GENOMIC BASIS OF EVOLUTIONARY RADIATION IN LAKE MALAWI CICHLIDS

A Dissertation

Presented to

The Academic Faculty

by

Chinar Patil

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy in Biology in the

School of Biological Sciences

Georgia Institute of Technology

December 2018

GENOMIC BASIS OF ADAPTIVE RADIATION IN LAKE MALAWI CICHLIDS

Approved by:

Dr. J T Streelman, Advisor

School of Biological Sciences

Georgia Institute of Technology

Dr. Michael Goodisman

School of Biological Sciences

Georgia Institute of Technology

Dr. Soojin Yi

School of Biological Sciences

Georgia Institute of Technology

Dr. Fredrik O. Vannberg

School of Biological Sciences

Georgia Institute of Technology

Dr. Reade B. Roberts

W. M. Keck Center for Behavioral Biology

Comparative Medicine Institute

North Carolina State University

Date Approved: [October 31, 2018]

Dedicated to Neha Ahirrao and Palvi Patil. One for the past, one for the future.

ACKNOWLEDGEMENTS

I would like to begin by thanking my advisor Todd Streelman for standing by me through thick and thin, teaching me valuable lessons and doing a lot of the heavy lifting required to make me the scientist I am today. I would also like to thank past and present members of the Streelman Lab, too many to name all, but in no particular order, Nick Parnell, Kawther Abdilleh, Jon Sylvester, Karen Pottin, Amanda Ballard, Natalie Haddad, Paula Lavantucksin, Zack Johnson, Teresa Fowler and many others.

Beyond just the Streelman lab, I have had the privilege of collaborating with other labs, learning from professors at Georgia Tech . I would like to thank Ryan York, Soojin Yi, Michael Goodisman, Patrick McGrath, Fred Vannberg, Reade Roberts, Joe Lachance among others. I would also like to specifically thank Shweta Biliya from the genomics core who was invaluable support as I fumbled my way through sequencing and somehow ended up with some nice data.

Life in graduate is impossible without friends who share in the journey. In addition to members of the Streelman lab, I would like to thank Linh Chau, Jessica Pruett, Shefali Harankhedkar, Samit Watve, Swetha Srinivasan, Jennifer Pentz, Lavanya Risheshwar, Laurel Jenkins and William Gignac for sharing listening to me complain, being my sound-board for (usually unsound) ideas and most importantly, giving me company while I eat.

Nothing is life is possible without family. I would like to thank my mom and dad for making me who I am in every which way, my brother and sister-in-law, who have been my pillars of support in Atlanta and my niece who is an all-round bundle of joy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ANR	Anterior Neural Ridge
CNE	Conserved Non-coding Element
EG	Early gastrula
hpf	Hours Post Fertilization
InDel	Insertion / Deletion
LG	Late Gastrula
LG	Linkage Group
LoF	Loss of Function
MB	Megabases
MG	Mid gastrula
NGS	Next Generation Sequencing
SNP	Single Nucleotide Polymorphism

SUMMARY

Understanding the genomic basis and origin of phenotypic variation is one of the fundamental questions of biology. Next generation sequencing technologies have afforded evolutionary biologists unprecedented access to whole genome sequences allowing for in-depth investigations into the bases of phenotypic divergence in non-model organisms.

Adaptive radiations provide a window into recent and ongoing adaptive phenotypic evolution. Lake Malawi was colonized by a single haplochromine lineage diversifying and colonizing the lake in unprecedented numbers. Lake Malawi has about 500-800 recently evolved, closely related cichlid species reflecting phenotypic diversity along the lines of habitat, trophic levels and communication. This genetic system has long served as a screen of phenotypic variation that can be used as 'natural mutants' to elucidate the basis of crucial aspects of tooth diversity, craniofacial development, brain development, retina development, opsin diversity and many other phenotypes in Lake Malawi cichlids. The ability to make viable and fertile interspecific hybrids in the laboratory lends additional strength to this experimental model. In the following study, I examine in detail the basis and origin of phenotypic diversity in Malawi Cichlid lineages along two different axes of adaptive evolutionary differentiation of habitat and communication.

First, I demonstrate a way to identify all the functional variation associated with an older divergence within Lake Malawi. Lake Malawi cichlids are broadly divided into rock dwelling and sand dwelling cichlids based on their habitat. A comparison of 8 rock and 14 sand dweller genomes sequenced at high coverage reveals the genetic variation associated

with divergence along the rock versus sand branches of the Lake Malawi cichlid phylogeny. Divergent variants are (a) more significantly found in regions with a high conservation score indicating functional regions and are also (b) enriched in intergenic regions indicating regulatory differences. Genes near divergent variants are significantly enriched for pathways related to early brain development and adult behavior. F_1 rock-dam X sand-sire hybrid males can perform either rock parent like or sand parent like mating behavior depending on social context. Differentially expressed genes between socially rock and socially sand males are enriched for some of the same brain and behavior related pathways revealed by the genomic comparison. A key early development gene , *irx1b* has an alternately fixed deletion that segregates rock and sand and also shows spatial and temporal patterning differences that define the neural plate border leading to differences in the eye field versus the telencephalon. Genomic comparison indicates that variants are significantly associated with genes involved in behavior and early brain development that reflect behavioral differences between rock and sand and also early developmental differences between rock and sand.

Second, using a more recently evolved phenotype, the bower-building behavior, I uncover the genomic basis of mating behavior. Within the sand dweller lineage, males from many species build typical species specific bower for mating display. I compare 9 males from 9 castle building species to 11 males from 11 pit digging species within the sand dweller lineage. Genomic comparisons show that the bower building has evolved multiple times with thousands of genetic variants strongly associated with pit digging and castle building, suggesting a highly polygenic architecture. F_1 hybrid males of pit and castle species sequentially first dig a pit and then build a castle bower. Whole brain transcriptomes

of behaving F_1 showed that genes near behavior-associated variants display behavior-dependent allele-specific expression with preferential expression of the pit-species allele during pit digging, and of the castle-species allele during castle building. These genes are highly enriched for functions related to neurodevelopment and neural plasticity.

CHAPTER 1. INTRODUCTION

Identifying genes contributing to observed phenotypes is a fundamental question in biology. Phenotypic evolution is inextricably linked to genomic variation and diversification[1]. The Next Generation Sequencing [NGS] revolution has given everyone unprecedented access to large amounts of genomic data[2, 3] vastly enlarging our understanding of these processes on a whole genome scale[4]. With this understanding comes the realization that the question of control of complex traits itself has complex answers [5]. In this framework, we can ask what set of traits are shaped by speciation that define species divergence.

Adaptive radiations, usually driven by ecological opportunity, are exceptional models for understanding the genetic underpinnings of adaptive phenotypic diversity[6, 7]. Young radiations are characterized by species with low genotypic divergence with intermediary or transitionary phenotypic variation[8]. Adaptive radiations are known to follow a three stage model that describes the sequential diversification of a lineage as it colonizes a new habitat[9]. A diversifying lineage colonizing a new niche will sequentially diverge along the axes of habitat, followed by trophic levels and feeding strategies followed by the most recent divergence along the axis of communication.

Studies looking for genetic underpinnings of phenotypic divergence have revealed one of two genomic patterns associated with a given set of traits, depending on the stage and age of the radiation. Every time the marine threespine stickleback colonizes freshwater lakes in North America, the *eda* locus is targeted independently in the genome associated with the loss of armor and spines[10]. Carrion crows in Europe are divided into all-black

in the west and grey-hooded in the east with a small region of overlap in the middle of the continent. In the face of gene-flow, phenotypic divergence between the two types of crows is maintained almost entirely by a < 2 Megabases region in the genome that contains within it genes associated with pigmentation and visual perception[11]. A genome-wide comparison of divergent regions associated with beak shape diversity in the iconic Darwin's finches shows, in addition to one highly characterized 240kb region , peaks of divergence all throughout the genome[8]. In a younger radiation along the crater lakes surrounding the larger lakes of the East Africa Rift Valley Lake system of diversifying lineages of cichlid fishes we see islands of speciation in the genome, many regions characterized by divergence higher than baseline[12]. Depending on the age of divergence of an adaptive radiation and their place in the three stage model of divergence we see a pattern of either one or two big regions of the genome defining the phenotypic divergence or these divergent regions of the genome spread out over the entire genome.

Lake Malawi cichlids consist of over 500 unique cichlid species most of which come from a single haplochromine lineage colonizing the lake between 1-5 million years ago[13-15]. The diversification of the Lake Malawi cichlid lineage is characterized first along habitat followed by trophic levels and then communication axes. This starts with the division of the lineage into rock dwelling and sand dwelling cichlids that further diversify into food acquisition adaptations followed by differentiation into a wide array of mating and display strategies[9, 16]. Lake Malawi cichlids fit the requirement of having a large enough pool of natural phenotypic variants with low genotypic variation[13, 17] . This naturally occurring variation is a powerful screen for phenotype association studies. The

presence of naturally occurring variation can be used as a natural variant screen for phenotypic association studies[18]. “Natural mutants” have been used to uncover the basis and mechanics of specific divergent phenotypes in Lake Malawi cichlid system to uncover the basis of craniofacial diversity[19], dental diversity[20], visual pigment diversity[21], sex determination and color[22], forebrain diversity[23, 24]. Here I leverage the power of the Lake Malawi cichlid system using whole genome re-sequencing to uncover a majority of the genomic differences associated with a phenotype.

In this thesis, I lay out two genomic comparisons within the Lake Malawi cichlids I target the fundamental divergence in the Lake Malawi where the original cichlid lineage diverged into rock dwellers and sand dwellers based on habitat. I try to link the genomic variation in the older rock-sand divergence to observable phenotypic differences between the two lineages. Within the sand dweller lineage, males of many species build species specific typical bowers. I use these bowers as an extended phenotype and link the variation between species that build two quantifiably different bower types to the genetic variation between them. I sequence a total of 28 male individuals from 28 species (Table 1-1) from Lake Malawi and use the genomic variation associated with the phenotypes to delve deeper into the evolutionary trajectory of the diversification of the Lake Malawi cichlid lineage along the three stages of adaptive radiation.

.

CHAPTER 2. GENOME-ENABLED DISCOVERY OF FUNCTIONAL VARIANTS IN BRAIN AND BEHAVIOR[1]

The genomic basis of adaptive divergence in species divergence is a central question in evolutionary biology. Lake Malawi cichlids are a powerful adaptive divergence model due to the large number of species with extreme phenotypic diversity along axes of evolutionary divergence but very low genomic differences. We compare whole genomes of 8 rock dwelling and 14 sand dwelling species reflecting the fundamental rock versus sand divergence in the Malawi cichlid lineage to reveal genetic variants diverging between rock and sand and 4484 genes lie in or around them. Divergent variants are also enriched for intergenic regions with a high conservation score indication functional divergence in regulatory regions. Genes near divergent variants are significantly enriched for pathways related early brain development and adult behavior. *irx1b* is a key brain development gene found near a divergent region and we show it defines rock and sand differences in early development stages in spatial and temporal patterning. Independently, in F_1 rock-dam X sand-sire hybrid males we demonstrate that context dependent rock-like and sand-like behavior is accompanied by differential transcriptional expression among genes near divergent variants. We have leveraged the 'natural mutant' screen in our model to recover a majority of the functional differences between species using whole genome sequencing with confirmation from independent targeted experiments. We show that the rock versus

---

[1] **Patil C**, Sylvester JB, Abdilleh K, Malinsky M, Norsworthy M, Pottin K, Bloomquist RF, McGrath PT, Streelman JT (*in prep*) Genome-enabled discovery of functional variants in brain and behavior

CP Contribution: Genome sequencing, Genome assembly, behavior assay, brain transcriptomics, differential expression analysis

sand divergence in Lake Malawi is to a large extent associated with early developmental patterning in the brain and gene pathways associated with adult behavior.

2.1 Introduction

While simpler traits can be linked to their genetic basis by direct association, we do not understand fully how emergent complex traits are linked to genetic control. The most complex trait whose genomic basis we do not fully understand is behavior. The question of linking genes to behavior was framed in 1974 as the "dual encoding problem" [25]. Complex traits like behavior are driven by two layers of control, both contributing to the ultimate behavioral phenotype. A broad range of genes expressed in a developing brain establishes cellular neural architecture. Interactions between various cellular components of the brain is what gives rise to behavior. Only genetic analysis or only cellular level experiments are not sufficient for explaining the basis of a behavior although both are clearly necessary. To understand behavior, therefore, we must untangle both effects, how genes lay out a nervous system and how the nervous system network gives rise to specific behavioral phenotypes.

Even though behaviors often have a strong genetic basis, it has been difficult to identify natural *causative* genetic changes. A significant problem is the complexity and non-linearity of genetic interactions among causative variants, which make their identification a significant challenge. In fact, this is a complication for nearly all complex traits, not just behaviors. Upwards of 93% of human disease related variants – the complex traits for which we have the most data from genome wide association studies (GWAS) – reside in noncoding DNA sequence. Many of these noncoding variants are regulatory, that is, they

affect the expression of genes[26]. GWAS loci often impact genes near the variant itself – in this case, they are termed cis-acting expression quantitative trait loci[27, 28]. Overall, it is unclear how the genome is activated to produce phenotype, and this may be more vexing for *context-dependent traits* of the nervous system like behaviors. Over the past decade, a number of systems have been developed to identify the neural and/or genetic basis of natural behaviors (e.g., those that evolve in the wild, and not in the laboratory) among vertebrates. These systems and behaviors include vole and field mouse parental care[29, 30], bird song[31] and stickleback schooling[32]. Conceptually, researchers seek to know if the genetic basis of behavior follows general rules, like for instance, the incorporation of *cis*-regulatory logic in the genetic basis of development[33]. Our goal is to identify genetic variants associated with the dual encoding of behavior, using the Lake Malawi cichlid system as a model.

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[15], such that the genomes of individuals across species boundaries remain highly similar. For example, nucleotide diversity across the Malawi species assemblage (0.26%)[17] is less than that among lab strains of the zebrafish (0.48%)[34], comparable to that of chimpanzees (0.24%)[35] and humans (0.11%)[36] and contrasts with the ~1.2% divergence between chimp and human[37]. An appreciable fraction (~50%) of genetic polymorphism identified in Malawi species segregates deeply in cichlid lineages from throughout East Africa -- suggesting that ancient genetic variation fuels diversification of the Malawi flock[17]. Set against this background of genome similarity, Malawi cichlids exhibit staggering diversity in phenotypes including pigmentation, sex determination[22,

38], craniofacial[20, 39, 40] and brain patterning [23, 24]. Recent work has focused on the genomic and early developmental underpinnings of this diversity, in rock- *vs*. sand-dwelling species[23, 24, 39, 41] .

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)[9]. 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[41]. Malawi rock-dwellers are strongly territorial and aggressive; they breed and feed at high density in complex 3D 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 breed on leks where males build sand 'bowers' to attract females. 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 (Appendix A Figure 1).

We target the fundamental divergence in the Lake Malawi where the original cichlid lineage diverged into rock dwellers and sand dwellers based on habitat. We compare whole genomes of rock dwelling and sand dwelling species in an unbiased screen looking for divergent regions. We then delve deeper into more details of the mechanisms of key variants associated with rock/sand divergent genomic regions. We sequenced genomes of 8 male individuals from 8 rock dwelling species and 14 male individuals from 14 sand dwelling species (Table 1-1) .

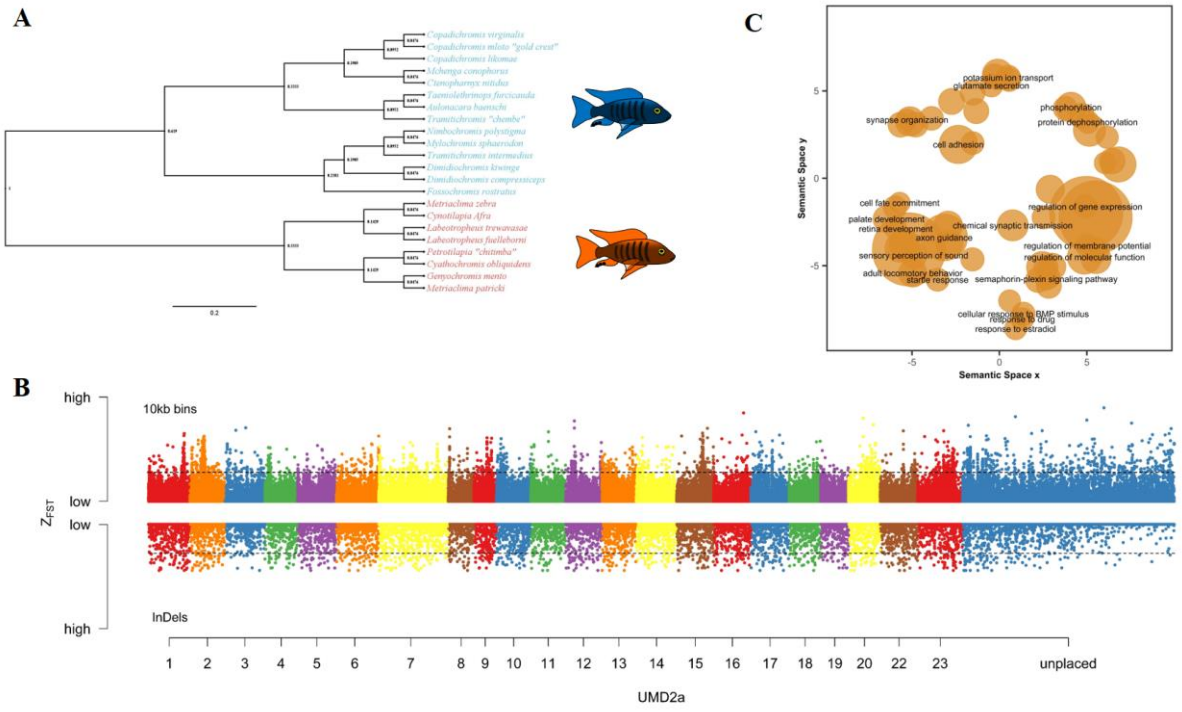


Figure 2-1 : Rock Sand Genomic Variation | (A) A maximum likelihood phylogeny of 8 rock and 14 sand species based on informative SNPs throughout the genome. (B) A plot of Z-F_{ST} across the genome for individual genomes in 10kb windows going up and InDels going down. Threshold lines indicate 2.5% FDR. (C) Gene Ontology(GO) enrichment for genes near variant regions with the GO terms scaled in two dimensions based on semantic similarity.

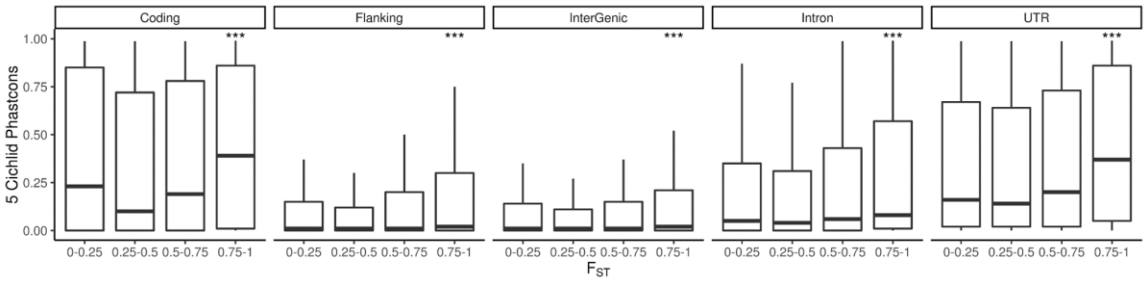
2.2 Results and Discussion

2.2.1 The genomic signature of rock-sand divergence

We compared whole genomes of 8 rock dwellers and 14 sand dwellers aiming to uncover the divergent regions associated with rock versus sand evolutionary diversification. We aligned genomes to a reference genome of nearly 1 gigabase[42]. We identify approximately 22 million Single Nucleotide Polymorphisms [SNPs] and 200,000 Insertion-Deletions [InDels]. We use FST per SNP, averaged across 10kb windows and for each InDels to calculate the divergence between the 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 are mapped to linkage groups (chromosomes), it is apparent that each chromosome carries the signature of rock- vs. sand- divergence(Figure 2-1).

A total of 4,484 genes lie within 25 kb of either an alternately fixed variant, a highly divergent 10kb window [2.5%FDR] or a highly divergent InDel region [2.5% FDR]. Pathway enrichment analysis [43] of human homologs/analogs for these genes reveals categories implicating early embryonic development, brain development, synaptic transmission and neuronal function (Table 2-1)(Figure 2-1C). Divergent genes are significantly enriched for factors implicated in human neurological disease like Autism Spectrum Disorder (SFARI, Fisher's exact test p value < 2e-16) and disorders related to the neural crest [44], (Fisher's exact test p value < 2e-16, Table 2-1). Highlights amongst these lists include



avpr1a, *cntnap2*, GABA receptors, glutamate receptors and members of the Fox, Hox, Dlx Irx, Hh, Wnt and Bmp families/pathways.

