrath, H. B. Fraser, R. D. Fernald and J. T. Streelman (2018). "Behavior-dependent cis regulation reveals genes and pathways associated with bower building in cichlid fishes." Proceedings of the National Academy of Sciences **115**(47): E11081-e11090.

Zou, D., L. Chen, D. Deng, D. Jiang, F. Dong, C. McSweeney, Y. Zhou, L. Liu, G. Chen and Y. Wu (2016). "DREADD in parvalbumin interneurons of the dentate gyrus modulates anxiety, social interaction and memory extinction." Current Molecular Medicine **16**(1): 91-102.

956 **Table 1**

957

958

Species	Habitat	Sample Source	Accession number	UMD2a_map_percent	UMD2a_meancov	UMD1_map_percent	UMD1_meancov
Metriaclima zebra	Rock	Lab bred	SRR6322515	97.50	42.63	96.81	46.68
Labeotropheus trewavasae	Rock	Streelman Finclip #2238	SRR6314224	98.23	14.89	97.53	16.42
Labeotropheus fuelleborni	Rock	Lab bred	SRR6314225	97.78	27.02	97.09	29.79
Cynotilapia afra	Rock	Lab bred	SRR6314226	98.89	23.88	98.24	26.33
Cyathochromis obliquidens	Rock	Streelman Finclip #2368	SRR6314227	98.52	13.95	97.82	15.38
Metriaclima patricki	Rock	Lab bred	SRR6314228	97.78	37.41	97.20	41.21
Petrotilapia 'chitimba'	Rock	Lab bred	SRR6314229	95.04	27.82	94.20	30.60
Genyochromis mento	Rock	Streelman Finclip #2170	SRR6314230	91.77	8.83	91.10	9.74
Mchenga conophorus	Sand	Lab bred	SRR5438125	98.39	35.44	97.70	39.05
Ctenopharynx nitidus	Sand	Martin Genner [#17]	SRR5438124	97.21	30.22	96.39	33.26
Copadichromis likomae	Sand	Martin Genner [#52]	SRR5438110	96.72	27.06	96.08	29.83

Copadichromis sp. “mloto goldcrest" *sensu Konings*	Sand	Lab bred	SRR5438123	99.20	21.68	98.57	23.93
Tramitichromis "chembe"	Sand	Ryan York	SRR5438109	87.84	16.50	87.28	18.20
Taeniolethrinops furcicauda	Sand	Martin Genner [#25]	SRR5438122	87.54	15.48	86.96	17.07
Tramitichromis intermedius	Sand	Lab bred	SRR5438118	97.81	35.60	97.04	39.29
Copadichromis virginalis "yellow blaze nkanda"	Sand	Lab bred	SRR5438119	97.96	34.54	97.28	38.07
Fossochromis rostratus	Sand	Streelman finclip stock [2549]	SRR5438116	97.32	28.95	96.77	31.97
Aulonocara baenschi	Sand	Lab Bred	SRR5438115	89.31	28.04	88.71	30.91
Dimidiochromis compressiceps	Sand	Lab Bred	SRR5438114	95.78	24.62	95.18	27.16
Mylochromis sphaerodon	Sand	Lab Bred	SRR5438113	97.37	24.63	96.60	27.12
Nimbochromis polystigma	Sand	Ryan York	SRR5438113	81.57	16.77	81.06	18.49
Dimidiochromis kiwinge	Sand	Streelman finclip stock [2384]	SRR5438112	96.46	16.33	95.93	18.03

