Genome-enabled discovery of evolutionary divergence in brains and behavior

Chinar Patil1,2, Jonathan B. Sylvester1,2, Kawther Abdilleh1, Michael W. Norsworthy3,4, Karen Pottin5, Milan Malinsky6,7, Ryan F. Bloomquist3,8, Zachary V. Johnson1, Patrick T. McGrath1 & Jeffrey T. Streelman1

Lake Malawi cichlid fishes exhibit extensive divergence in form and function built from a relatively small number of genetic changes. We compared the genomes of rock- and sand-dwelling species and asked which genetic variants differed among the groups. We found that 96% of differentiated variants reside in non-coding sequence but these non-coding diverged variants are evolutionarily conserved. Genome regions near differentiated variants are enriched for craniofacial, neural and behavioral categories. Following leads from genome sequence, we used rock- vs. sand-species and their hybrids to (i) delineate the push–pull roles of BMP signaling and irx1b in the specification of forebrain territories during gastrulation and (ii) reveal striking context-dependent brain gene expression during adult social behavior. Our results demonstrate how divergent genome sequences can predict differences in key evolutionary traits. We highlight the promise of evolutionary reverse genetics—the inference of phenotypic divergence from unbiased genome sequencing and then empirical validation in natural populations.

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

Over the past two decades, systems have been developed to identify the genetic basis of traits from nature. Amongst vertebrate animals, these traits include body armor, color, head and jaw shape, parental care, song and coordinated movement. The take home message from this work has been that a small number of genes from recognizable pathways explain a considerable proportion of phenotypic variance. Yet, these studies may be biased in interpretation and limited in inference space. The focus is typically on one or two species and one trait at a time, often using hybrid pedigrees founded by a small number of individuals, and candidate gene or QTL approaches. Here, we explored a different strategy in a system of many species with many divergent traits. We sought to determine the genetic differences between closely related groups of species and then to focus experiments on leads from genome divergence. In essence, we’ve asked the genome which traits to follow. This
approach is similar to genome-first examples in mammals, wherein unbiased analysis of genome data selects the traits rather than a trait-first model like QTL mapping.

The Malawi cichlid system is an appropriate one for our research aims. The assemblage comprises hundreds of closely related species that have diverged in the last 500,000 to one million years, such that the genomes of individuals across species boundaries remain highly similar. An appreciable fraction of genetic polymorphism identified in Malawi species is shared with cichlid lineages from throughout East Africa—suggesting that ancient genetic variation fuels diversification of the Malawi flock. Set against this background of genome similarity, Malawi cichlids exhibit staggering diversity in phenotypes including pigmentation, sex determination, craniofacial and brain patterning, and social behavior. Previous work has focused on the genomic and early developmental underpinnings of this diversity, in rock- vs. sand-dwelling species.

Rock- vs. sand-species form ecologically distinct groups similar to other ecotypes in well-known evolutionary radiations (marine vs. freshwater sticklebacks; tree vs. ground finches and anoles). 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. Malawi rock-dwellers tend to be strongly territorial and aggressive; they breed and feed at high density in complex rock-reef habitats. Most eat algae from the substratum with strongly reinforced jaws packed with teeth. Adult rock-dweller brains exhibit enlarged anterior components, telencephalons and olfactory bulbs. Sand-dwellers are less site-specific and less aggressive. They often breed on communal leks where males build sand ‘bowers’ to attract females. 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 (Supplementary Fig. S1). We aimed to understand evolutionary divergence between rock- and sand-dwelling lineages by identifying the number, type and spectrum of genetic variants that separate these groups.

To target the primary axis of evolutionary divergence, we compared whole genomes of one male individual each from 8 rock-dwelling and from 14 sand-dwelling species (Supplementary Table S1), to an average of 25X coverage per individual. Species were chosen to represent the diversity present within each of the rock- and sand-groups (Fig. 1A), in terms of body size, color pattern, ecology and phylogenetically defined lineages within the sand- species group.

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

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

4484 genes lie within 25 kb of either an alternately fixed variant or a highly divergent 10 kb window (2.5% FDR). Pathway enrichment analysis of human homologs/analogs for these genes reveals categories spanning early embryonic development, craniofacial morphogenesis, brain development, synaptic transmission and neuronal function (Supplementary Table S2). In particular, rock-sand divergent genes are enriched for GO Biological Process terms ‘telencephalon development’ (p < 1.7e−18), ‘adult behavior’ (p < 2e−14), ‘synaptic plasticity’ (p < 1.4e−12), ‘odontognathogenesis’ (p < 3.7e−11), ‘response to BMP’ (p < 3.2e−09), ‘gastrulation’ (p < 5.6e−06), ‘face morphogenesis’ (p < 8.9e−08), ‘neural crest cell differentiation’ (p < 4.3e−13), and ‘eye development’ (p < 1.3e−15). Over-represented gene families included nuclear hormone receptors (p < 3.0e−08), HOXL subclass homeoboxes (p < 1.4e−07), TALE class homeboxes (p < 4.2e−04) and Forkhead boxes (p < 9.33e−04; for novel expression domains in cichlid foxp2 see Supplementary Fig. S2). We observed enrichment for the mouse phenotypes ‘abnormal cognition’ (p < 3.3e−15), ‘abnormal learning and memory’ (p < 2.6e−15), ‘abnormal craniofacial morphology’ (p < 4.9e−11) and ‘abnormal social/conspecific interaction’ (p < 7.1e−14). We used the list of differentiated genes to query an Allen Brain Atlas dataset that reports gene expression in hundreds of brain regions. Rock- sand-divergent genes were enriched for the basomedial nucleus of the amygdala, a sub-region of the telencephalon (pFDR = 0.001) that regulates fear, anxiety, and physiological response to territorial intruders, and has been linked to Social Anxiety Disorder in humans. Finally, we queried databases of genes involved in human disease. Genes near divergent variants are significantly enriched for factors implicated in neurological disease like Autism Spectrum Disorder (SFARI, Fisher’s exact test p value < 2e−16) and disorders related to the neural crest (Fisher’s exact test p value < 2e−16).