Figure 2-2 Functional Variants more likely to be Divergent | PhastCons scores across the genome subdivided by regions of the genome and further into bins of F_{ST} . Flanking and Intergenic regions have low conservation scores compared to Coding regions, Introns and UTRs. PhastCons, , indicating functionality, is significantly higher for higher bins of F_{ST} values (Wilcoxon rank sum p value < 2e-16)

96% of all variants we discovered are predicted to be non-coding. Among fixed variants, 3.5% were found in coding regions, ~17% in intergenic regions (gene deserts), 38% in introns, 38% in flanking regions (within 25kb up- or downstream of a gene), and 3% in annotated UTRs. Rock vs. sand fixed variants were more likely to be missense/loss-of-function (LoF, 72.6%) than silent (27.3%). We aligned published cichlid reference genomes [13] and calculated a PhastCons score across the genome, measuring evolutionary conversation of each nucleotide position and hence putative functionality of the region. For both coding and non-coding portions of the genome, more divergent genetic variants had higher PhastCons scores (Figure 2-2), suggesting that the variants we have discovered are enriched for function.

Given the degree of craniofacial divergence between rock- and sand- groups, and genome wide enrichment for craniofacial and neural crest biology, we examined published datasets of neural crest and craniofacial enhancers [45, 46] in mammals. These data allow us to identify (i) craniofacial and neural crest cell enhancers conserved between mammals and cichlids and (ii) fixed SNPs between rock and sand species within conserved enhancer elements. A total of 275 craniofacial enhancer elements and 234 human neural crest cell enhancers are evolutionarily conserved between mammals and cichlids. Fixed SNPs were found within the enhancer elements of key genes integral to CNCC development and migration (Table 2-2). Notably, from both enhancer datasets, fixed SNPs were found within the enhancer elements of the gene *nr2f2*, a nuclear receptor gene that co-localizes with the master neural crest regulator *tfap2a*.

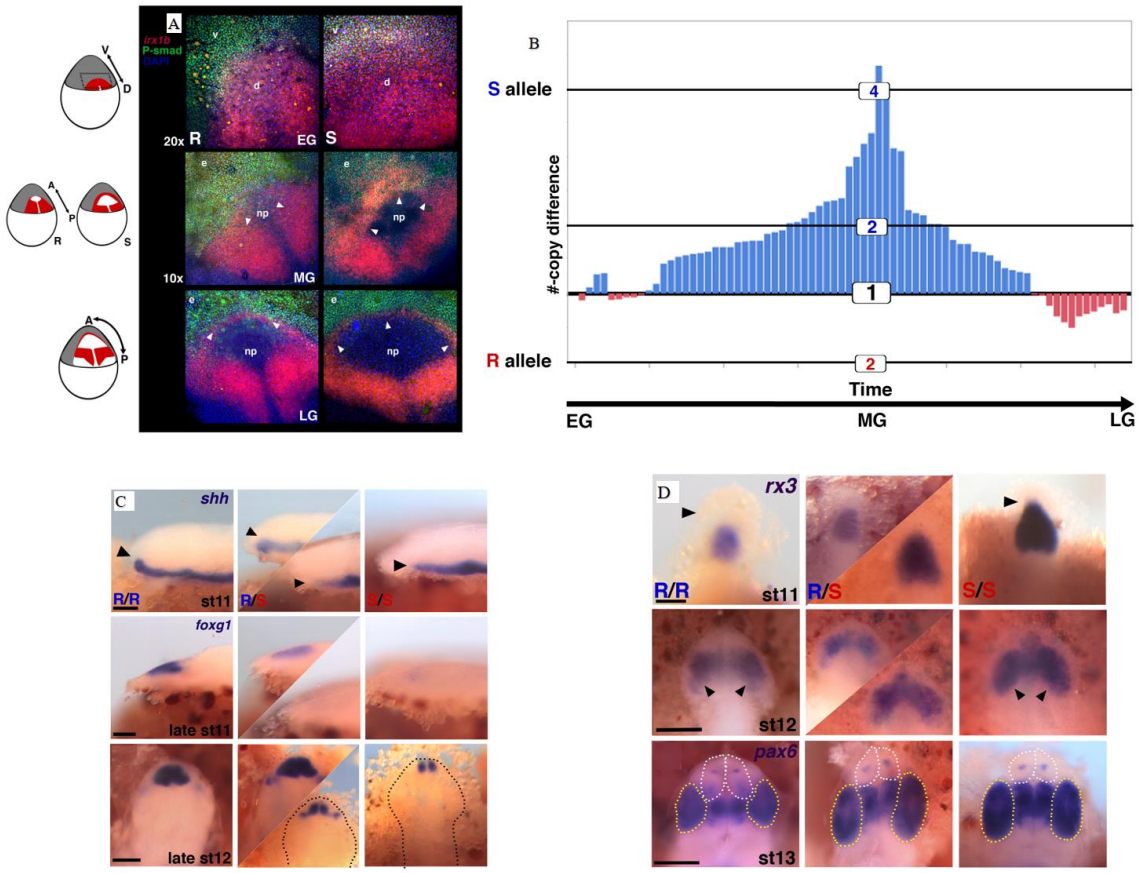


Figure 2-3 Rock Sand *irx1b* expression patterns | (A)Relative spatial expression patterns of *irx1b* in red, P-SMAD in green and DAPI in blue during EG, MG and LG. note the differences marked in the schematic marking the neural plate and the larger field of *irx1b* in the anterior domain. (B)Distribution of copy number from RT-PCR of rock(red) and sand(blue) alleles in F_2 embryos heterozygous for *irx1b*. *irx1b* expression peaks in the sand allele earlier during MG (C) *In-situ* hybridization showing relative differences in *shh* activity, *foxg1*(telencephalon marker) activity and (D) *rx3*(eye field marker) activity in F_2 individuals index for the *irx1b* allele.

We recently examined genome-wide divergence amongst sand-dweller groups that construct pit versus castle bowers, sand-made structures used to attract females for mating[47]. We asked whether rock-sand and pit-castle genomes had diverged similarly. Out of 3070 genes that are targeted by 10kb High F_{ST} regions in the rock vs. sand comparison, 483 overlap with 1090 genes that are targeted by High F_{ST} regions in the pit vs. castle

comparison (p-value < 2e-9, Fisher's exact test). Lake Malawi evolutionary radiations have likely targeted similar genetic modules across evolutionary timescales, although the specific mutations differ.

Genome-wide divergence of rock vs. sand Malawi cichlids involves a relatively small percentage of genetic variants. Divergent variants are (a) evolutionarily conserved and (b) enriched for genes and pathways involved in embryonic development, brain development and function, human neurological disorders and diseases of the neural crest. 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 likely to differ between rock- and sand- groups.

2.2.2 A gastrula-stage map of rock-sand divergence

The complexity of the vertebrate brain is first laid out in the neural plate, a single-cell thick sheet of cells that forms between the non-neural ectoderm and the germ ring at gastrulation stage. Neural crest cells are specified just outside of the presumptive neural plate, under the control of bone morphogenetic protein (BMP) and Wingless (Wnt) signals[48]. The neural plate, in turn, is polarized under the influence of rostral (BMP) and caudal (Wnt) signals[49]. Thus, during gastrula and neurula stages, neighboring cellular territories are defined one from the other, in part by interacting BMP and Wnt gradients, to generate neural vs. neural crest (craniofacial) precursors. IRX genes act as transcriptional repressors of BMP signal in gastrulation, and also function to specify the neural plate[50].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 early eye[49]. Given alternatively fixed SNPs and InDels in the *irx1b* gene (above and[23]), expected interactions with BMP signal in the early embryo and known telencephalon vs. eye size differences between rock- vs. sand- species[23, 24], we examined and quantified the early activity of *irx1b* and BMP in rock- vs. sand- embryos.

We designed a custom microfluidic device to orient and image cichlid embryos in toto at gastrula and neurula stages[51]. In early gastrula (EG), *irx1b* (red) and BMP signal (green) delineate complementary dorsal and ventral domains of the embryo (Figure 2-3A). 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. This second domain, overlapping with BMP activity, is the pre-placodal region and will give rise to sensory placode precursors and rostral-most populations of migratory neural crest cells[49, 52, 53]. 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, sand-dwellers (S) express more *irx1b* in the anterior domain than rock-dwellers at EG and MG and then define the boundary of the neural plate earlier in LG. As a consequence, BMP signal has a longer-lasting influence on the neural plate in rock-dwelling species, which is predicted to result in a relatively larger telencephalon and smaller eye field[49].

We developed a panel of rock- x sand- hybrid crosses to formally evaluate the role of *irx1b* in forebrain diversification. We used quantitative RT-PCR to measure allele-specific expression (ASE) in heterozygous rock- x sand- F_2 hybrids, across gastrulation. We observed that the sand- *irx1b* allele was expressed at significantly higher levels (average

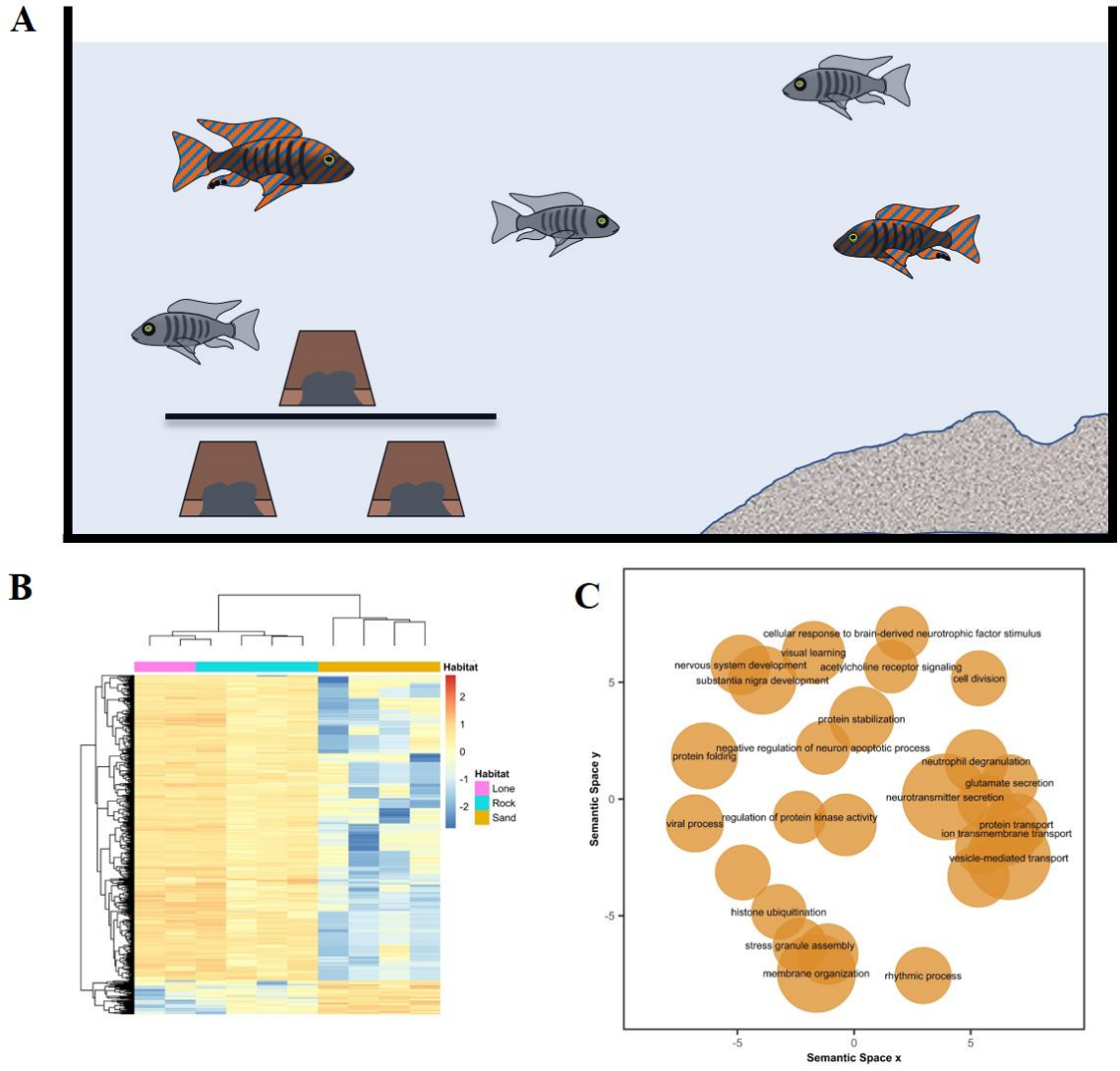


Figure 2-4 Contextual Differential Gene Expression | (A) A Schematic of the rock/sand behavioral paradigm. Terracotta pots simulate rocks on one side with sand on the other side separated by empty tank space. (B) Differential gene expression heatmap for whole brain transcriptomes of F_1 males behaving in either rock or sand social contexts. Note that the individuals cluster by social context and not cross relation. (C) Gene Ontology (GO) terms enriched for differentially expressed genes separated onto two axes based on semantic similarity.

of 2.5-fold) and that this difference was largely confined to MG (Figure 2-3B). Rock- x sand- F_2 hybrids, indexed for *irx1b* genotype, were raised to neurula and somitogenesis stages and we examined the activity of *shh*, *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,* which marks the presumptive telencephalon, and a smaller *rx3* domain (Figure 2-3C D). Note that these phenotypic differences between rock- vs. sand-*irx1b* genotype match the known expression divergence observed amongst rock- vs. sand-species (Figure 2-1)[23, 24]. Finally, when we raised rock- x sand- F_2 to juvenile stage and compare the relative size of the telencephalon among *irx1b* genotypes, individuals homozygous for rock- alleles exhibit larger telencephala (Appendix A Figure 2). We conclude that genetic variants in and around the *irx1b* gene contribute to divergent specification of the Malawi cichlid forebrain. This is notable because the differences in gene expression and BMP activity we detect are noticeable before brain structures are apparent.

2.2.3 The genomics of social challenge and opportunity

One of the main differences between rock- and sand-dwellers is their means of male display and courtship behavior. Rock- males defend caves year round and are highly aggressive, they court females in these caves; sand- males breed seasonally on leks and build and bowers to attract females and mitigate male-male aggression. Given this difference, and genome-wide enrichment for categories related to adult behavior and neuronal function (Table 2-1), we sought to evaluate the brain gene expression profiles of male courtship display. To assess male display and courtship behavior, we designed a courtship preference assay. We evaluated social interactions between males and females using a 40-gallon tank design with a 'rock' habitat at one end and 'sand' at the other, separated by glass bottom (Figure 2-4A). When parental rock- species are placed in this tank paradigm, males court females over the rocks. Males of sand- species court females over sand and construct species-appropriate bowers. When single rock- x sand- F_1 males were placed in this set up

with F_1 females, males invariably courted females over the 'rock' habitat, suggesting genetic dominance. When two rock- x sand- F_1 males were allowed to compete for F_1 females in this tank paradigm, something interesting happened. One male, typically the larger, courted females over the rock habitat, and the other simultaneously constructed bowers to court females in the sand. We detected no difference in GSI (gonadal-somatic index) between F_1 males behaving as 'socially rock' vs. 'socially sand.' This observation of divergent behavior among interacting F_1 brothers suggests an interaction between the genome and the social environment in these males.

We used RNA-seq to investigate the context dependent behavior of rock- x sand- F_1 males in the courtship preference assay (described above). Whole brains of F_1 males tested singly (n=2 lone) as well as F_1 brothers assayed in dyads (n=4 dyads) were dissected after sacrifice during courtship and interrogated using RNA-seq. Genes were considered significantly differentially expressed between "socially rock" and "socially sand" brains if they exhibited both a fold change ≥ 2 and crossed the threshold of Padj < 0.05. Based on this criterion, we found 832 genes differentially expressed between brain transcriptomes (Figure 2-4B, Table 2-3). Gene expression profiles clustered not by fraternal relatedness, but rather by behavior (Figure 2-4B). 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. Using GeneAnalytics[43], we observed significant functional enrichment for brain regions (e.g., cerebral cortex); mental disorders (e.g., Schizophrenia and Frontotemporal Dementia); neural pathways (e.g., Transmission Across Chemical Synapses, Axon Guidance, Neurotransmitter Release, Oxytocin Signaling, Synaptic Transmission and Brain Development)

and brain phenotypes (e.g., Ataxia, Abnormal Spatial Learning and Abnormal CNS Synaptic Transmission). Roughly 38% of differentially expressed genes also contained genetically differentiated SNPs between rock- and sand- species (p-value < 2e-6, Fisher's exact test), implying considerable cis-acting genetic variation. These context-dependent differences in brain gene expression are striking and imply rapid and concerted changes in brain expression modules as males encounter changing social challenge and opportunity[47, 54].

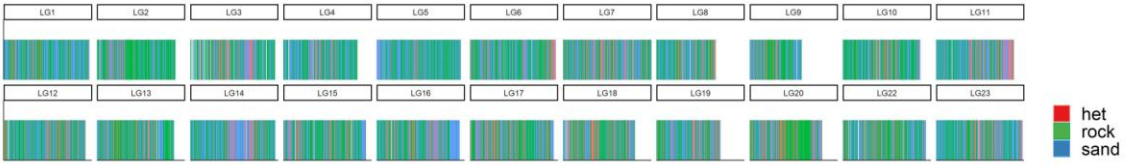


Figure 2-5 *Astatotilapia calliptera* state | 13000 informative SNP markers across the genome on all Linkage Groups marked green if homozygous for the rock allele, blue if homozygous for the sand allele and red if polymorphic

2.2.4 Genome-enabled discovery of natural variants in brain and behaviour

African cichlids are textbook examples of evolutionary diversification. One of the hallmark features of cichlid genomes is allele sharing between species, due in part to gene flow across premating species barriers and the retention of ancestral polymorphism [13]. We sought to explore where the genetic variants differentiating rock- from sand- Malawi species came from. To do so, we compared rock- and sand- genomes to the genome of *Astatotilapia calliptera*, a riverine species located within, but not endemic to, Lake Malawi. Previous research suggests that *A calliptera* genomes from within and just outside Lake Malawi are composites of rock and sand alleles[17]. We painted A. calliptera chromosomes

to identify the state of rock- vs. sand- divergent variants (Figure 2-5). For 42% of differentiated variants, rock- species share the same allelic state with *A. calliptera*, in 11.5% of cases, *A. calliptera* is polymorphic (heterozygous). Notably, we observed numerous regions (e.g., chromosome 14), some as large as 4 Mb, with predominant runs of either rock- or sand- alleles in the *A. calliptera* genome. These large regions are evidence of either very strong divergent selection between rock- and sand- groups, and/or secondary contact between *A. calliptera* and rock- or sand- lineages.

Our goal in this work was to identify genetic variants associated with differences in brain and behavior, because this has been difficult to do in other vertebrate systems. We followed an approach of evolutionary forward genetics[55], wherein we sequenced and compared the full genomes of rock- vs. sand- Lake Malawi species groups, known to differ in brain and craniofacial features, as well as social and courtship behaviors. We found a small percentage (0.06%) of total variants to be genetically differentiated, but nonetheless these variants were enriched for functional categories related to brain and craniofacial development, neuronal function and behavior. This list of variants constitutes a starting place for follow up study in the Malawi system. To illustrate that point, we carried out experiments to uncover new biology in early brain development and later brain function (behavior) – the two components of Brenner's dual encoding problem from genes to behavior.

2.3 Methods

2.3.1 Genome sequencing

We used genomic DNA from the fin clips (Qiagen DNeasy, Cat #69504) from 8 rock dwelling and 14 sand dwelling Lake Malawi species (Table 1-1) obtained from fish

located in-house or from field samples. We made libraries using the Illumina Nextera Library prep kit to perform paired-end sequencing on the Illumina Hi-Seq 2500 at Georgia Tech. Genome assembly, variant discovery and annotation was done using the *Metriaclima zebra* reference genome version MZ_UMD2a [42] using standard BWA[56] and GATK practices[57]. The maximum likelihood tree in Figure 2-1A was made using SNPhylo[58]on the variant data.

2.3.2 Genetically Divergent Regions

Vcftools[59] was used to calculate and 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 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[60] and GeneAnalytics [43] was used to test the genes within 25 kb of significant variants for enrichment of functional categories.

2.3.3 Conserved elements

A comparative genomic approach was used to identify putatively conserved craniofacial and neural crest CNEs between mammals and *M. zebra*. Experimentally verified and published genome-wide craniofacial and neural crest enhancers active during early embryonic stages that play a role in shaping the development of neural crest and craniofacial structures in mammals were identified[45, 46]. We used the liftOver tool [61], which identifies conserved orthologous genomic regions between species to identify conserved CNEs between human, mouse and the M. zebra genome. Starting from mouse and human,

we identified orthologous craniofacial and neural crest CNEs in Nile Tilapia and *Metriaclima zebra*.

2.3.4 RNA Extraction and Sequencing

Adult F_1 rock male♂ x sand female♀ hybrids were reared in tanks with a simulated rock habitat on one side and simulated sand habitat on another separated by empty tank space. As soon as a pair of brothers in this tank set-up exhibited territoriality (danced for females) on both sides at the same time, they were rapidly decapitated, and their brains were fixed in RNA-Later. The males were designated as 'Socially Rock' or 'Socially Sand' depending on the side of the tank they used as their territory. Two of the males, designated as 'Lone' were the only males in their tank. A total of 10 males were sacrificed:

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* ♀	

Tissues were frozen in liquid nitrogen, homogenized using a mortar and pestle and placed in trizol. Following standard chloroform extraction RNeasy mini columns (Qiagen) were utilized to purify RNA for storage at -80^{O}C. 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.