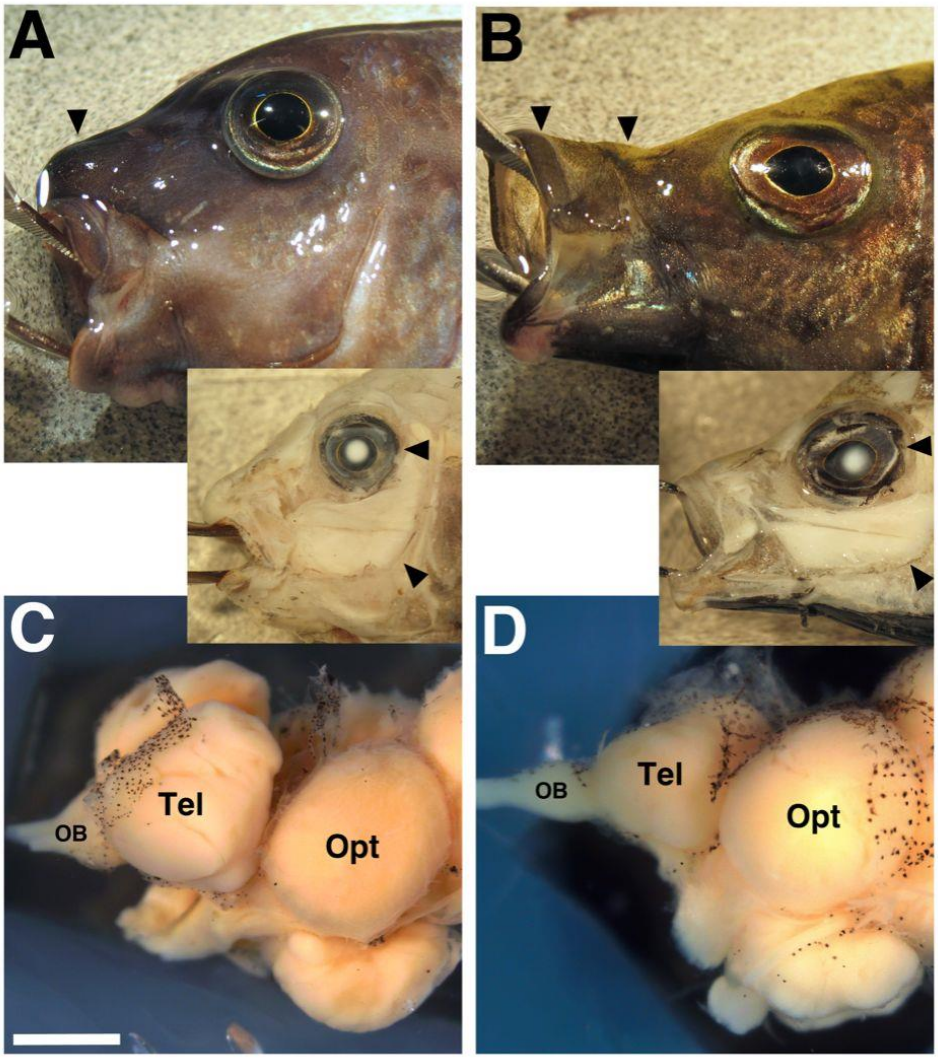
959

960

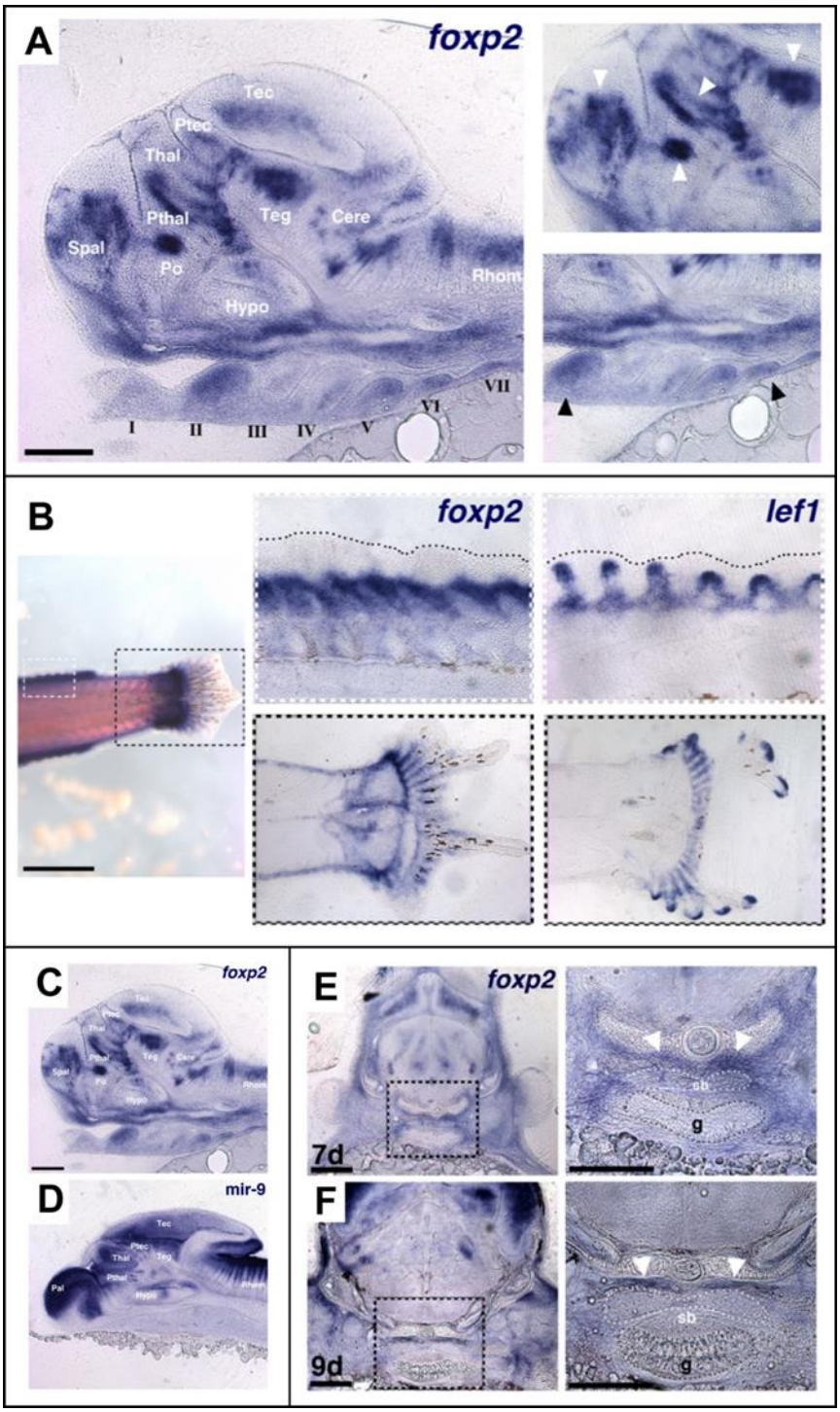
961 **Table 5**

Socially Rock F_1 Hybrid Males	**Socially Sand F_1 Hybrid Males**	**Socially Lone F_1 Hybrid Males**
Metriaclima zebra ♂ *X Tramitichromis intermedius* ♀	*Metriaclima zebra* ♂ *X Tramitichromis intermedius* ♀	*Metriaclima zebra* ♂ *X Mchenga conophoros* ♀
Petrotilapia nigra ♂ *X Aulonacara baenschi* ♀	*Petrotilapia nigra* ♂ *X Aulonacara baenschi* ♀	*Labeotropheus feulleborni* ♂ *X Mchenga conophoros* ♀
Petrotilapia nigra ♂ *X Mchenga conophoros* ♀	*Petrotilapia nigra* male ♂ *X Mchenga conophoros* ♀	
Labeotropheus feulleborni ♂ *X Mchenga conophoros* ♀	*Labeotropheus feulleborni* ♂ *X Mchenga conophoros* ♀	