2.3.5 Differential Gene Expression Analysis

Raw sequence reads from "Social Rock" and "Social Sand" samples were quality controlled using the NGS QC Toolkit[62]. Raw reads with an average PHRED quality score below 20 were filtered out. The remaining reads were further trimmed of low-quality bases at the 3' end. High quality sequence reads were aligned to the M.zebra reference genome MZ_UMD2a [42] using TopHat v2.0.9[63]. 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 [64]. Sorted sam files were used as input for the HTSeq-count v0.6.1 program to obtain fragment counts for each locus[65]. Fragment counts were scale-normalized across all samples using the calcNormFactors function in the edgeR package v3.6.8[66]. Relative consistency among replicates and samples was determined via the Multidimensional scaling (MDS) feature within the edgeR package in R. 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 the limma package v3.20.9[67]. Pairwise contrasts were constructed between socially rock and socially sand samples. After correcting for multiple comparisons using the Benjamini-Hochberg method[68], genes were considered differentially expressed between socially rock and socially sand samples if they exhibited both a fold change ≥ 2 and $P_{adj} < 0.05$.

2.3.6 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[24]. The rock-dweller species included *Cynotilapia afra* (CA, planktivore), *Labeotropheus fuelleborni* (LF, algivore) and *Maylandia 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 were measured starting from the earliest eyes can be differentiated from the forebrain (mid-somitogenesis, stage 12) and at each stage until the forebrain has defined prosomeres (early pharyngula, stage 14)[23]. 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.

2.3.7 Staging during gastrula

Cichlid gastrulation was split into three sub stages within the gastrula stage 9. Gastrulation lasts 8 to 12 hours, depending on the species, and is defined as after the shield (as described in zebrafish) stage until the presence of the first somite at the beginning of the neurula stage 10. Embryos were classified as early gastrula (EG) by an asymmetry in epiboly after shield stage until the formation of a ridge that is analogous to the anterior neural ridge (ANR) in chick and mouse and the anterior neural border in zebrafish. At that point embryos are classified as mid gastrula (MG). MG lasts until the formation of the dorsal-ventral axis, defined by further lengthening of one side of the embryo, which begins to thicken as epiboly progresses. This is the dorsal side of the embryo, and the side opposite the ANR is classified the ventral side of the embryo. At this point the embryos are defined as late gastrula (LG). LG ends with the specification of the neural plate, which appears as a portion of the dorsal embryo that is raised relative to ventral side, usually in line with the ANR. In addition to these morphological markers, EG, MG, and LG can be identified via gene expression of *dx3b, irx1b, tlc,* and *sox2,* which all have recognizably different expression domains at each sub-stage.

2.3.8 Immunohistochemical staining

Embryos were harvested at 24 hours post fertilization (hpf) from each of the rock- and sand-dwelling cichlid species that were measured, along with *Metriaclima patricki* for rock- and *Tramitichromis intermedius* for sand-dwellers. The embryos were cultured until they reached gastrula stage, approximately 36 to 40 hpf, then fixed at intervals throughout gastrula until neurula. The embryos were then treated with auto-fluorescence reducer

(1.55mL 5M NaCl, 250ul Tris-HCl, pH 7.5, and 95mg NaBH4) overnight, and 10% 2-mercaptoethanol for 1 hour. Next, whole mount *in situ* hybridization was done, using a modification of methods previously published in Fraser *et al.* 2008. Each gene was visualized using Fast Red (naphthol chromogen, Roche Diagnostics), which fluoresces at near red wavelengths (500-650 nm). After *in situ* hybridization, embryos were immunostained for pSMAD 1,5,8 protein, using protocols published in Tucker *et al.* 2008. Embryos were then bathed in Vectashield (Vector Labs) containing DAPI and placed in a specially built mold that accommodates the large yolk and holds the embryo upright. Embryos were then scanned using a Zeiss LSM 700-405 confocal microscope and processed using LSM 700 software and Image J.

2.3.9 Quantitative PCR

A subset of the embryos cultured for IHC were fixed in RNALater (Qiagen), a total of 12 individuals for each species of rock- and sand-dweller. The 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 *bmp4*, *irx1b*, or *sox2* was quantified using a one-step RT qPCR kit (Express One-Step SYBR GreenER kit, Invitrogen) and RT-PCR Machine (Mastercycler by Eppendorf). Each gene of interest was standardized against *beta-actin* [69] using the equation 2^-(*gene of interest – beta-actin*) to generate delta Ct. Each individual had a total of three replicates for each gene, and the experiment was repeated at least once.

2.3.10 Rock-Sand hybridization and genotyping

Two rock-sand crosses, one between *Copadichromis borleyi* (CB, sand-dweller sire) and *Maylandia 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 either RNA-extracted (stages 9 and 10) or DNA- extracted (stages 11-14). RNA extraction followed the same protocol as normal embryos, DNA extraction was performed by fixing the embryos 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.

2.3.11 F_2 Analysis

The F_2 at stages 9 and 10 had their *irx1b* quantified using the same protocol as the normal embryos. The data was then separated by genotypic class and tested with an ANOVA, followed by a Tukey's multiple comparison test to determine significance between classes. In individuals heterozygous for the *irx1b* allele, 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). The F_2 at the later stages (stages 11 – 14) were either sectioned and measured for eye/forebrain using the same protocol for normal embryos or treated with *in situ* hybridization to visualize genes involved in the formation of the forebrain and eye.

2.3.12 PhastCons analysis

Pairwise alignments generated using lastz v1.02[70], 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 Jim Kent's axtChain tool with -minScore=5000 for cichlid-cichlid and -minScore=3000 for cichlid-other teleost alignments. Additional tools with default parameters were then used following the UCSC whole-genome alignment paradigm (http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto) in order to obtain a contiguous pairwise alignment. Multiple alignment were generated from pairwise alignments using the multiz v11.2[71] program using default parameters and the following pre-determined phylogenetic tree: (((((((*M. zebra*, *P. nyererei*), *A. burtoni*), *N. brichardi*), *O. niloticus*), medaka), stickleback), zebrafish), in agreement with Brawand et al.[13].Sequence conservation scores were then obtained using the phastCons[72] with a phylogenetic model estimated by the phyloFit[73] program, both from the PHAST software package (v.1.3). The model fitting was done using default parameters. The phastCons was run in two iterations, first to obtain the free parameters of the model (--estimate-trees and --no-post-probs) and then using the output from this we run phastCons again the conservation scores. For the 'cichlid-only' phastCons runs, we used -

-target-coverage 0.3 --expected-length 100, while for the broader teleost dataset we specified --target-coverage 0.125 --expected-length 20, which resulted in 53% coverage of exon sequences by conserved elements.

CHAPTER 3. BEHAVIOR-DEPENDENT CIS-REGULATION REVEALS GENES AND PATHWAYS ASSOCIATED WITH BOWER BUILDING IN CICHLID FISHES[2]

Many behaviors are associated with heritable genetic variation. Genetic mapping has revealed genomic regions or, in a few cases, specific genes explaining part of this variation . However, the genetic basis of behavioral evolution remains unclear. Here I investigate the evolution of an innate extended phenotype, bower building, among cichlid fishes of Lake Malawi. Males build bowers of two types, pits or castles, to attract females for mating. we performed comparative genome-wide analyses of 20 bower building species and found that these phenotypes have evolved multiple times with thousands of genetic variants strongly associated with this behavior, suggesting a polygenic architecture. Remarkably, F_1 hybrids of a pit-digging and a castle-building species perform sequential construction of first a pit and then a castle bower. Analysis of brain gene expression in these hybrids showed that genes near behavior-associated variants display behavior-dependent allele-specific expression with preferential expression of the pit-species allele during pit digging, and of the castle-species allele during castle building. These genes are highly enriched for functions elated to neurodevelopment and neural plasticity. Our results suggest that natural behaviors

[2] York RA†, **Patil C**†, Abdilleh K, Johnson ZV, Conte MA, Genner MJ, McGrath PT, Fraser HB, Fernald HB, Streelman JT (2018) Behavior-dependent cis regulation reveals genes and pathways associated with bower building in cichlid fishes , *Proceedings of the National Academy of Sciences* Nov 2018, 201810140; DOI: 10.1073/pnas.1810140115
CP contribution: Sequenced Genomes, assembled genomes, ancient / new allele analysis and gene enrichment.

are associated with complex genetic architectures that alter behavior via cis-regulatory differences whose effects on gene expression are specific to the behavior itself.

3.1 Introduction

Understanding behavioral evolution requires identifying the genetic and regulatory architectures encoding neural development and function. To characterize the evolution of a complex social behavior, we focused on the remarkable bower building feats performed by ~200 cichlid fish species in Lake Malawi that live on sandy substrate. Bowers are species-specific sand structures that serve as signals in male-male competition and female mate choice[74]. Malawi cichlid species build two basic bower types: 1) pits, which are depressions that resemble nests in the sand, and 2) castles, which resemble miniature volcanoes[75]. Bower building requires highly repetitive activity in which males perform hundreds of scoop-spit bouts with their mouths per hour, interspersing construction with the courtship of females and aggressive encounters with conspecific males (Figure 3-1a)[75, 76]. To dig pits, males collect sand from the center of the pit and spit it elsewhere, while to build castles, males gather sand from elsewhere and spit it in a targeted location (Figure 3-1b, c). Pit and castle bower types are distributed widely across the Malawi cichlid sand-dweller phylogeny, suggesting that parallel evolution and/or hybridization may be responsible (4). Furthermore, bower building is innate. Naïve males born in an aquarium, who have neither experienced sand nor other males, perform species-specific bower behavior when housed with sand and gravid females.

3.2 Results

3.2.1 Extensive genetic differences exist between pit and castle species

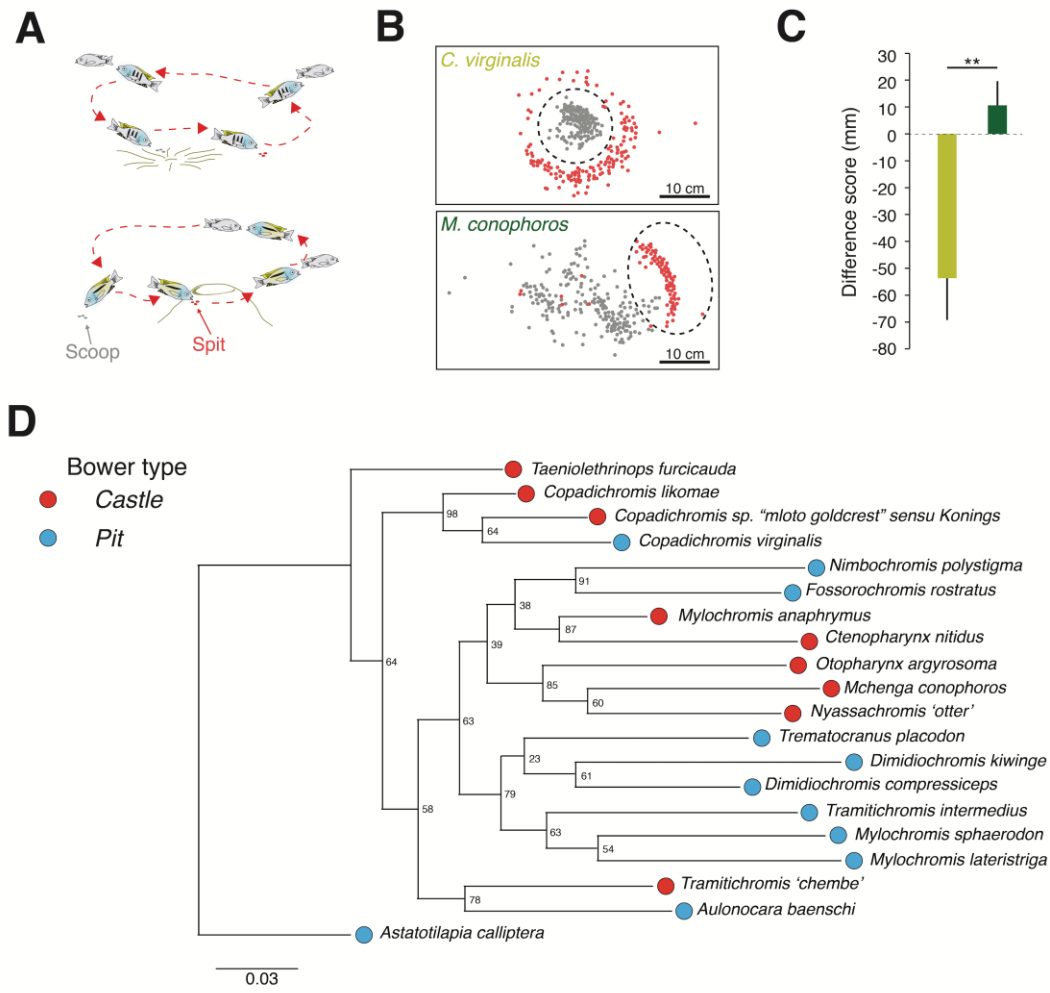


Figure 3-1 Bower building | (A) Characteristic behavioral patterns associated with pit (top) and castle (bottom) bower building. **(B)** Average locations of scoops (grey) and spits (red) during bower building trials in the pit species *Copadichromis virginalis* and the castle species *Mchenga conophoros.* Consensus locations of the bowers are indicated by the hashed black circles. **(C)** Results of a Student's t-test (two-tailed, p = 0.006) comparing difference score (mm) between *C. virginalis* (pit; yellow) and *M. conophoros* (castle; green). **(D)** Maximum likelihood phylogeny from genome-wide variants of the species sequenced in this study, numbers at nodes are bootstrap support values.

To identify genetic variants associated with bower building, we sequenced the genomes of 20 male individuals from 20 sand-dwelling Lake Malawi cichlid species: eleven pit-digging species and nine castle-building species (Table 1-1). Species in these two groups construct either pits or castles despite differences in color pattern, ecology, feeding

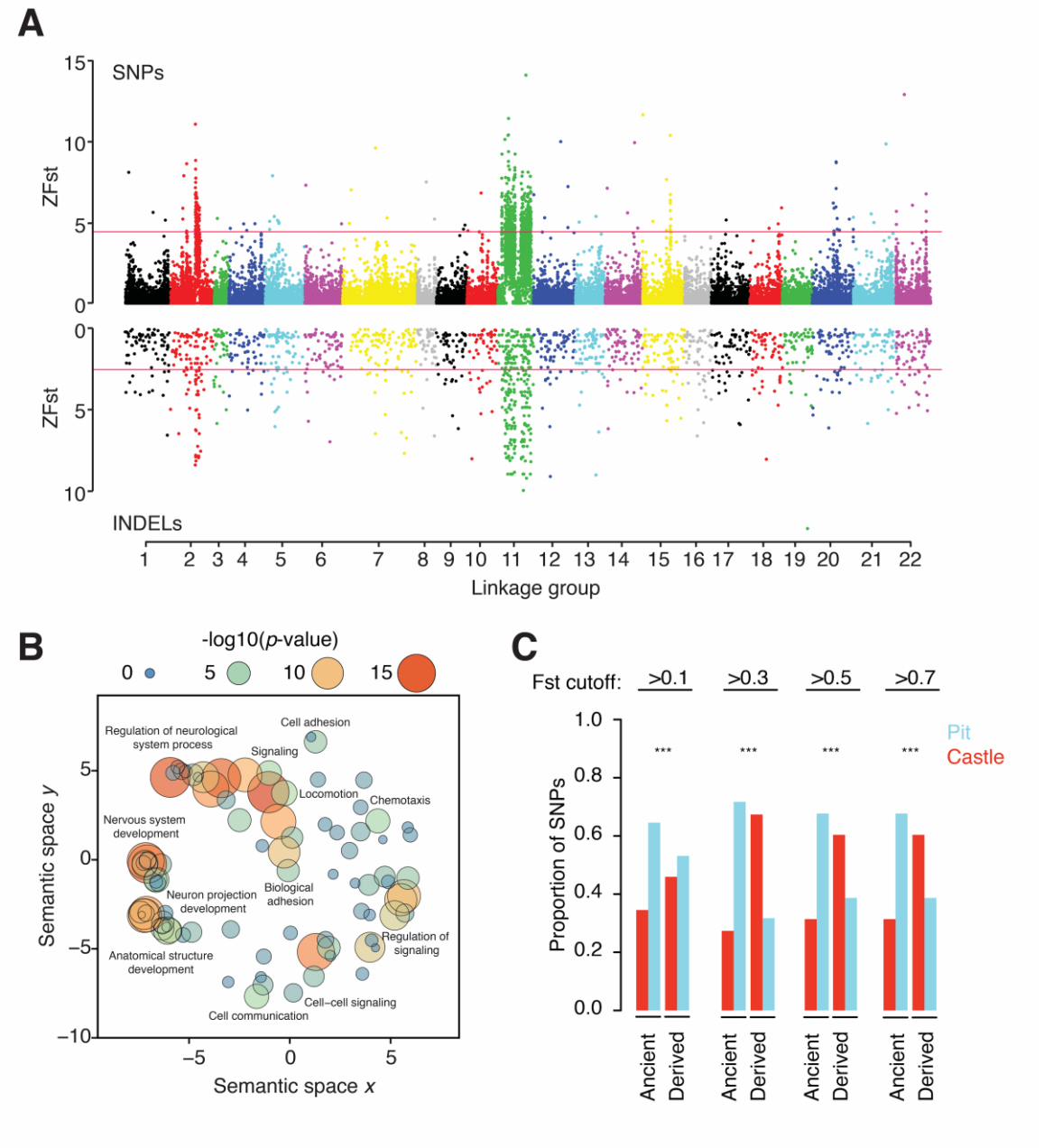


Figure 3-2 Genome-wide divergence associated with bower building | (A) Manhattan plot of genome-wide ZF_{ST} for SNPs and INDELs between pit and castle species. **(B)** Semantic similarity of Gene Ontology Biological Process terms enriched for high F_{ST} variants. **(C)** Barplot of SNP proportions per F_{ST} cutoff for ancestral and derived SNPs. SNPs in which castle species possess the alternate allele are colored red, those in which pit species possess the alternate are colored blue. *** = Fisher's exact p <0.001.

mode and other characteristics[75]. Sequence reads from each species (mean coverage

~25X) were aligned using the Malawi cichlid reference genome[77] and mapped to Malawi linkage groups (Methods). A maximum likelihood phylogeny based on variant sites (Figure 3-1d) is consistent with repeated evolution of pit-digging and castle-building behavior in our sampled species.

Like other recently evolved species flocks[8], East African cichlids share genetic polymorphisms because of incomplete lineage sorting and hybridization[13, 17, 41, 78]. Therefore, we used both population-based and phylogeny-based analytical approaches to understand genomic correlates of bower building. We applied the population-based fixation index (F_{ST}) to identify genomic regions differentiating pit vs. castle species. We observed ~15.5 million single nucleotide polymorphisms (SNPs) and ~130,000 insertion/deletions (indels) in the sample set. 1.5% of variable sites were notably divergent between pit vs. castle groups ($F_{ST} > 0.2$, compared to 0.08 genome-wide mean). We compared patterns of F_{ST} divergence across the genome (Figure 2a) to a population-structure corrected GWAS on bower behavior and found that the two are strongly correlated (Appendix B, Figure 1).

We next identified outlier 10-kilobase (kb) regions based on mean F_{ST} of SNPs (10% FDR, $F_{ST} > 0.2$) and individual insertions or deletions (indels; 10% FDR, $F_{ST} > 0.1$; Figure 2a). Outlier regions were observed on every linkage group, but peaks on LG2 and LG11 are striking for their size (1.3 megabases [Mb] and 6 Mb, respectively) and consistency across SNP and indel data. Broad peaks of differentiation on LGs 2 and 11 could be caused by structural changes (e.g., inversions[79]) associated with bower behavior or, for instance, other traits like male sex determination[22] (T. Kocher, pers. comm.). However, sex determination systems are not known for these species, and we could not identify

structural variants that could explain the broad peaks of genetic differentiation on LGs 2 or 11[80, 81] (Methods).

3.2.2 Characterizing variants associated with bower type

We hypothesized that if high-F_{ST} regions across the genome are the product of selection on bower building or associated behavioral traits, then they should be enriched near genes involved in brain function and development. To test this, we examined 1563 genes, located within 25 kb of an outlier 10 kb region (~30% of these genes are located on LGs 2 and 11), for functional enrichment[43]. This gene set was significantly enriched for tissue types, pathways, human disorders and phenotypes associated with brain development and behavior, including axon guidance, synaptic transmission, autism spectrum disorder and spatial learning (Appendix B Figure 2b, Table 3-1). Together, these analyses identify genomic regions and genetic variants associated with bower behavior and demonstrate that genes near these variants are strikingly enriched for putative function in brain and behavior.

To assess the role of ancestral variation in differentiation of pit-digging vs. castle-building behaviors, we classified SNPs as either new if they were only found in sand-dwellers or ancient if they were also found in the genomes of non-sand-dwellers, including species from other African rift lakes. For new variants found only in sand-dweller species, we marked alleles as derived if they were not shared with the rock-dweller *Metriaclima zebra* reference genome (Appendix B Figure 2a). Such standing genetic variation has been recruited for rapid adaptation in sticklebacks and other cichlid fish species[10, 13]. We used odds ratios from a Fisher's exact test comparing SNP-level allele counts to infer whether pit-digging and castle-building preferentially associated with ancestral or derived

alleles or not (Appendix B Figure 2b). We found that among more genetically diverged SNPs, castle-building species tended to have derived alleles ($p < 0.0001$) while pit-digging species were enriched for ancestral alleles ($p < 0.0001$; Figure 2c). Indeed, increasing F_{ST} was related to greater divergence in odds ratios between derived and ancestral alleles (Appendix B Figure 2c). Furthermore, when comparing the overall distribution of F_{ST} measures, 9% of all SNPs were ancient, but in high F_{ST} variants ($F_{ST} \geq 0.2$), this proportion was elevated to 20% (Chi-squared test, p-value < 2.2e-16; Appendix B Figure 3). These data suggest that standing genetic variation, as well as derived alleles shared by castle-building species, both contribute to the overall genetic architecture of bower behavior.

3.2.3 Allele sharing amongst bower building species may be due to introgression

The above observations, along with low bootstrap support values for some nodes on the whole-genome maximum likelihood (ML) phylogeny (Figure 3-1d), suggest that sand-dweller genomes may have been subject to evolutionary processes leading to species tree contraventions, such as incomplete lineage sorting and introgression. To test this, we constructed ML phylogenies from non-overlapping windows of 10,000 SNPs (1,927 in total)[82]. The resulting local phylogenies demonstrate that a variety of tree topologies are present (Figure 3-3a). Using TWISST[83], a method that measures the "weights" of various tree topologies genome-wide, we found moderate to strong support for a variety of trees including several that group species by bower phenotype (Figure 3-3b; Appendix B Figure 4a-b). Furthermore, support for trees that group by bower phenotype varied across linkage groups with an extremely strong increase in weights on LG 2 and 11, reflecting strong genetic divergence seen via F_{ST} in these regions (Figure 3-3c-d; Appendix B Figure 4c).

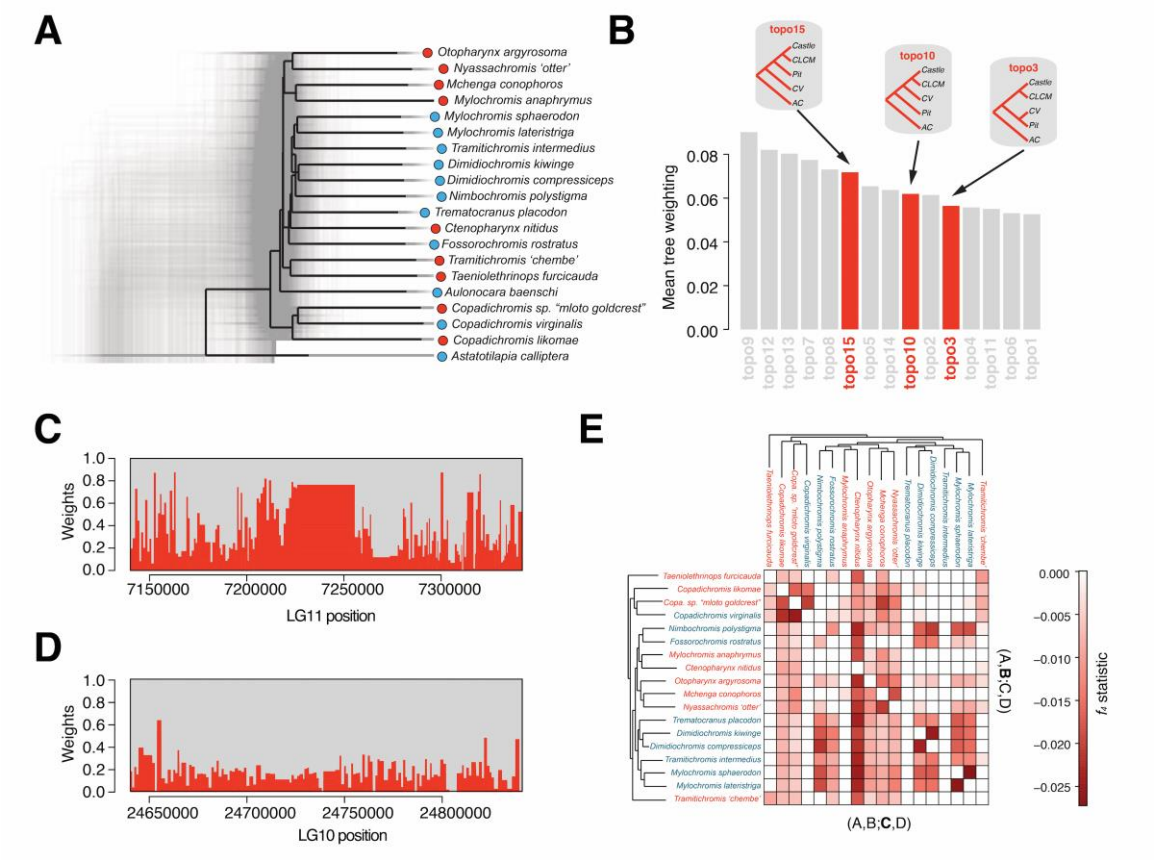


Figure 3-3 Complex phylogenetic relationships among sand-dwelling Malawi cichlids | (A) 1,927 phylogenies resulting from non-overlapping 10,000 SNP windows were plotted using DensiTree. The consensus phylogeny produced by DensiTree is colored black. (B) Barplot of mean genome-wide weightings for the fifteen tree topologies tested with Twisst. Trees grouping clades by bower phenotype (topos 15, 10, and 3) are highlighted. See SI Appendix, Figure S4 for visualizations of all fifteen topologies. (C) Stacked plot of topology weightings along an example region of linkage group 11 with strong support for groupings by phenotype. (D) Example stacked plot of a mixed weight region on linkage group 10. (E) Heatmap of the most significant f_4 values. Species in the x and y axes were either B or C in the form f_4(A,B;C,*A. calliptera*). Pit and castle species names are colored blue and red, respectively. A darker red square indicates more signal of gene flow between the species pairs in the respective row and column.

We next tested for signals of admixture by comparing the observed similarity in allele frequencies among species pairs to those expected given their phylogenetic relatedness (as established by comparisons to pairs of outgroup species)[84]. We found many

species pairs with signals of gene flow that were stronger than expected by chance as evidenced by significantly negative admixture statistics (Figure 3-3e; Methods). Similarly, using TREEMIX[85] we found that analyses of admixture scenarios incorporating all species in our data set also supported multiple admixture events, though the predicted number and specifics of these events could not be confidently estimated (Appendix B Figure 5). We also observed that signals of admixture varied by genomic location and in many cases recapitulated regions of high divergence (Appendix B Figure 6)[12]. Given these species' high degrees of relatedness and the potential of natural selection acting on standing genetic variation, it is difficult to ascertain the importance of introgression in the evolution of bower building. Nonetheless, taken together these patterns suggest that gene flow has occurred across the sand-dwelling clade and may have impacted variants important for bower building. Our observations of complex evolutionary histories reflecting both segregation of ancestral polymorphism and gene flow between species are consistent with findings from other recent studies of African cichlid genome-wide divergence[12, 13]. Specific hypotheses of gene flow between bower building species could be tested by additional population and geographic sampling[86].

3.2.4 Bower building is associated with context dependent allele-specific expression

Behavioral traits can be associated with rapid transcriptional changes and distinct neurogenomic states[87] so we next asked how gene expression was activated in the brains of pit-digging vs. castle-building cichlids. To do this, we assayed whole-brain gene expression (RNA-seq) in interspecific hybrid males to measure allele-specific expression (ASE), an approach that can identify *cis*-regulatory divergence between closely related species[88, 89]. We crossed the pit-digging *Copadichromis virginalis* (CV; sire) with the castle-builder

Mchenga conophoros (MC; dam) based on previous laboratory observations confirming the viability of this cross. Remarkably, CVxMC F_1 hybrids produced an unusual intermediate "pit-castle" bower by carrying out parental behaviors in sequence. First, a pit is excavated for several days to weeks followed by a transition to construction of a castle (Appendix B Figure 7). This observation suggests that both pit-digging and castle-building behavioral control circuits are functional in the F_1 male brain. We took advantage of this sequential bower construction to compare brain RNA-seq data from CVxMC F_1 hybrid males during three behavioral contexts: pit-digging (n=2), castle-building (n=2), or in isolation without conspecifics or sand (control; n=2) (Figure 3-4a). Given that the terminology for behaviors observed in the hybrids overlaps with that used for the pure-species genomic comparisons above, we will from here on denote F_1 hybrid behavior with the suffix "-phase". Sequences were aligned using the *Metriaclima zebra* reference genome and ASE was measured from gene-level read counts.

We found the presence of ASE in the brain transcriptomes of fish in all three behavioral contexts. We identified 621 genes with significant ASE across replicates in at least one behavioral context (Table 3-2; Bonferroni corrected p <0.05). Because many of these genes may be unrelated to bower building, we reasoned that differential ASE (diffASE) across contexts—e.g. cases where one allele is more highly expressed than the other only during the pit-phase or castle-phase—would enrich for those involved in the behavior. We found robust variation in the number of CV- and MC-biased genes between behavioral contexts that, surprisingly, reflected a pattern of species bias (Figure 3-4b). Specifically, significantly more genes are MC-biased during the castle-phase compared to the pit-phase

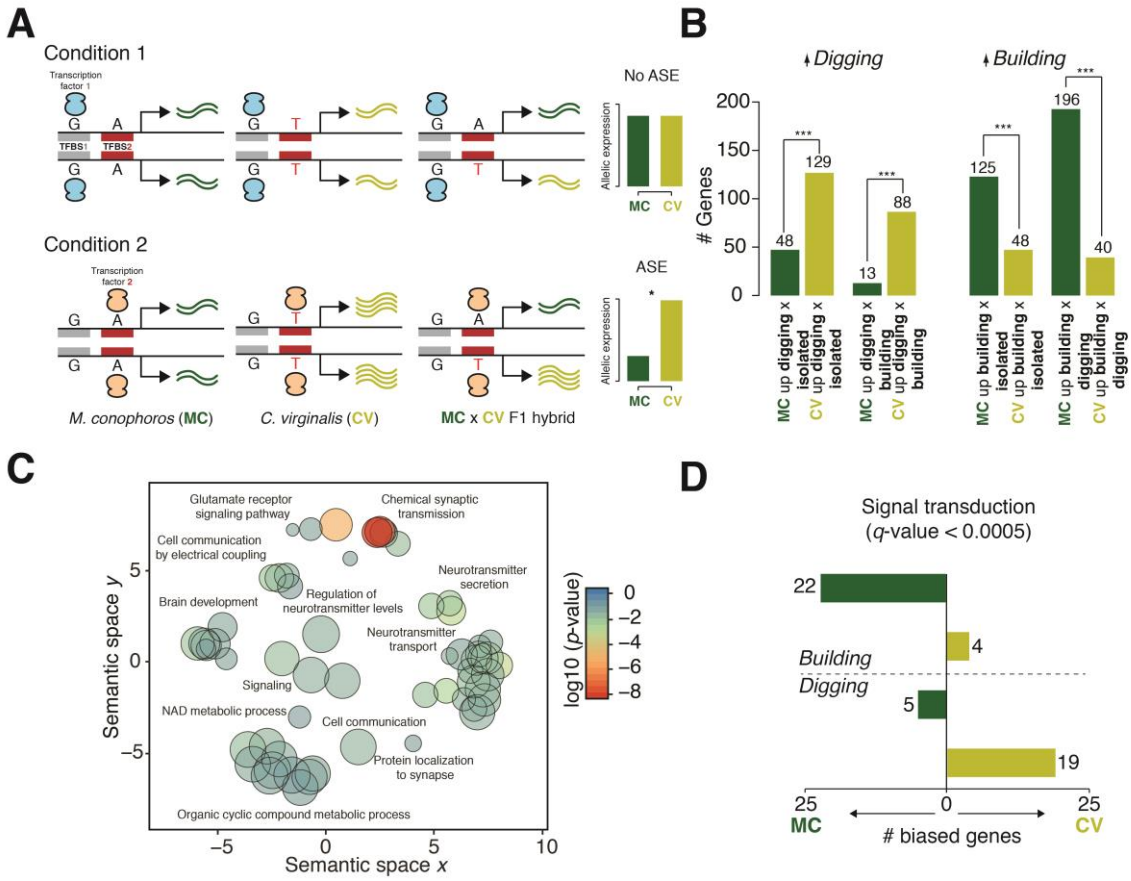


Figure 3-4 Behaviorally-dependent allele-specific expression | (A) Cartoon of allele-specific expression under different contexts. In context 1, the sequence in the transcription factor binding site (TFBS1) is identical between the two alleles, leading to an expected ~50:50 allelic ratio in the F_1 hybrid. In context 2, there is a variant between the species in TFBS2 leading to allele-specific expression (ASE) in the F_1 hybrid. (B) Barplots indicating the distribution of significantly differentially biased genes across building and digging contexts. Significance calculated using a Fisher's exact test (***; $p < 5 \times 10^{-5}$). (C) Semantic similarity of Gene Ontology Biological Process terms enriched for genes with differential allele-specific expression (diffASE). Node size is the log of the size of the category represented. Nodes are colored by the log10(*p*-value) of the enrichment. (D) Example result from a sign test comparing context-dependent allele-specific expression (Signal transduction; Reactome pathway R-HSA-162582).

(counts: 196 MC biased genes, 40 CV biased genes; Fisher's exact $p < 7.45 \times 10^{-19}$) or compared to isolation (counts: 125 MC biased genes, 48 CV biased genes; Fisher's exact $p =$

$1.37x10^{-7}$), while significantly more genes are CV-biased during the pit-phase compared to the castle-phase (counts: 13 MC biased genes, 88 CV biased genes; Fisher's exact $p = 3.03x10^{-11}$) and during the pit-phase compared to isolation (counts: 48 MC biased genes, 129 CV biased genes; Fisher's exact $p = 2.58 \times 10^{-8}$). Furthermore, we identified a number of individual genes with "discordant ASE" because their direction of allelic bias switched between behavioral contexts (Appendix B Figure 8a-b). Notable examples of this phenomenon include the genes *atp1b4* (Digging: CV allele 2.98-fold higher expression; Building: MC allele 10.83-fold higher), an ion pump with brain-specific expression in fish[90], and *dgcr8* (Digging: CV allele 2.55 fold higher; Building: MC allele 2.75 fold higher), a core component in microRNA biogenesis that is required for inhibitory synaptic function[91] (Appendix B Figure 8; Full results in table 3-2).

These patterns of differential allelic expression in F_1 animals indicate an unexpected amount of dynamic genomic regulation associated with behavior. Notably we observed that, within the same brain containing alleles from both parental genomes, castle-building *cis*-regulatory elements are specifically activated during the castle-phase and vice versa for the pit-phase . These results add to a growing body of literature illustrating context dependent transcriptional response with experience or changes in behavior[92, 93] and further suggest that evolutionary differences between species in relation to brain function and behavior may arise from variation in such context-dependent regulation of gene expression.

3.2.5 Context and lineage-specific induction identifies behaviour dependent genes and pathways

We reasoned that if the context-dependent ASE observed above is biologically relevant then genes with shared patterns of allelic induction across contexts should be enriched for similar functional roles. To this end we first identified genes with significant differential ASE (diffASE) –varying ratios between CV and MC alleles dependent on context - between at least two of the three behavioral contexts (Methods; 435 genes). We performed gene set enrichment analysis using all other genes detected in at least one of the three contexts (9,703 genes) as background. This analysis identified a number of enriched categories spanning various aspects of neural structure and function (Table 3-3). For example, genes with diffASE were significantly enriched in a number of neural specific Reactome pathways such as transmission across chemical synapses (R-HSA-888590; *q*-value = 8.13×10^{-6}) and GABA synthesis, release, reuptake and degradation (R-HSA-112315; *q*-value = 5.38×10^{-7}). Enriched gene ontology biological processes categories were predominately related to synaptic function, neurotransmitter regulation and signaling, and ion transport and binding (Figure 3-4c; Table 3-3) while cellular component sets included structures and loci important for neuronal function such as clathrin-sculpted vesicles (GO:0060198; *q*-value = 1.06×10^{-5}) and postsynaptic densities (GO:0014069; *q*-value = 3.46×10^{-5}).

We further refined these tests by identifying genes that displayed significant differential induction of one or both alleles during building or digging behaviors (Methods; *n* building upregulated genes = 171; *n* digging upregulated genes = 174). These genes differ from those with diffASE in that the comparison of interest is the expression of individual

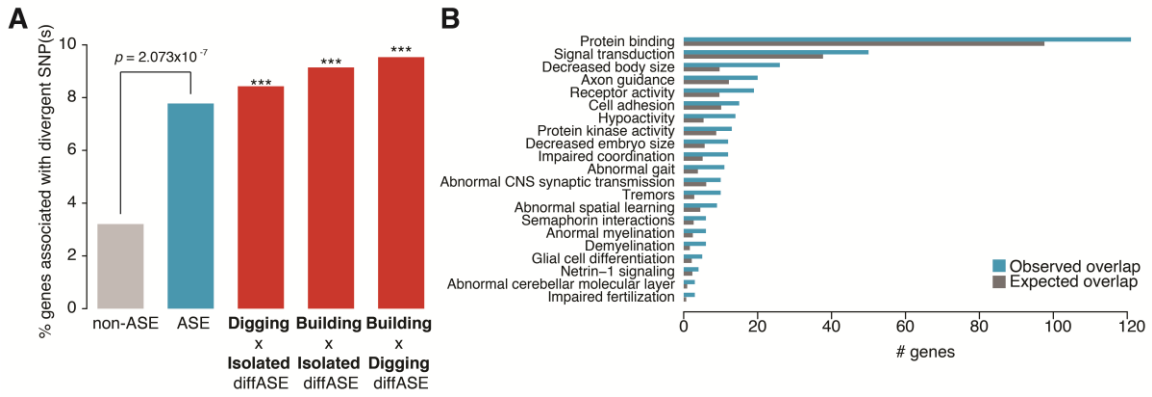


Figure 3-5 Intersection of genome-wide SNPs and ASE | (A) Barplot comparing the number of ASE, non-ASE, and diffASE (across all contexts) genes associated with highly divergent SNPs between pit and castle species. *p*-values computed with a Fisher's exact test comparing genes overlapping SNPs and genes not overlapping SNPs (***; $p < 1 \times 10^{-4}$) (B) Categories in which the observed amount of observed overlap (blue) between genes associated with highly divergent SNPs and with ASE is significantly greater than expected (grey).

alleles *across* rather than *within* contexts. While diffASE is ascertained by identifying different ratios between alleles, differential induction is more similar to traditional differential expression tests in that it is concerned with the expression of individual alleles across contexts. Analyzing these genes might then provide insight into the divergence of context-dependent regulation of alleles between CV and MC. To assay the roles of genes with differential induction we performed gene set enrichment tests as above. Genes upregulated during building were enriched for neurotransmitter release (R-HSA-112310; *q*-value = 4.70×10^{-2}) and ion channel transport (R-HSA-983712; *q*-value = 1.97×10^{-7}). Similarly digging induced genes were associated with ion homeostasis (R-HSA-5578775; *q*-value = 2.90×10^{-2}) and chemical synaptic transmission (GO:0007268; *q*-value = 3.20×10^{-2}). These results indicate that genes with differential ASE and induction are coherently enriched for specific neural processes in comparison to the rest of the transcriptome, adding evidence

to the idea that context-dependent gene regulation may support distinct neural states related to behavior.

We extended this analysis to explore the role of lineage-specific differences in gene regulation via consideration of independent regulation of CV and MC alleles across behavior states. To do so we applied a sign test to the directionality of CV and MC alleles between the pit- and castle-phases [89, 94] (Table 3-2). We compared the allelic counts of individual genes across phases (pit- vs. castle), avoiding a bias for the sampled tissue as opposed to a typical gene ontology enrichment test which would compare the complete focal gene list to a background. We found several hierarchically organized gene sets that matched a pattern of significant differential induction of CV and MC alleles between digging and building (Table 3-3). The identified gene sets were largely involved in cell signaling and communication (example plotted in Figure 3-4d). These observations suggest that lineage-specific selection may have played a role in producing differential regulation of neural signaling genes in CV and MC related to their species-specific bower behavior.

3.2.6 Bower-associated SNPs and cis-regulatory variation

Finally, we asked the extent to which genomic divergence amongst bower building species is associated with *cis*-regulatory changes inferred from the CVxMC intercross. If natural selection had acted on regulatory variants associated with pit-digging and castle-building, it would result in enrichment of highly differentiated SNPs proximal to genes that display ASE. Indeed, we found that ASE genes were significantly enriched near high F_{ST} SNPs (Figure 3-5a; 48/621 ASE genes vs. 332/10221 non-ASE genes; Fisher's exact p =

2.07×10^{-7}). Furthermore, this enrichment increases when considering those genes displaying diffASE (Figure 3-5a). We also detected over thirty pathways that showed significant overlap between ASE and high F_{ST} gene lists (Figure 3-5b; Bonferroni corrected $p < 0.05$; hypergeometric test). As in previous analyses of ASE and F_{ST} alone, many of these pathways are involved in neurodevelopment, neuroplasticity, and behavioral regulation, suggesting concordance between patterns of genetic and regulatory divergence among bower building species.

3.3 Discussion

By combining genome sequencing across many closely related species with analysis of allele-specific expression in the brains of behaving hybrid animals we here provided a genome-wide view of how a complex behavior has evolved. Our results suggest that the evolution of bower building was associated with polygenic selection on old and new genetic variants that regulate genes involved in neural activity and synaptic plasticity in specific behavioral contexts. The observation of context dependent ASE associated with sequential pit-digging and castle-building behavior in F_1 males suggests how *cis*-regulatory divergence across many genes may combine to produce the evolutionary divergence of a complex behavior.