962



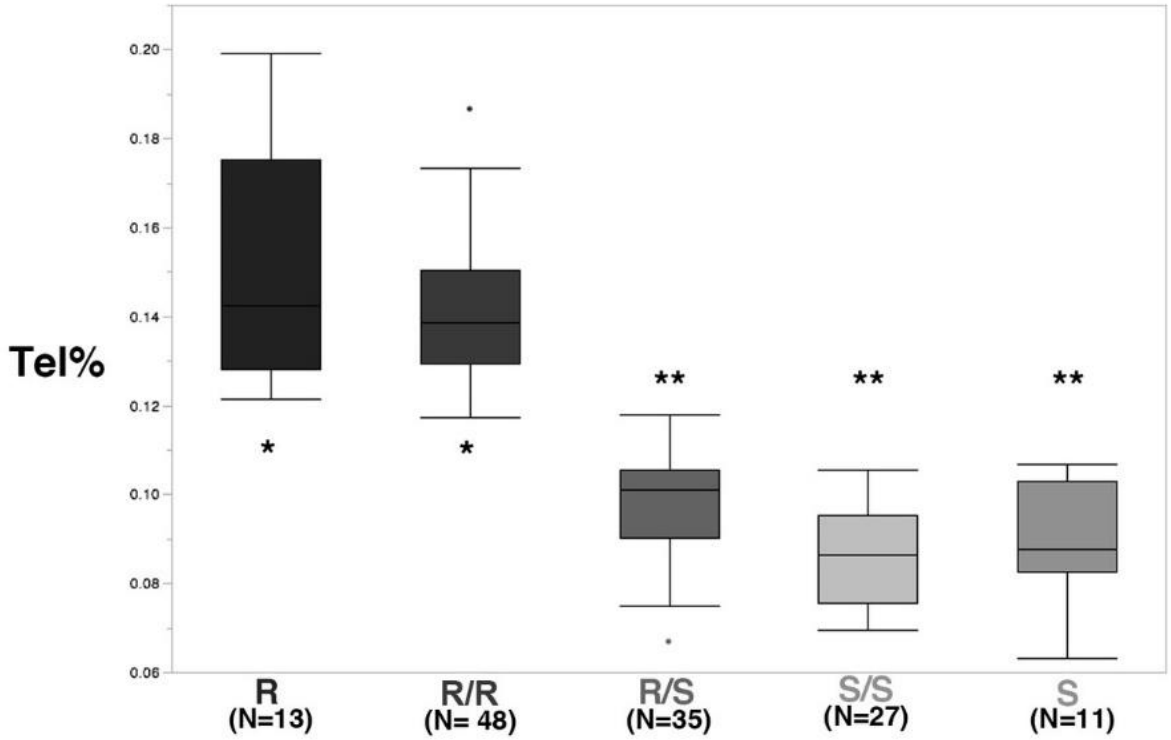
963
964
965 SI Figure 1: Differences in the face (A-B) and brain (C-D) between rock- vs. sand-dwelling
966 Lake Malawi cichlids | Rock-dwellers (A, C) have strongly reinforced jaws, smaller eyes,
967 typically larger cheek muscles, larger olfactory bulbs and telencephala. Sand-dwellers
968 have kinematic, gracile jaws, larger eyes, less robust cheek musculature, and large optic
969 tecta.
970



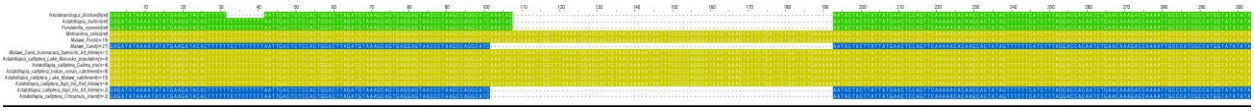
SI Figure 2: Malawi cichlid *foxp2* is expressed in the brain and in all sonic organs | We previously noted novel expression domains for cichlid *foxp2* (Bloomquist, Fowler et al. 2017), elaborated here (A) Expression of *foxp2* throughout the brain (regions labeled in white) and developing pharyngeal arches (labeled I – VII) at 5 days post fertilization (dpf).