The elevated F_{ST} values at thousands of variants among the 20 diverse bower building species examined revealed that bower building is associated with a complex, yet phylogenetically consistent, genetic architecture. The observation that these sites are both ancestral – polymorphisms shared with species outside of Lake Malawi – and derived parallels similar findings from studies of Malawi cichlids[12, 13, 41] and other recently evolved

species flocks such as sticklebacks [10] and finches[8]. Notably, we found that castle-building species tended to possess derived variants at more genetically diverged sites suggesting that, at least from a genomic perspective, castle-building is the younger, derived behavior. The observation that the species believed to be similar to the common ancestor of Malawi cichlids, *Astatotilapia calliptera*, digs pits and is positioned at the base of the Lake Malawi phylogeny lends credence to this idea (Figure 3-1a). The phylogenetic distribution of bower building suggests that repeated instances of selection, be it on standing variation or introgressed alleles, may have acted to differentially fix the genetic architecture associated with castle bowers in a number of sand-dwelling species. That we detect potential genomic signatures of gene flow supports the notion that introgression may have played a role in the propagation of the derived castle-building behavior among sand-dwelling cichlid species. The importance of gene flow across species boundaries has been highlighted before in cichlid fish adaptive evolution [13, 95], but bower building represents a special case of this general phenomenon, as the behavior is sex-specific and unlikely to increase male survival. It will be interesting to test specific hypotheses of gene flow between particular bower-building species using more targeted sampling and genetic methods.

Given the large regions of increased genetic divergence identified on linkage groups 2 and 11, it is intriguing to consider the possibility that there may be 'supergene'-like elements underlying bower building, similar to those that have been found to be associated with other animal behaviors such as male reproductive morphs in the ruff[96, 97] or insect social organization [98]. Our observation that F_1 hybrids of a cross between pit and

castle species can build both structures, and do so in a mutually exclusive, sequential fashion may further support the notion that the varying bower types require genomic regions working in a modular and independent fashion. Such a finding would not be without precedent among the cichlids of Lake Malawi. The orange-blotch (OB) coloration phenotype, found among >20 species of rock-dwelling species, is associated with a tightly-linked genomic locus resembling a supergene in its size and inheritance and has independently arisen at least three times[38].

Our approach to sequence the genomes of individuals from pit-digging vs. castle-building species has identified numerous genetic variants associated with bower behavior, not unlike the genetic architecture of other complex traits including human neurological disorders[99]. Integrating genome sequencing with RNA-seq from F_1 hybrid brains pairs our strategy with the complexity of the bower trait. Our results fit the general pattern observed in other complex traits: numerous genetic associations with the phenotype[5], and an expectation that many of these variants exert their effects via context dependent cis-regulation of gene expression[27, 100]. Models like Malawi cichlids may thus occupy a 'sweet spot' in complexity - combining a rich genetic, evolutionary, and phenotypic profile with tractable biology – that could reveal novel insights into the origins of behavioral diversity. Of utmost importance will be continued efforts toward correct phenotypic categorization of such complex traits. In this study, we decided to characterize bower building into two qualitative categories and used this definition to perform genome-wide tests across twenty diverse species. Yet our species cohort may share other traits correlated with bower type, adding noise to our measures of genetic association, or may represent more than two behavioral strategies (though this scenario seems unlikely given the behavioral and genetic

findings presented here). In general, the use of bower building as a model complex trait will be benefited by careful work on the roles of ontogeny, intra-specific variation, and behavioral variability in the regulation of this behavior.

The extensive, context-dependent transcriptomic divergence associated with bower building in F_1 hybrids provides intriguing insights into the regulatory basis of behavioral evolution. For example, it appears that bower building is defined by modularity at multiple levels of biological organization. The transition in CV x MC F_1 hybrids from pit to castle bowers may be considered as a shift between distinct behavioral modules associated with distinct behavioral patterns reflective of the respective parental species. The finding that these phases are associated with distinct transcriptomic states suggests that the pit and castle-alleles function modularly based on behavioral context, best reflected by the significant number of genes displaying discordant ASE across these behavioral modules. This is in opposition to other scenarios in which regulatory divergence might be static across behavioral conditions or minimal in the context of brain function and behavior. Instead, the transcriptomic states associated with the pit- and castle-phases in F_1 hybrids appear to be more similar to those found between tissues in an organism, arising from potentially distinct regulatory or epigenomic cellular environments. In this case, given the genomic and evolutionary signatures identified by the whole-genome analyses, it appears that these differences in regulation are at least partly due to extensive sequence-level variation in a number of functionally related regulatory elements associated with behavior. Notably, this may suggest that the genome harbors regulatory loci specifically involved in the dynamic coordination of behavior and brain function analogous to well-known genetic modulators of morphology and that these loci underlie the evolution of behavioral diversity.

3.4 Methods

3.4.1 Bower behavioural measurements

Individual adult, reproductive subject males (*Copadichromis virginalis*, n=4; *Mchenga conophoros*, n=4) were each housed with 1-3 adult, reproductive stimulus females of the same species in 43.2 cm x 91.4 cm x 40.6 cm (160 liters) glass aquariums maintained on a 12:12 hour light:dark cycle. In each tank, a 5.1 cm deep, 35.6 cm diameter tray (Dynamic Design; Newbury Black Poly Saucer, SA1412BK) was placed and filled with sand (CaribSea Inc.; Sahara Sand, 00254). For each subject, 90 to 120 minute videos were recorded between 3-8 hours after lights on during periods of high bower building activity using a GoPro camera (GoPro; Hero4 Silver, CHDHY-401) housed in a waterproof compartment (GoPro; Clear Standard Housing, AHSRH401) and placed top-down directly above the sand-filled tray. Behavior of male subjects was scored using The Observer XT 12 software (Noldus) according to the following definitions: "scooping" was defined as opening of the mouth and collection of sand, and "spitting" was defined as expulsion of sand from the mouth. For each individual, screenshots were captured at the precise moment of every scoop and spit event, and these screenshots were exported to Paint (Microsoft) to extract spatial coordinates of the mouth for every scoop and spit event.

Bower building difference scores were calculated to measure the spatial dispersion of scoops compared to the spatial dispersion of spits for each subject male. To calculate the difference score for each subject, we first determined the coordinates for each subject's average scoop location and average spit location. Using these coordinates, the absolute distance was then calculated from the average spit location to each individual spit location,

and from the average scoop location to each individual scoop location. In order to generate a quantitative metric of spatial dispersion, these distances were then averaged, yielding two distance "scores" for each subject, one for scoops and one for spits. To compare the spatial dispersion of scoops versus spits, the difference between these distance scores (scoop distance score minus spit distance score) was calculated for each subject, thus providing an estimate of differences in spatial patterns of scooping and spitting sand for each animal. A Student's t-test (two-tailed) was used to compare these difference scores between species.

3.4.2 Genome sequencing, alignment and variant identification

We chose diverse representative species of pit-digging and castle-building groups, selected from multiple genera across the phylogenetic tree of sand-dwellers[75]. Genomic DNA was extracted from fin clips of 20 individuals collected in Lake Malawi (11 pit-diggers and 9 castle-builders, Table 1-1) using the Qiagen DNeasy kit (Qiagen Cat # 69504).

Libraries were constructed following the Illumina TruSeq DNA library preparation protocol. Paired-end sequencing (2x100) was performed on the Illumina Hi-Seq2500 at Georgia Tech. Raw sequence reads were quality controlled using the NGS QC Toolkit[62]. Quality control was performed as follows: first, raw reads with an average PHRED quality score below 20 were removed. The remaining reads were further trimmed of low-quality bases at the 3' end. QC reads for each of the genomes were aligned to the new *Metriaclima zebra* reference genome[77] using bwa-mem (version 0.7.4) and default parameters[56].We used Picard Tools (https://broadinstitute.github.io/picard/) to mark PCR duplicates. Sequences were mapped to 87% of the reference genome on average, with mean coverage of 24.31X. Variant discovery and filtering was performed using HaplotypeCaller

within the Genome Analysis Toolkit(GATK) program according to GATK best practice recommendations[101-103].

Deletions were identified using modifications to a previously published approach[104]. Candidate deletions were first identified using 'chimeric' reads as identified by bwa[56] (i.e. reads that included the SA tag) where each alignment mapped to the same contig and the same strand. These reads were used to infer the breakpoints and insertion sequence of a candidate deletion. Candidate deletions that were also present in the sequencing of *M. zebra* were excluded as likely errors in the reference sequence. Each candidate deletion was then genotyped in each of the pit and castle species by collecting all the reads with primary alignments that fell within 10bp of the candidate deletion and a mapping quality score greater than 10. These reads were realigned to both the reference and the candidate deletion sequence using a striped Smith Waterman Alignment from the scikit-bio Python library. Reads were classified as reference, mutant, or undetermined based upon their mapping score to the two alleles. Deletions were categorized as homozygous mutant if 80% or more of the reads were categorized as mutant, homozygous reference if 80% or more of the reads were identified as reference, and heterozygote if they fell in between. Candidate deletions that were reference homozygous in all species were filtered from the dataset.

3.4.3 Tests of genetic divergence and enrichment

We excluded sites with more than 50% missing genotypes from the whole genome sequencing data. We calculated F_{ST} per variant and in 10kb windows using the `--weir-fst-pop` parameter from the VCFTOOLS program[59] with the flags `--fst-window-`

`size 0` for individual sites and `-fst-window-size` 10000 for 10kb windows. Nucleotide divergence was calculated using VCFTOOLS with the `-window-pi 10000` flag. Thresholds were estimated using R-package FDR-TOOL[105]. F_{ST} values were converted to a normalized scale for visualizing these data on linkage groups (Figure 2a) using Fisher's Z-transformation. To compare F_{ST} to a population-structure controlled genome-wide association analysis we employed GEMMA v 0.96[106]. We first computed a kinship matrix using the filtered sites from which F_{ST} was calculated and then performed the association test on bower type using a linear mixed model (-lmm flag) factoring in relatedness via the kinship matrix.

SNP and indel variants were annotated using the SnpEff (4.3i) program[60] and analyzed for functional enrichment using GeneAnalytics (geneanalytics.genecards.org)[43]. Using the binomial distribution, this algorithm tests the null hypothesis that there is no functional overrepresentation. A resulting score is presented for each match in the form of a $-\log_2$ transformed *p*-value corrected for multiple comparisons via the false discovery rate method. Scores are arranged into three significance categories: high (Padj < 0.0001, medium (Padj < 0.05) and low (Padj > 0.05). Semantic similarity plots for significantly enriched categories were produced with REVIGO[107].

3.4.4 Identifying structural variants

To predict structural variants on a genome wide basis, we used BreakDancermax(1.1)[80] with default parameters. As read pair mismatches in size (pairs are farther away than expected) and direction (pairs in the same orientation) can be used to identify

inversions, we also used the Integrated Genome Browser[81] to manually validate predicted structural variants from BreakDancer-max, particularly those on LGs 2 and 11. Thousands of structural variants were predicted for each species, but none consistently associated with the broad F_{ST} peaks on LGs 2 and 11.

3.4.5 Improved genome annotation

In the original NCBI release of the latest *M. zebra* reference genome[77], 15,361/26,490 predicted protein-coding genes were annotated as hypothetical, or without orthologs. To improve this annotation, we identified orthologs for genes in two additional ways: 1) a phylogenetic method using Treefam, a curated database of phylogenetic trees of animal genes and 2) via reciprocal blast against the human genome and 5 fish genomes[108]. The final annotation merged orthologous genes identified by both methods. 1900 hypothetical genes remain.

3.4.6 Assigning SNPs and genome contigs to linkage maps

We used Chromonomer (1.03)[109] to anchor the gap-filled "M_zebra_UMD1" assembly[77] to linkage groups (LGs) using two different genetic maps, both generated via traditional F_2 crosses and genotyped with RAD-seq. First, a genetic map from 160 F_2 from a cross of *Metriaclima zebra* and *M. mbenjii* resulting in 834 markers in 22 LGs and spanning 1,933 cM[110] was used to anchor the M_zebra_UMD1 assembly. This initial anchored assembly was subsequently re-anchored with Chromonomer using a second genetic map. The second genetic map was generated by genotyping 268 F_2 from a cross of *Labeotropheus fuelleborni* and *Tropheops* 'red cheek' resulting in 946 markers in 24 LGs and

spanning 1453.3 cM[19]. BWA mem (version 0.7.12-r1044)[111] was used in both Chromonomer runs to create the input SAM file by aligning respective map marker sequences to the appropriate assembly or intermediate assembly. A minimum of two markers was required to anchor a contig to a particular LG. The resulting FASTA file of the anchored M_zebra_UMD1 assembly was used for subsequent analysis.

3.4.7 Phylogenetic analysis

A maximum likelihood phylogeny was constructed with the variant data using the SNPhylo pipeline[58]. Default parameters were used with an additional flag -M 0.5.

3.4.8 Ancestral Allele Reconstruction

Pairwise whole-genome alignments of *Neolamprologus brichardi* (species belonging to an older radiation from Lake Tanganyika), *Astatotilapia burtoni* (a riverine species found in East Africa around Lake Tanganyika) and *Pundamilia nyererei* (species from a recent radiation in Lake Victoria) were constructed each against the latest Lake Malawi cichlid genome, *Metriaclima zebra* using the last alignment algorithm (*A.burtoni* against *M.zebra*, *P. nyererei* against *M.zebra*, *N. brichardi* against *M.zebra*)[112]. The generated .maf files from the alignment were converted to sam format using the maf-convert script within the last alignment package. Resulting sam files were converted to bam format via SAMtools[64] and used to add ancestral allele information into the vcf file obtained from variant discovery.

3.4.9 Detection of ancient/derived allele enrichment among pit and castle species

To assess biases in the presence of ancient (polymorphic within and outside of Lake Malawi) and derived (polymorphic only among sand-dwelling species) alleles among pit and castle species we first intersected our SNP-level F_{ST} measurements with the lists of ancient and derived SNPs as identified through the ancestral allele reconstruction methods outlined above. We then calculated a *p*-value and odds ratio on allele counts at each SNP using a Fisher exact test. We assessed both the degree of divergence at each SNP via F_{ST} values and the direction of divergence through the odds ratios (ORs) from the Fisher's exact test. For example, at derived SNPs, an OR <1 indicated that the castle-building species tended to possess the derived allele while an OR >1 indicated a bias toward the derived allele for the pit-diggers. For ancient SNPs, an OR <1 indicated that the castle-building species tended to possess the non-*M. zebra* allele (recall that variants were called in relation to the *M. zebra* reference genome) while an OR >1 indicated that pit-digging species tended to possess the non-*M. zebra* allele.

To assess systematic differences in the possession of derived or ancient variants between pit and castle species we performed Fisher's exact tests at various F_{ST} thresholds. These tests were applied to both the proportion of SNPs with an OR>1 and OR <1 for the derived and ancient lists (as seen in figure 2C) in addition to comparing the mean ORs at various *p*-value thresholds (as seen in SI Appendix, Figure S2c).

3.4.10 Analyses of gene flow and incomplete lineage sorting

Maximum Likelihood phylogenies were produced for 10kb genomic bins using the python function phyml_sliding_windows.py created by Simon Martin (available at

github.com/simonhmartin/genomics_general). This resulted in 1,927 trees that were subsequently visualized using DensiTree v2.2.5[82].

We used TWISST[83] to analyze the distribution of phylogenetic topologies across genomic windows. To do so the .vcf file containing genotype information for the pit and castle species was filtered to include only bi-allelic sites in which all species possessed genotypes. Indels were also removed. The function phyml_sliding_windows.py was then used to create ML phylogenies from these variants in windows containing 50 informative SNPs. TWISST was then run on these trees, testing for variation in the tree topologies in the following five clades:

Clade 1 (Pit-diggers): *Trematrocranus placodon*, *Dimidiochromis kiwinge*, *Dimidiochromis compressiceps*, *Tramitichromis intermedius*, *Mylochromis sphaerodon*, *Mylochromis lateristriga*

Clade 2 (*Copadichromis virginalis*): *Copadichromis virginalis*

Clade 3 (Copadichromis castle-builders): *Copadichromis* sp. “mloto goldcrest”, *Copadichromis likomae*

Clade 4 (Castle-builders): *Mylochromis anaphrymus*, *Ctenopharynx nitidus*, *Otopharynx argyrosoma*, *Mchenga conophoros*, *Nyassachromis* ‘Otter’

Clade 5 (*Astatotilapia calliptera*): *Astatotilapia calliptera*

3.4.11 Four population tests

TREEMIX was run on genotype counts using the settings `-k 2000` and `-m 1,2,4,6,8,10` to assess support for admixture events among bower building species. We extended this analysis by computing the f4 statistic for every possible four population combination using the fourpop function in TREEMIX with the setting –k 500. *p*-values were calculated for every four population comparison from the reported z-scores and adjusted using Bonferroni correction. In order to use *A. calliptera* as an outgroup in the detection of possible gene flow among sand-dwellers, results for just the combination (A, B; C, *A. calliptera*) were extracted. This led to 14,536 comparisons, of which 3,706 had significant (Bonferroni corrected $p < 0.05$) f4 statistics (Figure 3-3e; Table 3-4). The most significant comparisons for each species pair where the species were (A, B;C,D) and (A,B;C,D) were then collected.

We used the fd statistic to identify patterns of possible introgression across the genome. The fd statistic functions similar to f4 in that it compares genotype frequencies between four populations but, whereas f4 is a genome-wide measure, fd can be calculated locally within genomic windows and therefore allows for the detection of genomic regions with particularly strong signals of introgression19. We calculated fd using the python function ABBABABAwindows.py created by Simon Martin (available at github.com/simonhmartin/genomics_general) over 10kb windows containing at least 50 informative SNPs for the four population groups identified as most significant from analyses of the f4 statistic.

3.4.12 RNA Sequence library construction

To assess the allele specific expression, whole brains were obtained from the following animals:

1 *Copadichromis virginalis* (CV) male; digging (Sire of all analyzed F_1 hybrids)

2 CV x *Mchenga conophoros* (MC) F_1 hybrids; digging

2 CV x MC F_1 hybrids; building

2 CV x MC F_1 hybrids; isolated

For these experiments individual males were housed with 3-5 conspecific females and allowed to develop territories and initiate bower construction. We confirmed that bower behavior was being performed reliably (consistent for >24 hours) and on the evening before the experiment we separated focal males from females using a transparent divider and flattened the bower. At lights on the next morning the barrier was removed and the males' behavior observed. Males were sacrificed via decapitation 30 minutes after the initiation of consistent behavior and, to prevent mRNA degradation, brains were dissected into RNAlater (ThermoFisher Cat. #AM7020) less than 10 minutes after sacrifice. Whole brains were homogenized in TRIzol (ThermoFisher Cat. #12183555) using a pestle. RNA was isolated using a Qiagen RNeasy mini kit (Qiagen Cat. # 74104). RNA-seq libraries were constructed using Illumina TruSeq kits, following manufacturer protocols. All libraries were sequenced as multiplexed samples in one lane of an Illumina HiSeq 2000. Two biological replicates per context were chosen for analyzing allele-specific expression (ASE) following methods from previous publications[89] that found a similar sample size was sufficient for reliably detecting allelic biases from RNA-seq data

3.4.13 RNA-seq alignments and SNP calling

RNA-seq read quality was assessed using FastQC (bioinformatics.bbsrc.ac.uk/projects/fastqc/). Illumina adapters were removed using SeqPrep (github.com/jstjohn/SeqPrep). We obtained the *M. zebra* genome assembly and annotations from NCBI RefSeq (Assembly accession: GCF_000238955.2; Assembly name: M_zebra_UMD1). RNA-seq reads were aligned to the *M. zebra* genome using STAR 2.4[113] with the options `-- SortedByCoordinate`, `--outSAMattributes MD NH NM`, and `--clip5pNbases 6`.

Read groups were added with AddOrReplaceReadGroups.jar in Picard Tools 1.92 (github.com/broadinstitute/picard). The resulting bam files were sorted using SAMtools[64]. Duplicate reads were marked using MarkDuplicates.jar in Picard Tools. We then applied GATK 3.3 indel realignment and duplicate removal and performed SNP and INDEL discovery using UnifiedGenotyper following the suggested GATK Best Practices[101-103].

We filtered the resulting .vcf files to identify all heterozygous sites in the F_1 hybrid samples with quality scores greater than 30. We also produced a list of all homozygous sites in the CV parental sample with quality scores greater than 30. To allow proper phasing of heterozygous sites we filtered the F_1 hybrid list to just sites that intersected with the CV homozygous SNPs.

3.4.14 Detection and quantification of allele-specific expression (ASE)

The *M. zebra* reference genome was masked at high-confidence heterozygous sites using the perl script MaskReferencefromBED.pl (github.com/TheFraserLab/ASEr). To control for reference bias all hybrid samples were then re-aligned to this masked reference

using the same STAR options as above. Duplicates were marked using MarkDuplicates.jar in Picard Tools and the bam files were sorted using Sam Tools. SNP-level ASE was then calculated with the python script CountSNPASE.py (github.com/TheFraserLab/ASEr).

All downstream ASE analyses were conducted using R version 3.2.3[114]. After SNP-level ASE was calculated we filtered for >5 counts per allele for every gene within each sample. In order to conduct gene-level analyses the subset of SNPs that met this expression cutoff were then summed without normalization into gene-level counts using the gene coordinates in the ref_M_zebra_UMD1_top_level.gff3 annotation (NCBI *M. zebra* Annotation Release 102). For all genes in each sample ASE was calculated by taking the $\log_2$ ratio of the gene-level CV allele counts over the gene-level MC allele counts. After filtering and summing into genes we investigated the distribution of ASE ratios across all genes for each sample. Distributions consistently skewed toward either allele could be potentially indicative of biases in read alignment or other technical artifacts. Analysis of ASE ratio distributions showed each to be roughly normal and centered around a $\log_2$ ratio of zero, indicating a lack of evidence for strong bias toward either species' allele across samples (Figure S9).

We next calculated significance of ASE per gene. Since allelic counts from RNA-seq data are prone to overdispersion we identified significant ASE using a beta-binomial test comparing the CV and MC counts at each gene with the R package MBASED[115] (1- sample analysis; default parameters; run for 1,000,000 simulations). The resulting *p*-values were adjusted for multiple tests using Bonferroni correction. For a gene to be considered significant within a context we required that both replicates possessed Bonferroni

All *p*-values < 0.05 and that the direction of ASE (either CV biased or MC biased) was the same between replicates.

3.4.15 Identifying differential allele-specific expression (diffASE)

Differential allele-specific expression (diffASE) and differential allelic induction were identified using the 2-sample analysis in MBASED[115] (default parameters; run for 1,000,000 simulations) which, like in the 1-sample analysis used to assay ASE, employs a beta-binomial model of read counts to control for over dispersion. MBASED was used to compare all possible pairings of the three behavioral contexts (digging, building, and isolated). This produced the pairings: digging x building, digging x isolated, and building x isolated. Furthermore, since MBASED tests for significance in only one of the two contexts at a time, we also ran all three comparisons in their reciprocal directions (building x digging, isolated x digging, and isolated x building). To identify diffASE, the expression of the CV and MC alleles within each context were compared (i.e. CV $\text{allele}_{\text{context1}}$ vs. MC $\text{allele}_{\text{context1}}$ compared to CV $\text{allele}_{\text{context2}}$ vs. MC $\text{allele}_{\text{context2}}$). Significant cases represent scenarios in which the ratios between the alleles diverge between contexts (as represented in figure 4a). Differential induction was assayed by testing for significant variation between CV and MC alleles across contexts, using the ratio of each allele to itself for the test of significance (i.e. CV $\text{allele}_{\text{context1}}$ vs. CV $\text{allele}_{\text{context2}}$ compared to MC $\text{allele}_{\text{context1}}$ vs. MC $\text{allele}_{\text{context2}}$).

The resulting *p*-values for each pairing were combined across replicated using Fisher's method and adjusted with Bonferroni correction. We then compared these combined and adjusted *p*-values across the reciprocal contexts (e.g. digging x isolated vs. isolated x digging) and selected the lowest *p*-value for downstream analyses.

3.4.16 Gene set enrichment tests

For the enrichment analyses of diffASE genes and genes with allelic induction in building and digging the corresponding lists were produced as well as background sets corresponding to all other genes detected in one of the three behavioral contexts. Gene enrichments were calculated with PANTHER[116] and filtered for a *q*-value < 0.05. The allelic induction lists were produced in a similar method to those for diffASE by using MBASED but differed in that instead of comparing alleles within contexts, the analysis was run on alleles across contexts (e.g. comparing CV allele expression in digging vs. building). The resulting *p*-values would then reflect differential induction of each allele across pairs of contexts. For the gene set enrichment tests, genes were only selected that had significant induction (Bonferroni corrected *p*-value < 0.05) of at least one allele in either the building or digging contexts alone. Semantic similarity plots for significantly enriched categories were produced with REVIGO (45).

Lineage-specific variation was assayed by comparing the number of genes enriched within gene sets that possessed differential allelic induction of either the CV or MC alleles during building or digging. To do so a .gmt file containing gene sets from the human KEGG, GO, Msigdb, NCI, IOB, NetPath, HumanCyc, reactome, and Panther databases

was downloaded from the Bader lab website (http://download.baderlab.org/EM_Gene-sets/) in April 2018. Genes were then assigned to categories and a Fisher's exact test was performed on a 2x2 contingency table in which the rows represented digging and building and the columns were the CV and MC alleles. To limit the number of tests performed we ran the Fisher's exact test on each ontology independently required each gene set to have 20 genes represented. The resulting *p*-values were then adjusted for multiple tests using Bonferroni correction.

CHAPTER 4. DISCUSSION

4.1 Conclusions

Lake Malawi cichlids are an excellent model for studying the genetic basis of phenotypic diversity. I have leveraged the strengths of the system to answer fundamental questions about the origins and nature of the genomic variation that gives rise to the phenotypic diversity in the Lake Malawi cichlid assemblage. Malawi cichlids have extremely low genotypic divergence coupled with extremes of phenotypic variation. We have the ability to make interspecific hybrids and assay for the effects of the parental alleles in F_1 and F_2 genetic environments. These allow for powerful experimental designs backed by a large dataset of whole genome sequences I generated and aligned to a reference species within the lake.

The colonization of Lake Malawi follows the classic 3 stages model of adaptive radiation. First the cichlids diverge along the axis of habitat - rock and sand - then diversify along trophic levels to generate diversity in feeding apparatus and strategies followed finally by diversification along the axis of communication giving rise to the large variety of mating strategies color patterns and body morphs[9]. Using representative species from the cichlid assemblage, we have looked deeper into the genetic basis of phenotypic divergences along these axes of evolution.

The habitat differences of rock and sand dwelling cichlids are reflected in the wide range of associated phenotypes that define the two lineages. Chief among them are the fundamental differences in behavior. Rock species tend to be territorial and breed all year

round. Sand species tend to be seasonal and have defined bower building and lekking behaviors. Just like the genomic divergence defining beak shape diversity in Darwin's finches is spread throughout the genome[8], the peaks of divergent SNPs and InDels that segregate rock and sand are distributed on all linkage groups. The divergent variants between rock and sand are highly functional and are associated with gene sets that are enriched for pathways associated with early development and behavior. The observable behavioral differences between rock and sand species are associated with brain development and adult behavior, a scenario encapsulated by Sydney Brenner's [25] dual encoding problem of behavior. Early embryonic development patterns lead to differences in brain structure, which in turn leads to differences in neuronal expression patterns in adult behavior. Early and late stage embryos of rock and sand parents as well as hybrids indexed for the parental allele show clear spatial and temporal patterns that lead to differences in the relative sizes of the telencephalon and eye fields. Adult F_1 males across five independent and distinct rock X sand crosses cluster transcriptomes according to social context rather than cross identity.

According to the 3 stage model of adaptive radiation, a divergence along the lines of habitat is followed by divergence along trophic levels followed by communication. These levels aren't discretely defined along the axis of time to the exclusion of each other. An organism colonizing a new environment merely prioritizes diversification along a certain axis at different stages of evolutionary radiation. A rock lineage divergence, for instance, is also characterized by evolution of novel craniofacial morphology and different colors for camouflage and mating. These differences are *elaborated* in subsequent diversification along the process of colonization. A large sampling of species along the more fundamental rock-sand divergence will include within each lineage differences that evolved

after, within each lineage, along the axes of trophic levels and communication. As such it is difficult in the dataset described here to define the genotypic divergence in the case of craniofacial traits like development of neural crest cells for example as limited to rock sand divergence or a combination of signal from the rock sand divergence and the subsequent divergence along the lines of trophic levels. A detailed look at all the phenotypic differences between my choice of representative rock and sand species could resolve these questions and define each phenotypic variation with the myriad genotypic variants. A subset of the species considered here can be used in conjunction with other species that clearly diverge along the second level of evolutionary radiation and a comparison of the variants that segregate habitat differences versus those that segregate trophic adaptations should yield interesting patterns of the relative contribution of different evolutionary forces in shaping an extant lineage.

Within the sand dwelling lineage, moving to the third, most recent axis of adaptive divergence, we see males from many species construct typical bowers in the sand to attract females[75, 117, 118]. Many studies of young radiations of divergent phenotypes evolving in the face of gene flow have shown patterns in the genome of "islands " of diversification as soon in hooded versus black carrion crows[11], divergent cichlid ecomorphs in the small crater lakes in East Africa[12] among others. From the many bower building cichlids within Lake Malawi, we chose typical species from two distinct bower types. Pit diggers, typified by *Copadichromis virginalis* dig pits in the sand while castle-builders, typified by *Mchenga conophoros* construct sand-castles in the sandy lake bottom. The divergent SNPs and InDels are clustered around two prominent peaks on LG2 and LG11 along with smaller

peaks in other regions. Genes associated with the divergent variants are enriched for pathways associated with brain and behavior function. Using reference genomes from cichlid species outside of the sand dweller lineage, I show that ancestral variation has been retained and reused in the evolution of bower building. There is also evidence for gene flow and introgression within the sand dwelling lineage. Pit X Castle F_1 hybrid males, interestingly, show both parental behaviors separated by time. Allele specific expression analysis on behaving F_1 male brains provide evidence for cis-regulatory elements involved in castle building and pit digging. Context specific activation of digging or building gene sets that were enriched for categories of neural function indicates a role for lineage specific selection on these sets of genes.

Complex traits, like behavior, have complex genetic architecture[5]. Using a robust data set of whole genomes sequenced at high coverage I have used comparative genomics to isolate the variation associated with specific and typical behavioral differences in divergent lineages. Using targeted experiments to further define the role of these variants revealed interesting genes and pathways associated with the behavioral differences as well as the evolutionary mechanisms involved in the divergence of these lineages.

The 28 genomes sequenced at high coverage were chosen on the basis of maximum diversity epitomizing the Lake Malawi cichlid variation whenever possible. Although the approach treats divergent lineages as populations identify differentiating genes, the 28 genomes represent 28 species and not dichotomous populations. I operate under the assumption that a divergent variant present in the individual male representative of its species in many similar species is present at high frequency in each population. Rapid advances in genome sequences will make population level assays a reality in the near future and this

assumption can be tested. While the pit castle divergence explicitly looks at male behavioral differences, the major behavioral differences in the rock-sand lineage also implicitly concern males. All the individuals sequenced were males as were the brains analyzed for transcriptional differences. Female mate choice clearly plays a role in the evolution of differences in mating strategies and a wider population sample would include both sexes and account for sex linked and sex limited traits that diversify in the rock-sand and pit-castle lineages. This is an especially thorny issue with the pit-castle divergence where sexual selection plays a strong role in defining the divergence of bower types. Aligning the genomes to a newer reference genome resolves the peaks on LG2 and LG11 into one big peak on LG11. The peak of divergence on LG11 shows a signature of a characteristic inversion in the genome, indicating a putative sex determining locus. Bower building in males and female preference for bower types both could be sex-linked traits that diverge on the putative sex determining locus. Finer scaled genomic assays sampling the bower associated phenotypic variance in males and females should resolve the role of sex linkage in the evolution of this genomic signature on LG11.

Another assumption that underlies some of the inferences I draw above is an adaptive one. Starting with a scan that defines all divergence as involved in adaptive divergences ignores non adaptive forces that may have shaped a genome. Indeed, the distributed genomic variants across the genome for older divergences like the rock-sand split indicate some of these divergent variants could either be due to genetic drift or due to non-uniform mutation rates on different parts of the genome. I show that a majority of these variants are associated with an adaptive signal and associated with functional variation. We also have evidence from other experiments not shown here that genomic divergence for the rock-

sand and pit-castle genomic variants does not correlate to measure of nucleotide diversity. The adaptive nature of variants near genes of interest also rely on a reliable annotation of a high quality reference genome. Cichlids are a nontraditional model organism with a relatively new reference genome that is constantly undergoing improvements. With the work undertaken by the Streelman lab and the East Africa cichlid community as a whole, we will have detailed and fine scaled data in the future to address these questions that I have encountered and managed to only indirectly address.

We have an extremely information rich data set to apply the strategies laid down in this thesis for other traits. Following the three stages model of adaptive radiation, cichlids have also diversified along the axis of trophic morphology. We have the expertise in the lab to delve deeper into the evolution of diversity in tooth and jaw morphology, tooth shape and feeding strategies. We are capable of setting up high throughput behavioral assays in the lab asking detailed questions of fine scaled behavior evolution. New advances in sequencing technologies like single-cell sequencing allows us to tease apart the neuronal architecture and expression patterns therein associated with complex behavior. All of these new directions of research in the lab are greatly aided by having a convenient resource of a substantial number of individuals sequenced at high coverage. We have already started work on a number of these questions building on the platform laid out by the work in this thesis. Complex polygenic traits have complex mechanisms that need extensive multi-disciplinary approaches to solve. I have built upon the existing knowledge base in the Streelman Lab and have set up a foundation so that we can answer interesting questions going forward.

4.2 Publications

A list of the publications that represent the work I have as presented in this thesis:

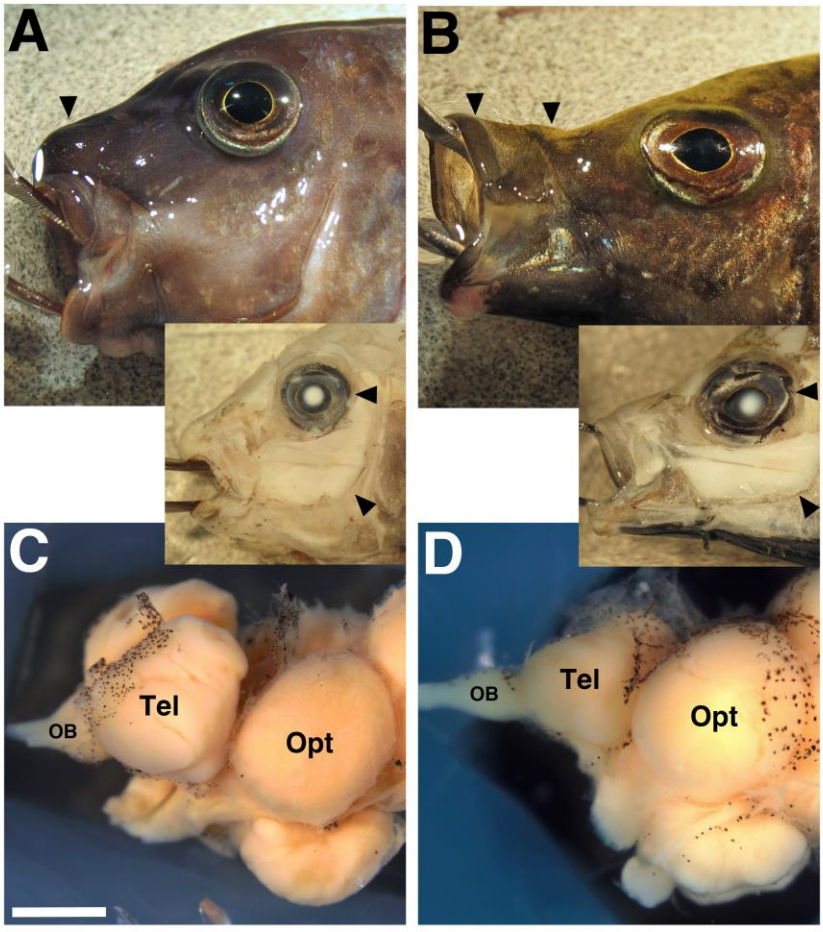
Patil C, Sylvester JB, Abdilleh K, Malinsky M, Norsworthy M, Pottin K, Bloomquist RF, McGrath PT, Streelman JT (*in prep*) **Genome-enabled discovery of functional variants in brain and behavior**

York RA†, **Patil C**†, Abdilleh K, Johnson ZV, Conte MA, Genner MJ, McGrath PT, Fraser HB, Fernald HB, Streelman JT (2018) **Behavior-dependent cis regulation reveals genes and pathways associated with bower building in cichlid fishes** , *Proceedings of the National Academy of Sciences* 201810140; DOI: 10.1073/pnas.1810140115

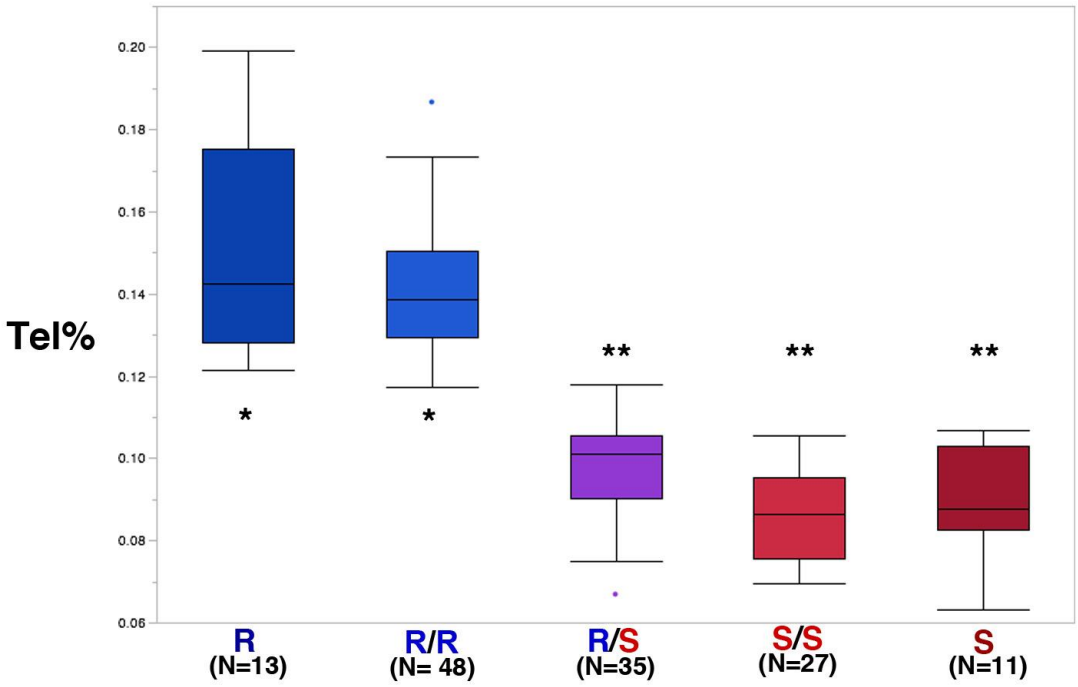
York RA, Byrne A, Abdilleh K, **Patil C**, Streelman JT, Finger TE, Fernald RD (2018) **Behavioral evolution drives hindbrain diversification among Lake Malawi cichlid fish** (*in review*)

York RA, **Patil C**, Hulsey CD, Anoruo O, Streelman JT, Fernald RD (2015) **Evolution of bower building in Lake Malawi cichlid fish: Phylogeny, morphology, and behavior**. *Frontiers in Ecology and Evolution*, 3-18.

APPENDIX A. SUPPLEMENTAL INFORMATION FOR CHAPTER 2

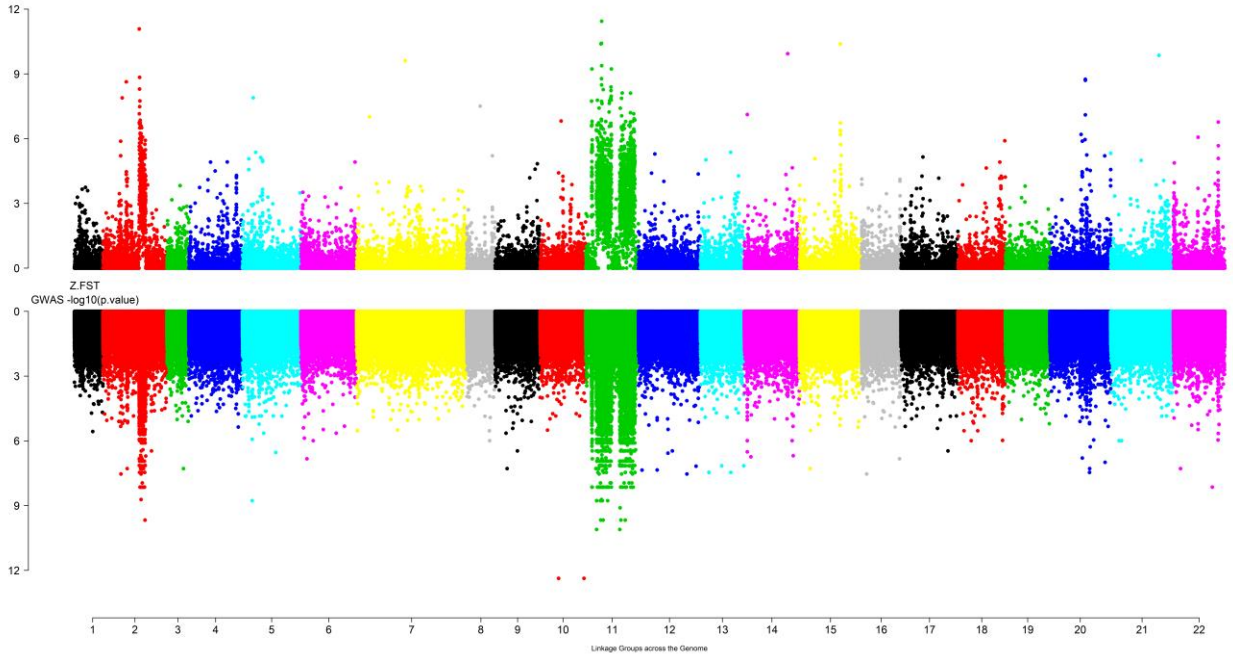


Appendix A Figure 1 **: Differences between in rock and sand** | Jaw shape differences between rock dweller (A) and sand dweller (B). Rock dweller eyes are smaller . Relative sizes of the optic tectum and the telencephalon in adult Rock dwellers(C) clearly different in the Sand dweller (D) which has a larger optic tectum.