977 On the far right, white arrows indicate expression in developing ganglia in the midbrain,
978 diencephalon, pre-optic region, and subpallium. Black arrows point to expression in
979 pharyngeal arches II and V. (B) *foxp2* expression, along with the WNT pathway
980 transcription factor *lef1*, in the developing dorsal and tail fins at 7 dpf. The white- and
981 black-dashed boxes on the right-most panels are zoomed in, midline sections of the
982 boxes on the left. Adjacent, overlapping expression of *foxp2* and *lef1* are indicative of
983 interaction between *foxp2* and the WNT pathway (Bonkowsky, Wang et al. 2008). (C) and
984 (D) show the expression of *foxp2* and *micro RNA-9* (*mir-9*). *mir-9* has been shown to
985 regulate *foxp2* activity in vertebrates (Shi, Luo et al. 2013) and the anti-correlated
986 expression patterns in cichlids suggest a similar interaction. (E) and (F) document
987 expression of *foxp2* in the developing swim bladder at 7 and 9 dpf. The black-dashed
988 boxed on the left panels indicate the zoomed panels on the right. *foxp2* is generally
989 expressed in the mesenchyme within and dorsal to the swim bladder at 7 dpf (white
990 arrows). Once the swim bladder epithelium forms by 9 dpf, *foxp2* expression is localized
991 dorsally (white arrows). All scale bars are 100μm. Abbreviations: Rhom =
992 Rhombencephalon, Cere = Cerebellum, Teg = Tegmentum, Tec = Optic Tectum, Ptec =
993 Pretectum, Thal = Thalamus, Pthal = Prethalamus, Po = Pre-optic area, Hypo =
994 Hypothalamus, Spal = Subpallium, sb = swim bladder, g = gut.