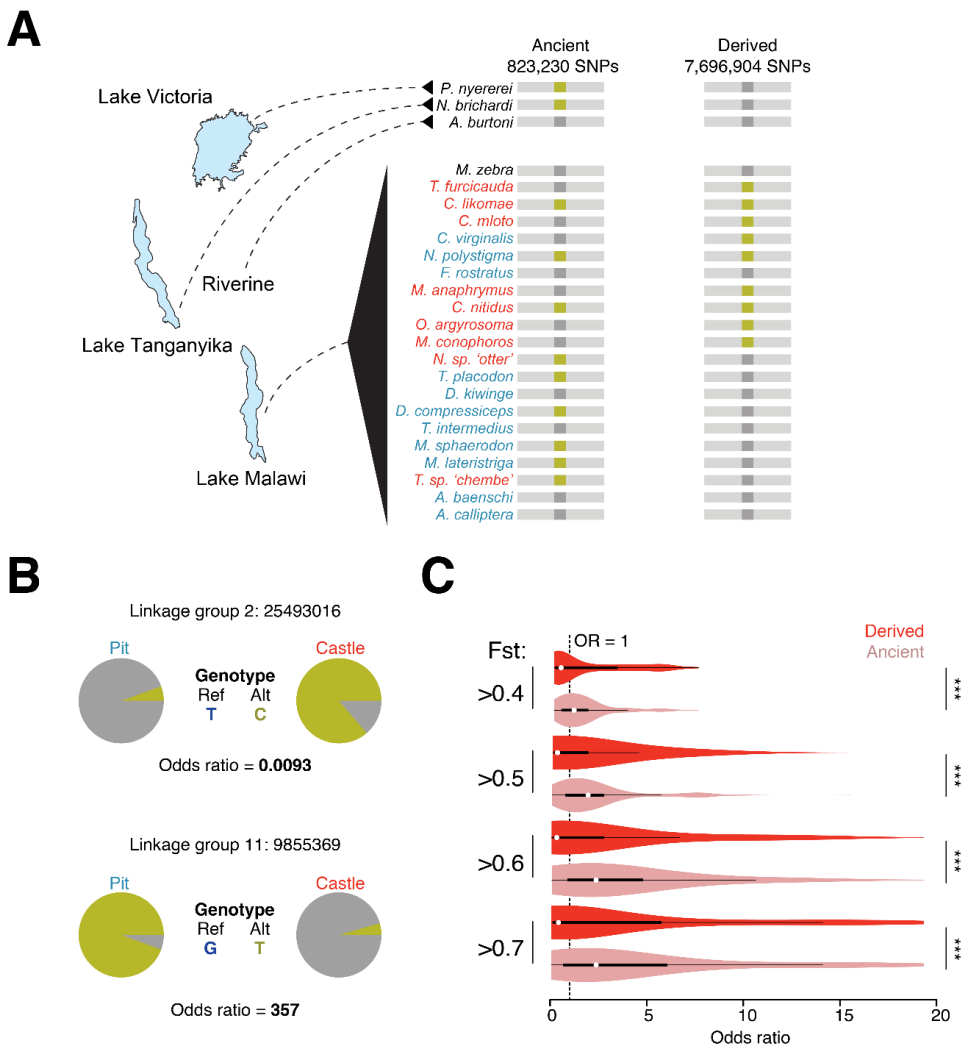


Appendix A Figure 2: **Telencephalon Differences in Rock Sand and Hybrids** | Grown F_2 individuals indexed for the *irx1b* allele compared to parental individuals for volumetric size of the telencephalon relative to the whole brain.

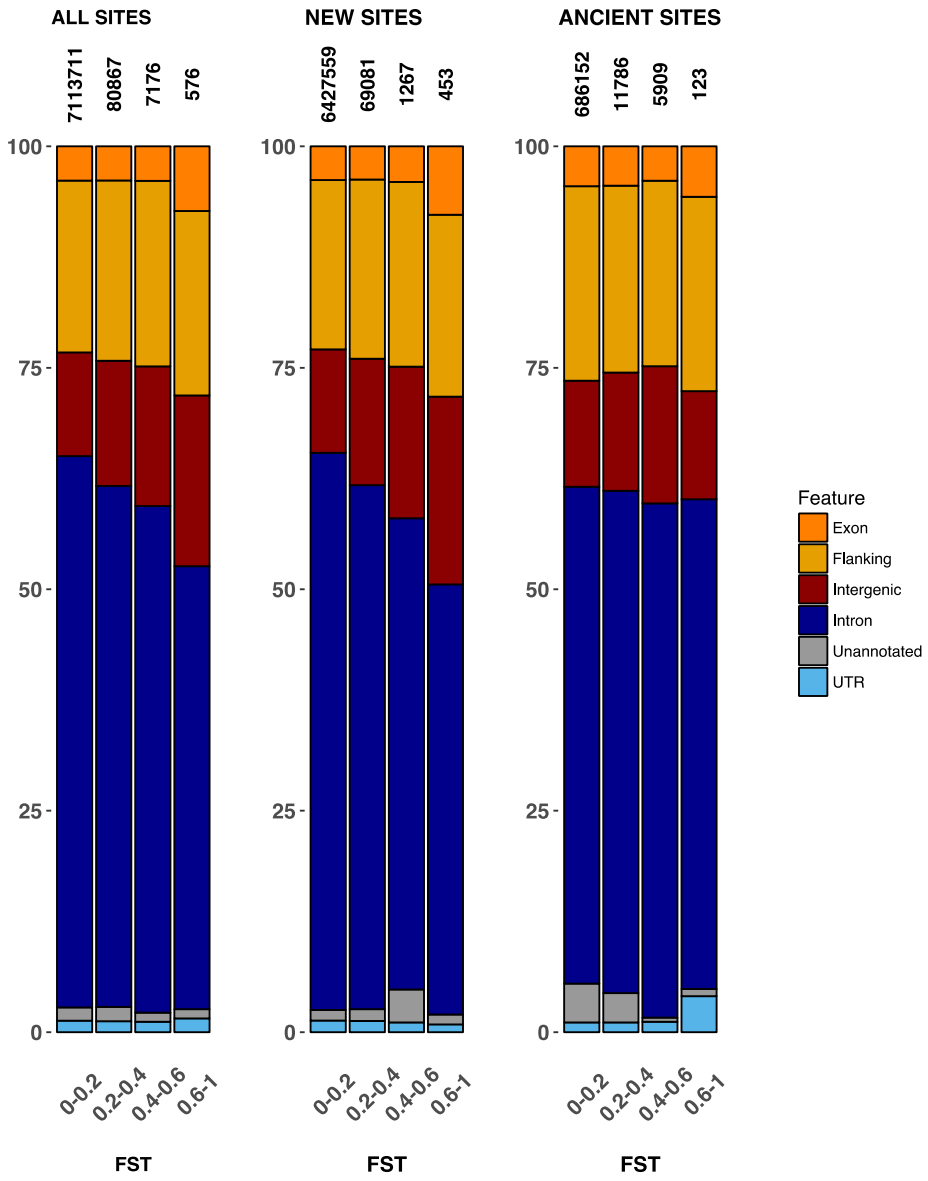
APPENDIX B: SUPPLEMENTAL INFORMATION FOR CHAPTER 3



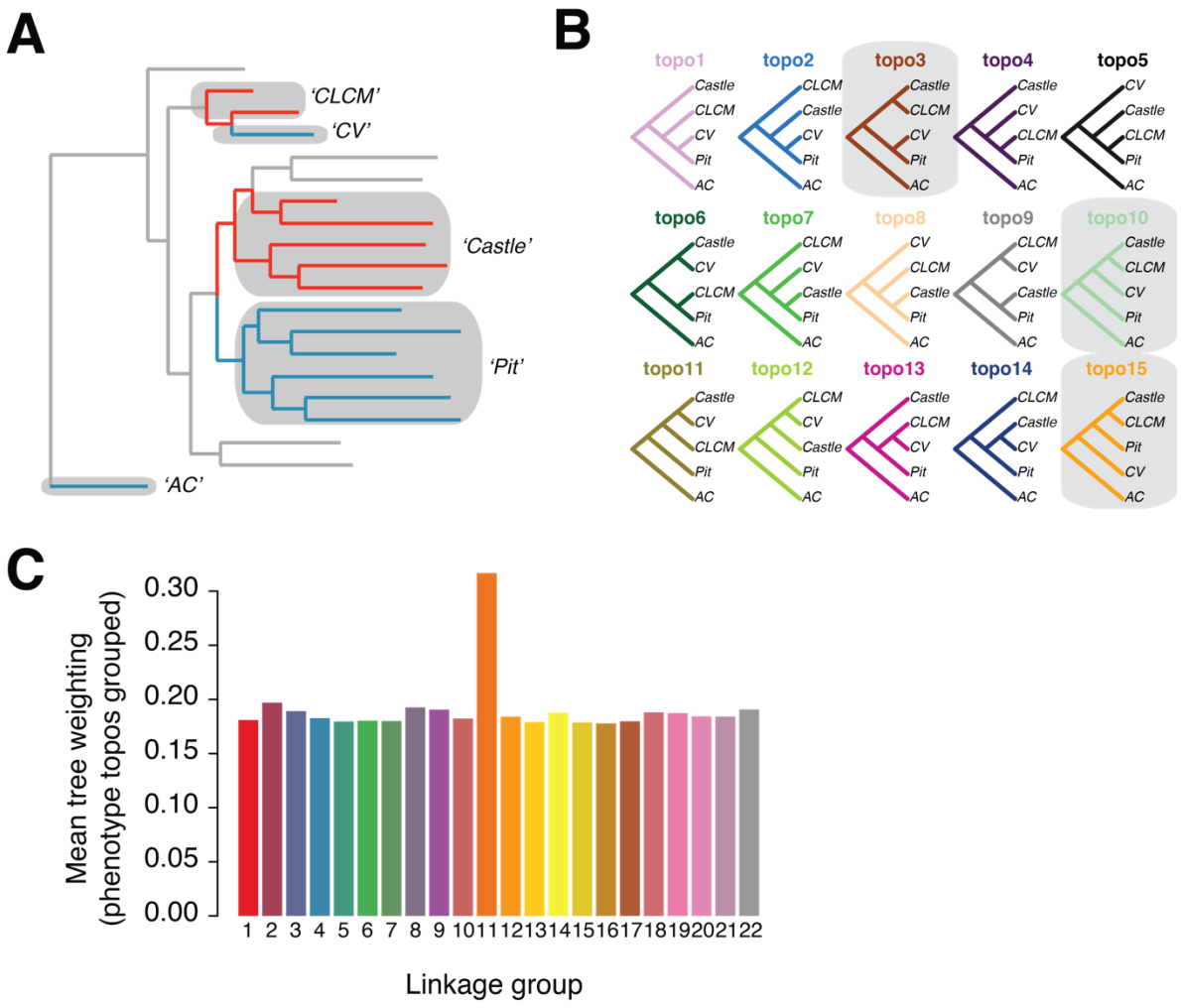
Appendix B Figure 1 : **Comparison of genetic divergence and association patterns across the genome.** Independent measures of pit versus castle divergence, F_{ST} [z-transformed F_{ST}] and ancestry-corrected GWAS [-log10 (pvalue)], across the genome show highly similar patterns.



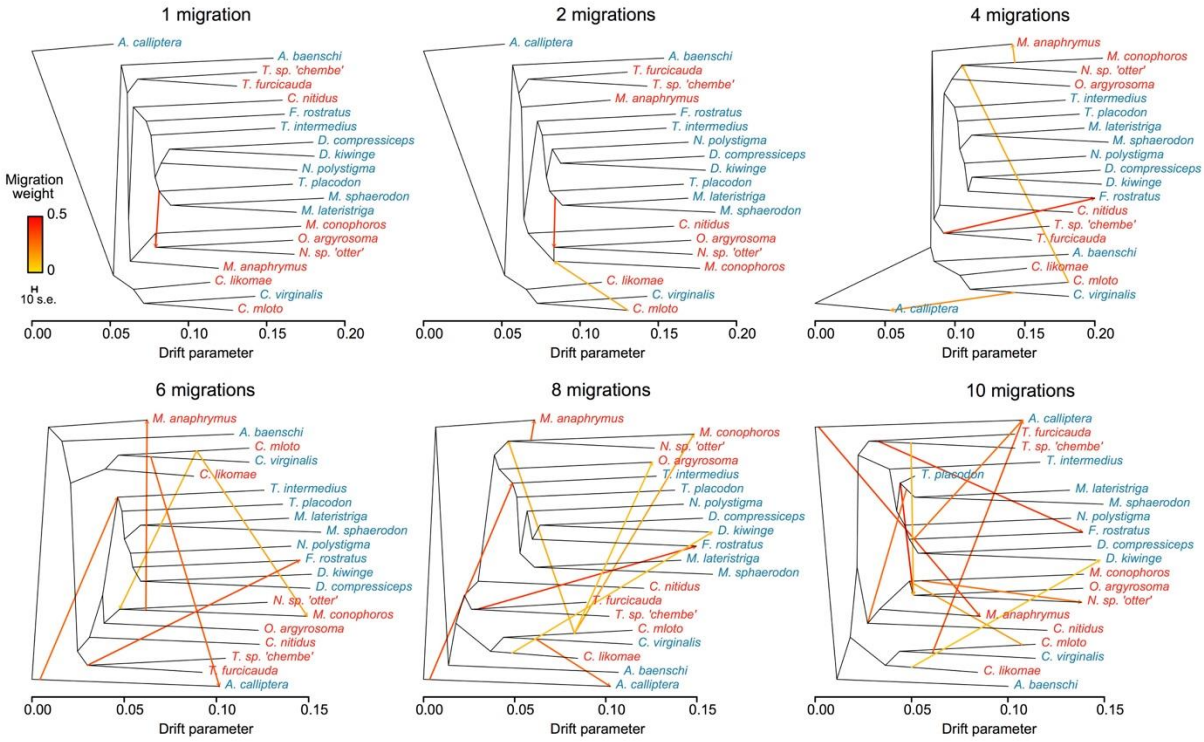
Appendix B Figure 2 :**Genome-wide ancestral and derived SNP enrichments | (A)** Cartoon describing the identification of derived and ancestral SNPs through whole-genome alignment with 5 non-sand dweller genomes **(B)** Representative high odds ratio and low odds ratio SNPs as identified by a GWAS on bower type. GWAS across the genome mirrors patterns of F_{ST} [Figure S1] **(C)** Violin plots of the GWAS odds ratio for SNPs at increasingly stringent *p*-value cutoffs, divided into derived (red) and ancestral (pink) groupings. *** = Kruskal-wallis p <0.0001.



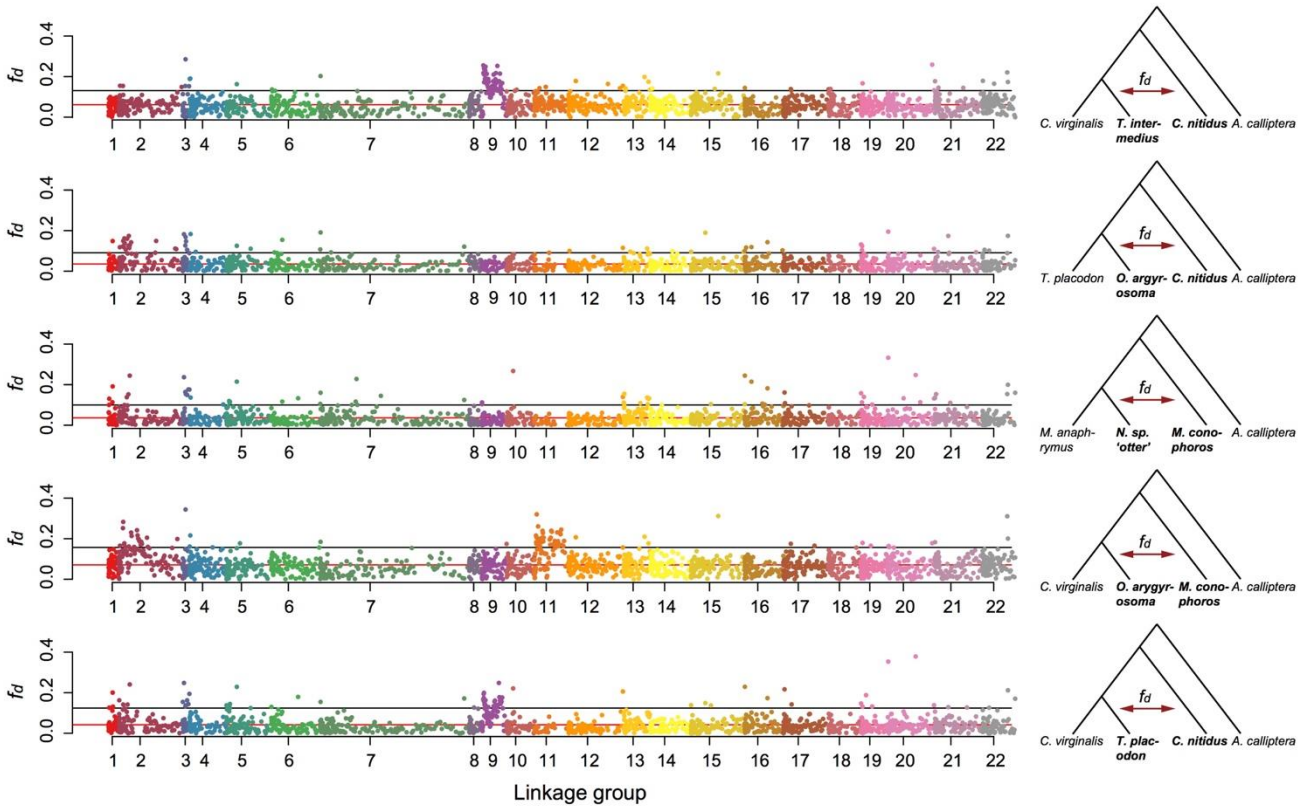
Appendix B Figure 3 : **Genomic distribution and FST of new and ancient SNPs |** Bars indicate proportion of genomic features represented by SNPs binned by FST values for new SNPs (polymorphic only within Lake Malawi), ancient SNPs (polymorphic within and outside of Lake Malawi, and all SNPs.



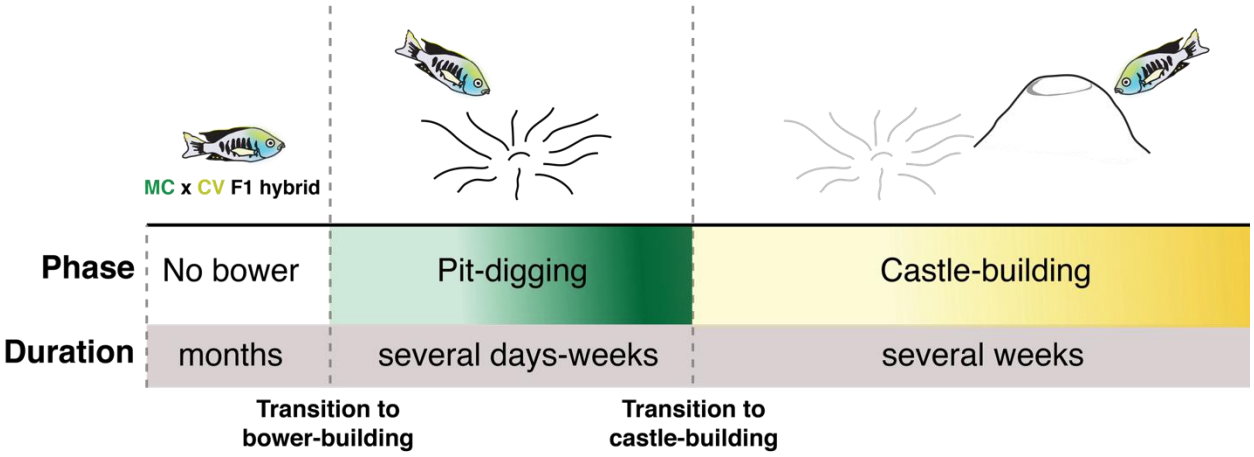
Appendix B Figure 4 : **Topology weighting with** TWISST | **(A)** Phylogeny of the clades used in the Twisst analyses. Branches are colored red and blue for pit and castle species, respectively. **(B)** The 15 topologies ("topos") weighted by Twisst with the phylogenies that group by bower phenotype highlighted with grey backgrounds. **(D)** Barplot of mean combined tree weightings for the three 'phenotype' topos (topo3, topo10, topo15) as binned by linkage group.