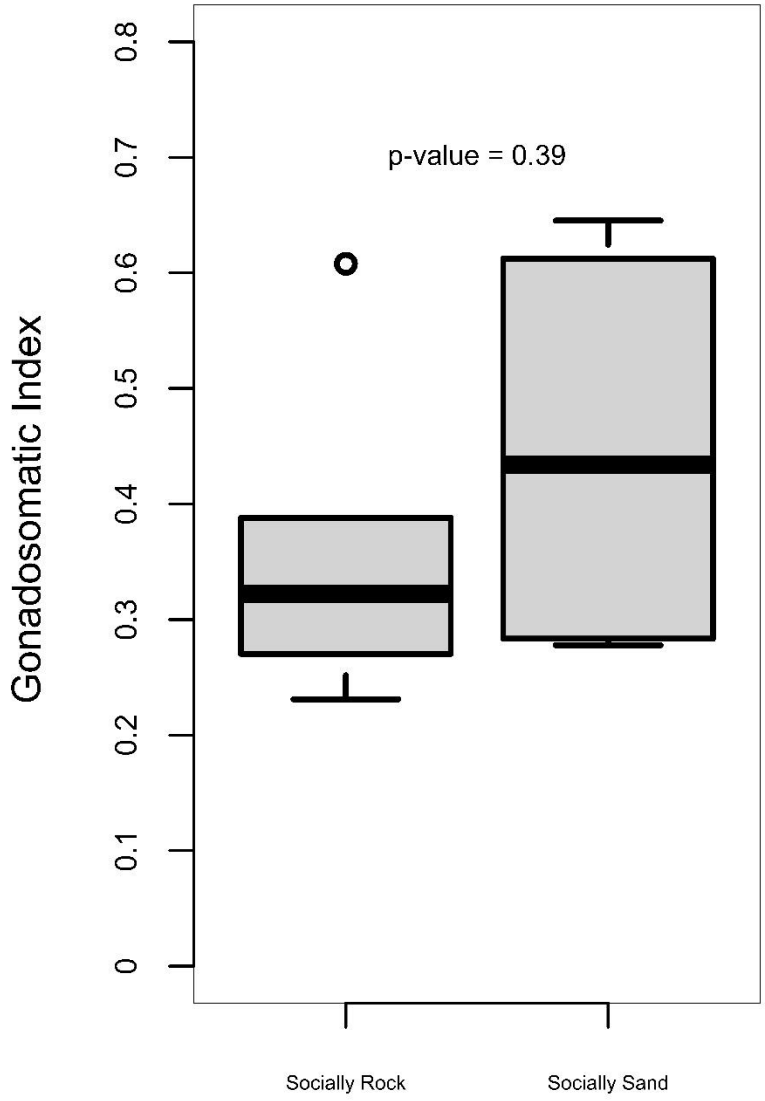
995



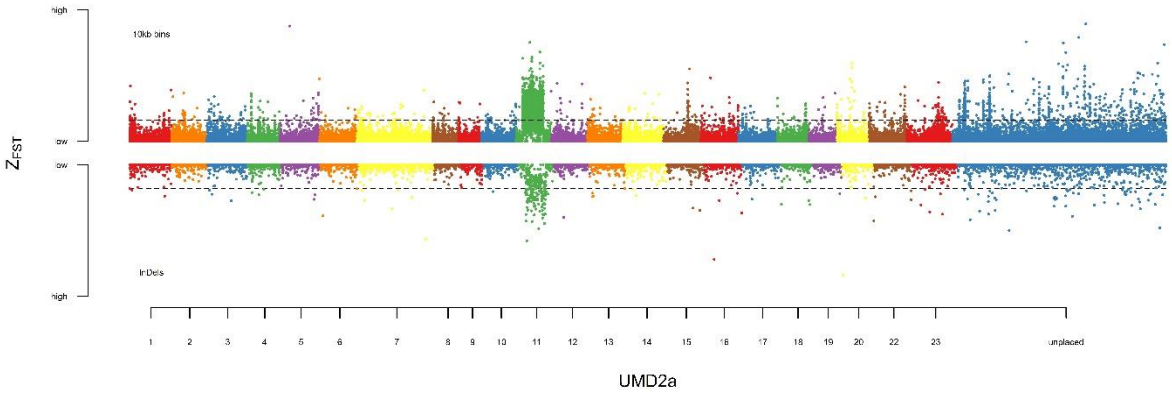
SI Figure 3: The relative size of the telencephalon differs in rock-, sand- and F_2 hybrids indexed for *irx1b* genotypes | Rock-, sand- and F_2 hybrid individuals indexed for *irx1b* genotype were sampled at stages 12-14. We calculated the volume of the telencephalon in each individual and express this as a percentage of total forebrain volume. See also Figure 2.



SI Figure 4: Schematic of an InDel located in the 3'UTR of the Malawi cichlid *irx1b* gene. Summary of allelic states: (1) rock-dwelling species possess an 85bp insertion (yellow) with similarity to Rex1 non-LTR retrotransposon, compared to outgroup species (shown in green); (2) sand-dwellers generally lack the insertion and exhibit a 6bp deletion (blue), compared to outgroup species. Note that *Aulonocara baenschi* is heterozygous. Most individuals of *Astatotilapia calliptera* carry the rock- insertion allele; however, individuals at Chizumulu Island and Itupi possess the sand allele.



SI Figure 5: Gonadosomatic index (GSI) does not differ amongst males behaving as social rock- or social sand-. See also Figure 3.



SI Figure 6: Genomic differentiation amongst sand-dwelling species that construct pit vs. castle sand bowers **|** Z-F_{ST} plot shows genome divergence amongst pit-digging vs. castle-building sand-dweller species, updated by mapping variants to the UMD2a reference genome (data from (York, Patil et al. 2018)). Threshold lines indicate 2.5% FDR.