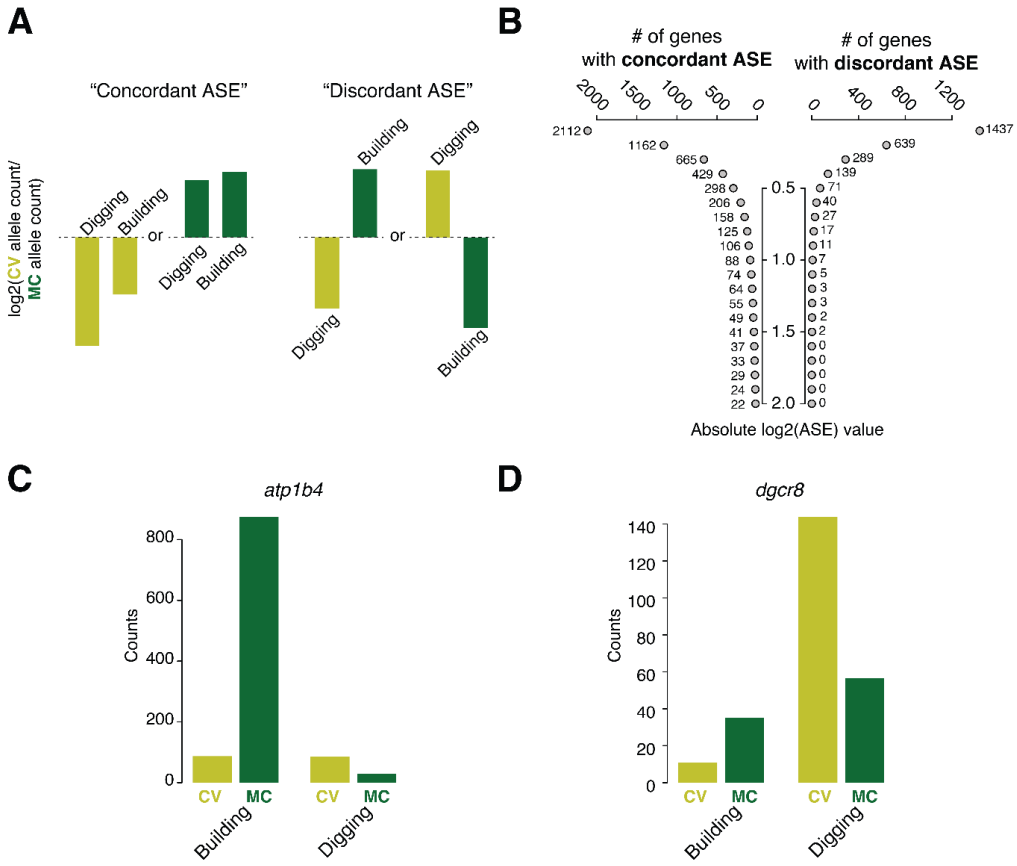
Appendix B Figure 5 : TREEMIX **scenarios |** Phylogenies are plotted with migration edges for 1, 2, 4, 6, 8, and 10 migrations. All plotted migration edges are significant. Pit and castle species names are colored blue and red, respectively.



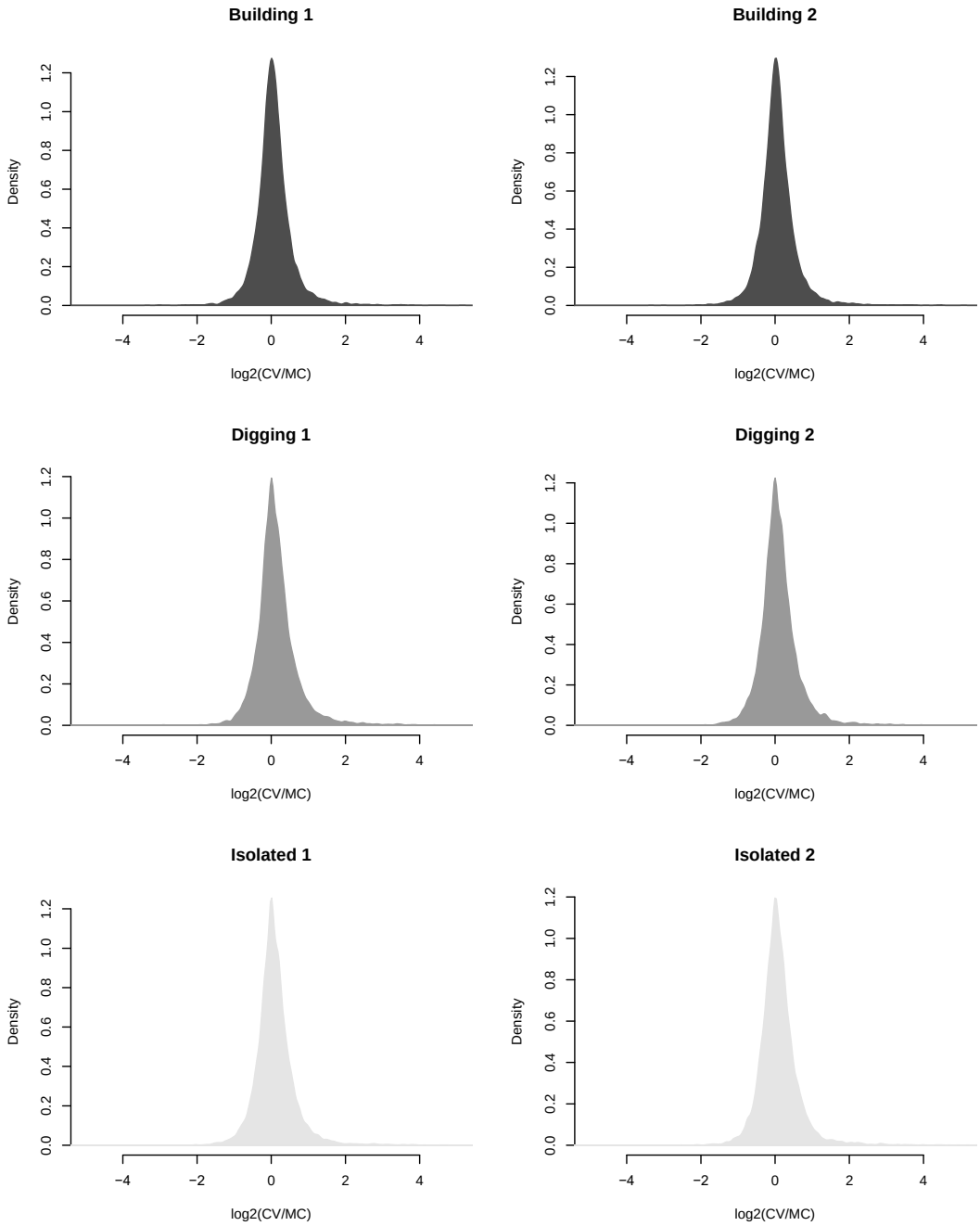
Appendix B Figure 6 :**Genome-wide f_d distribution |** Genome-wide scatterplots of f_d for the five most significant comparisons from the f_4 analyses (cartoon comparisons presented next to the scatterplots). Higher f_d values indicate greater support for introgression at that locus.



Appendix B Figure 7 : **Ontogeny of *Copadichromis virginalis* x *Mchenga conophoros* F_1 hybrid bower building |** The cartoon indicates the typical progression of bower building stages during a courtship "season", proceeding from the initiation of pit-digging to the transition to castle-building.



Appendix B Figure 8 : **Genes with discordant and concordant allele-specific expression (ASE) across behavioral states | (A)** Cartoon examples of the log2 ASE ratio between *C. virginalis* ("CV"; pit) and *M. conophoros* ("MC"; castle) alleles for genes showing concordant ASE (same direction in allelic bias between behaviors) and discordant ASE (different direction in allelic bias). **(B)** Scatterplot of the number of genes with concordant and discordant ASE at different log2 ASE ratio thresholds. **(C)** Barplot of gene-level RNA-seq expression counts for CV and MC alleles across digging and building behaviors for the gene *atp1b4*. **(D)** Barplot of gene-level RNA-seq expression counts for CV and MC alleles across digging and building behaviors for the gene *dgcr8*.



Appendix B Figure 9 : **The distribution of allele specific expression across F_1 hybrid samples and conditions |** Density plots representing the distribution of log2(CV allele counts/MC allele counts) for all genes measured after thresholding (see methods) in each sequencing sample.

REFERENCES

1. Nosil, P. and D. Schluter, *The genes underlying the process of speciation.* Trends Ecol Evol, 2011. **26**(4): p. 160-7.

2. Seehausen, O., et al., *Genomics and the origin of species.* Nature Reviews Genetics, 2014. **15**(3): p. 176-192.

3. Ellegren, H., *Genome sequencing and population genomics in non-model organisms.* Trends in Ecology & Evolution, 2014. **29**(1): p. 51-63.

4. Salzburger, W., *Understanding explosive diversification through cichlid fish genomics.* Nat Rev Genet, 2018.

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

6. Seehausen, O., *African cichlid fish: a model system in adaptive radiation research.* Proc Biol Sci, 2006. **273**(1597): p. 1987-98.

7. Wellborn, G.A. and R.B. Langerhans, *Ecological opportunity and the adaptive diversification of lineages.* Ecol Evol, 2015. **5**(1): p. 176-95.

8. Lamichhaney, S., et al., *Evolution of Darwin's finches and their beaks revealed by genome sequencing.* Nature, 2015. **518**(7539): p. 371-375.

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

10. Jones, F.C., et al., *The genomic basis of adaptive evolution in threespine sticklebacks.* Nature, 2012. **484**(7392): p. 55-61.

11. Poelstra, J.W., et al., *The genomic landscape underlying phenotypic integrity in the face of gene flow in crows.* Science, 2014. **344**(6190): p. 1410-4.

12. Malinsky, M., et al., *Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake.* Science, 2015. **350**(6267): p. 1493-1498.

13. Brawand, D., et al., *The genomic substrate for adaptive radiation in African cichlid fish.* Nature, 2014. **513**(7518): p. 375-81.

14. Turner, G.F., et al., *How many species of cichlid fishes are there in African lakes?* Molecular Ecology, 2001. **10**(3): p. 793-806.

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

16. Albertson, R.C., J.T. Streelman, and T.D. Kocher, *Genetic basis of adaptive shape differences in the cichlid head.* Journal of Heredity, 2003. **94**(4): p. 291-301.

17. Loh, Y.H.E., et al., *Origins of Shared Genetic Variation in African Cichlids.* Molecular Biology and Evolution, 2013. **30**(4): p. 906-917.

18. Albertson, R.C., et al., *Evolutionary mutant models for human disease.* Trends Genet, 2009. **25**(2): p. 74-81.

19. Albertson, R.C., et al., *Genetic basis of continuous variation in the levels and modular inheritance of pigmentation in cichlid fishes.* Molecular Ecology, 2014. **23**(21): p. 5135-5150.

20. Fraser, G.J., R.F. Bloomquist, and J.T. Streelman, *A periodic pattern generator for dental diversity.* Bmc Biology, 2008. **6**.

21. Hofmann, C.M., et al., *The Eyes Have It: Regulatory and Structural Changes Both Underlie Cichlid Visual Pigment Diversity.* Plos Biology, 2009. **7**(12).

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

23. Sylvester, J.B., et al., *Brain diversity evolves via differences in patterning.* Proceedings of the National Academy of Sciences, 2010: p. 201000395.

24. Sylvester, J.B., et al., *Competing signals drive telencephalon diversity.* Nature Communications, 2013. **4**.

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

26. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012: p. 1222794.

27. Degner, J.F., et al., *DNase I sensitivity QTLs are a major determinant of human expression variation.* Nature, 2012. **482**(7385): p. 390.

28. Gaffney, D.J., et al., *Dissecting the regulatory architecture of gene expression QTLs.* Genome biology, 2012. **13**(1): p. R7.

29. Okhovat, M., et al., *Sexual fidelity trade-offs promote regulatory variation in the prairie vole brain.* Science, 2015. **350**(6266): p. 1371-1374.

30. Bendesky, A., et al., *The genetic basis of parental care evolution in monogamous mice.* Nature, 2017. **544**(7651): p. 434.

31. Pfenning, A.R., et al., *Convergent transcriptional specializations in the brains of humans and song-learning birds.* Science, 2014. **346**(6215): p. 1256846.

32. Greenwood, A.K., et al., *Evolution of schooling behavior in threespine sticklebacks is shaped by the Eda gene.* Genetics, 2016: p. genetics. 116.188342.

33. Baran, N.M., P.T. McGrath, and J.T. Streelman, *Applying gene regulatory network logic to the evolution of social behavior.* Proceedings of the National Academy of Sciences, 2017. **114**(23): p. 5886-5893.

34. Guryev, V., et al., *Genetic variation in the zebrafish.* Genome research, 2006. **16**(4): p. 491-497.

35. Fischer, A., et al., *Evidence for a complex demographic history of chimpanzees.* Molecular biology and evolution, 2004. **21**(5): p. 799-808.

36. Consortium, I.H., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851.

37. King, M.-C. and A.C. Wilson, *Evolution at two levels in humans and chimpanzees.* Science, 1975. **188**(4184): p. 107-116.

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

39. Fraser, G.J., et al., *An ancient gene network is co-opted for teeth on old and new jaws.* PLoS biology, 2009. **7**(2): p. e1000031.

40. Powder, K.E. and R.C. Albertson, *Cichlid fishes as a model to understand normal and clinical craniofacial variation.* Developmental biology, 2016. **415**(2): p. 338-346.

41. Loh, Y.H.E., et al., *Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids.* Genome Biology, 2008. **9**(7).

42. Conte, M.A., et al., *Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes.* bioRxiv, 2018.

43. Ben-Ari Fuchs, S., et al., *GeneAnalytics: An Integrative Gene Set Analysis Tool for Next Generation Sequencing, RNAseq and Microarray Data.* Omics-a Journal of Integrative Biology, 2016. **20**(3): p. 139-151.

44. Piñero, J., et al., *DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants.* Nucleic Acids Research, 2017. **45**(D1): p. D833-D839.

45. Attanasio, C., et al., *Fine Tuning of Craniofacial Morphology by Distant-Acting Enhancers.* Science, 2013. **342**(6157).

46. Rada-Iglesias, A., et al., *Epigenomic Annotation of Enhancers Predicts Transcriptional Regulators of Human Neural Crest.* Cell Stem Cell, 2012. **11**(5): p. 633-648.

47. York, R.A., et al., *Behavior-dependent cis regulation reveals genes and pathways associated with bower building in cichlid fishes.* Proceedings of the National Academy of Sciences, 2018(in press).

48. Steventon, B., et al., *Differential requirements of BMP and Wnt signalling during gastrulation and neurulation define two steps in neural crest induction.* Development, 2009. **136**(5): p. 771-779.

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

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

51. White, D.E., et al., *Quantitative multivariate analysis of dynamic multicellular morphogenic trajectories.* Integrative Biology, 2015. **7**(7): p. 825-833.

52. Houart, C., et al., *Establishment of the telencephalon during gastrulation by local antagonism of Wnt signaling.* Neuron, 2002. **35**(2): p. 255-265.

53. Streit, A., *The cranial sensory nervous system: specification of sensory progenitors and placodes.* 2008.

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

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

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

57. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**: p. 11.10.1-33.

58. Lee, T.H., et al., *SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data.* Bmc Genomics, 2014. **15**.

59. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-2158.

60. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3.* Fly, 2012. **6**(2): p. 80-92.

61. Kent, W.J., et al., *The human genome browser at UCSC.* Genome research, 2002. **12**(6): p. 996-1006.

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

63. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

64. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.

65. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

66. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

67. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

68. Hochberg, Y. and Y. Benjamini, *More powerful procedures for multiple significance testing.* Stat Med, 1990. **9**(7): p. 811-8.

69. Gunter, H.M., et al., *Identification and characterization of gene expression involved in the coloration of cichlid fish using microarray and qRT-PCR approaches.* Journal of molecular evolution, 2011. **72**(2): p. 127-137.

70. Harris, R.S., *Improved Pairwise Alignmnet of Genomic DNA.* 2007.

71. Blanchette, M., et al., *Aligning multiple genomic sequences with the threaded blockset aligner.* Genome research, 2004. **14**(4): p. 708-715.

72. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.

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

74. Keenleyside, M.H.A., *Cichlid Fishes: Behavior, Ecology and Evolution*. 1st ed. 1991: Chapman and Hall.

75. York, R., et al., *Evolution of bower building in Lake Malawi cichlid fish: Phylogeny, morphology, and behavior.* Frontiers in Ecology and Evolution, 2015. **3**.

76. Magalhaes, I.S., G.E. Croft, and D.A. Joyce, *Altering an extended phenotype reduces intraspecific male aggression and can maintain diversity in cichlid fish.* Peerj, 2013. **1**: p. 17.

77. Conte, M.A. and T.D. Kocher, *An improved genome reference for the African cichlid, Metriaclima zebra.* Bmc Genomics, 2015. **16**.

78. Meier, J.I., et al., *Ancient hybridization fuels rapid cichlid fish adaptive radiations.* Nature Communications, 2017. **8**.

79. Kirkpatrick, M. and N. Barton, *Chromosome inversions, local adaptation and speciation.* Genetics, 2006. **173**(1): p. 419-434.

80. Chen, K., et al., *BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.* Nat Meth, 2009. **6**(9): p. 677-681.

81. Freese, N.H., D.C. Norris, and A.E. Loraine, *Integrated genome browser: visual analytics platform for genomics.* Bioinformatics, 2016. **32**(14): p. 2089-2095.

82. Bouckaert, R.R., *DensiTree: making sense of sets of phylogenetic trees.* Bioinformatics, 2010. **26**(10): p. 1372-1373.

83. Martin, S.H. and S.M. Van Belleghem, *Exploring Evolutionary Relationships Across the Genome Using Topology Weighting.* Genetics, 2017. **206**(1): p. 429-438.

84. Martin, S.H., J.W. Davey, and C.D. Jiggins, *Evaluating the Use of ABBA-BABA Statistics to Locate Introgressed Loci.* Molecular Biology and Evolution, 2015. **32**(1): p. 244-257.

85. Pickrell, J.K. and J.K. Pritchard, *Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data.* Plos Genetics, 2012. **8**(11).

86. Mims, M.C., et al., *Geography disentangles introgression from ancestral polymorphism in Lake Malawi cichlids.* Molecular Ecology, 2010. **19**(5): p. 940-951.

87. Robinson, G.E., R.D. Fernald, and D.F. Clayton, *Genes and Social Behavior.* Science, 2008. **322**(5903): p. 896-900.

88. Wittkopp, P.J. and G. Kalay, *Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence.* Nature Reviews Genetics, 2012. **13**(1): p. 59-69.

89. Fraser, H.B., et al., *Systematic Detection of Polygenic cis-Regulatory Evolution.* Plos Genetics, 2011. **7**(3).

90. Pestov, N.B., et al., *Evolution of Na,K-ATPase beta m-subunit into a coregulator of transcription in placental mammals.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(27): p. 11215-11220.

91. Hsu, R., et al., *Loss of microRNAs in pyramidal neurons leads to specific changes in inhibitory synaptic transmission in the prefrontal cortex.* Molecular and cellular neurosciences, 2012. **50**(3-4): p. 283-292.

92. Chandrasekaran, S., et al., *Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(44): p. 18020-18025.

93. Hrvatin, S., et al., *Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex.* Nature Neuroscience, 2018. **21**(1): p. 120-+.

94. Orr, H.A., *Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data.* Genetics, 1998. **149**(4): p. 2099-2104.

95. Irisarri, I., et al., *Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes.* Nat Commun, 2018. **9**(1): p. 3159.

96. Küpper, C., et al., *A supergene determines highly divergent male reproductive morphs in the ruff.* Nature Genetics, 2016. **48**(1): p. 79.

97. Lamichhaney, S., et al., *Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax).* Nature Genetics, 2016. **48**(1): p. 84.

98. Schwander, T., R. Libbrecht, and L. Keller, *Supergenes and complex phenotypes.* Current Biology, 2014. **24**(7): p. R288-R294.

99. Geschwind, D.H. and J. Flint, *Genetics and genomics of psychiatric disease.* Science, 2015. **349**(6255): p. 1489-1494.

100. Moyerbrailean, G.A., et al., *High-throughput allele-specific expression across 250 environmental conditions.* Genome research, 2016(26): p. 1627-1638.

101. Auwera, G.A., et al., *From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline.* Current protocols in bioinformatics, 2013: p. 11.10. 1-11.10. 33.

102. McKenna, A., et al., *The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Research, 2010. **20**(9): p. 1297-1303.

103. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nature Genetics, 2011. **43**(5): p. 491-+.

104. McGrath, P.T., et al., *Parallel evolution of domesticated Caenorhabditis species targets pheromone receptor genes.* Nature, 2011. **477**(7364): p. 321-U92.

105. Strimmer, K., *fdrtool: a versatile R package for estimating local and tail area-based false discovery rates.* Bioinformatics, 2008. **24**(12): p. 1461-1462.

106. Zhou, X. and M. Stephens, *Genome-wide efficient mixed-model analysis for association studies.* Nature Genetics, 2012. **44**(7): p. 821-U136.

107. Supek, F., et al., *REVIGO summarizes and visualizes long lists of gene ontology terms.* PloS one, 2011. **6**(7): p. e21800.

108. Ramakrishnan Varadarajan, A., et al., *Genome-wide protein phylogenies for four African cichlid species.* BMC Evolutionary Biology, 2018. **18**.

109. Amores, A., et al., *A RAD-Tag Genetic Map for the Platyfish (Xiphophorus maculatus) Reveals Mechanisms of Karyotype Evolution Among Teleost Fish.* Genetics, 2014. **197**(2): p. 625-U307.

110. O'Quin, C.T., et al., *Mapping of pigmentation QTL on an anchored genome assembly of the cichlid fish, Metriaclima zebra.* Bmc Genomics, 2013. **14**: p. 8.

111. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* arXiv preprint arXiv:1303.3997, 2013.

112. Kielbasa, S.M., et al., *Adaptive seeds tame genomic sequence comparison.* Genome Research, 2011. **21**(3): p. 487-493.

113. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner*. Bioinformatics, 2013. **29**(1): p. 15-21.

114. R Core Team, *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2016.

115. Mayba, O., et al., *MBASED: allele-specific expression detection in cancer tissues and cell lines.* Genome Biology, 2014. **15**(8): p. 405.

116. Thomas, P.D., et al., *PANTHER: a library of protein families and subfamilies indexed by function.* Genome research, 2003. **13**(9): p. 2129-2141.

117. Konings, A., *Malawi Cichlids in their Natural Habitat . 2007*. El Paso, TX, US: Cichlid Press Google Scholar.

118. Kidd, M.R., C.E. Kidd, and T.D. Kocher, *Axes of differentiation in the bower-building cichlids of Lake Malawi*. Molecular Ecology, 2006. **15**(2): p. 459-478.