Given the prevalence of evolutionarily conserved, non-coding, divergent rock-sand variants and genome-wide enrichment for craniofacial and neural crest biology, we examined overlap with published datasets of.
Figure 1. The genomic substrate for rock vs. sand evolution. (A) A maximum likelihood phylogeny of eight rock- and 14 sand-dwelling species, based on variable sites (informative SNP and InDels) identified throughout the genome. "Sand" species contain representatives of the lineages "shallow benthic, deep benthic and utaka" from Ref.32, while the "rock" species correspond to the "mbuna" lineage. (B) A plot of Z-FST (FST normalized using Fisher's Z-transformation) across the genome (ggplot2 v3.3.3 https://ggplot2.tidyverse.org/), plotting genomic divergence between rock- vs. sand-dwelling groups. Single nucleotide polymorphisms (SNPs) summed over 10 kb bins and insertion-deletion mutations (InDels) are shown on the same scale. Numbers along the x-axis refer to linkage groups (i.e., chromosomes) and threshold lines indicate 2.5% FDR. (C) Evolutionary conservation (PhastCons) scores were calculated for each nucleotide across the genome, subdivided by genome annotation and plotted by bins of increasing FST. The PhastCons score for each genome category is significantly higher for increasing bins of FST (pairwise Wilcoxon rank sum p value < 2e−16 for each pair of bins).
mammalian neural crest and craniofacial enhancers. These data allow us to identify craniofacial and cranial neural crest cell (CNCC) enhancers conserved between mammals and cichlids and fixed variants between rock and sand species within these conserved regulatory elements. A total of 275 craniofacial enhancer elements and 234 human CNCC enhancers are evolutionarily conserved between mammals and cichlids. We found divergent rock-sand mutations within the enhancer elements of key genes integral to CNCC specification and migration (Supplementary Table S2). Notably, from both datasets, fixed rock-sand variants were found within the enhancer region of the gene irx3 (a nuclear receptor and master neural crest regulator). Rock-sand divergent variants were similarly located within craniofacial enhancers of three genes (yap1, fat4, cere) that function in the Hippo signaling pathway, as well as within enhancers of irx3 and axin2. These data linking rock-sand fixed SNP/InDels to evolutionarily conserved, experimentally verified enhancers further underscore the importance of non-coding variation in the craniofacial evolution of rock- and sand-lineages.

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

A gastrula-stage map of forebrain diversification. Rock- vs. sand-dwelling Malawi cichlids exhibit divergence in or near genes enriched for BMP signalling, gastrulation, eye and telencephalon development, as well as the TALE (Irx) gene family. To explore the developmental consequences of this differentiation, we investigated early forebrain specification in rock- and sand-embryos, building upon our previous studies and interest in irx1b and early brain development. During development, the complexity of the vertebrate brain is first laid out in the neural plate, a single-cell thick sheet of cells that forms between non-neural ectoderm and the germ ring at gastrulation. Irx genes act as transcriptional repressors of BMP signal in gastrulation, and function to specify the neural plate. BMPs, in turn, are protective of the anterior-most region of the neural plate, which will ultimately give rise to the telencephalon, and suppress the eye field. Given alternatively fixed variants in the irx1b gene, expected interactions between Irx and BMP signaling in the early embryo and known telencephalon vs. eye size differences between rock- vs. sand-species, we examined and quantified the early activity of irx1b and BMP in rock- vs. sand-embryos.

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

We developed a panel of rock- vs sand-hybrid crosses to formally evaluate the role of irx1b in forebrain diversification. First, we used quantitative RT-PCR to measure allele-specific expression (ASE) in heterozygous rock- vs sand-irx1b hybrids, across the stages of gastrulation. We observed that the sand-irx1b allele was expressed at significantly higher levels (average of 2.5-fold; p = 4.5e−10) and that this difference was largely confined to MG (Fig. 2B). Next, we used hybrid embryos to chart the development of the telencephalon and the eye field. Rock- vs sand-irx1b hybrids, indexed for irx1b genotype, were raised to neurula and somitogenesis stages and we examined the expression of shh (which induces foxg1 and the ventral forebrain), foxg1 (a marker of the telencephalon), and rx3 (a marker of the eye field), by in situ hybridization. F1 individuals homozygous for rock-irx1b alleles exhibited a larger and more rostral domain of shh expression, an earlier and larger domain of foxg1 and a smaller rx3 domain (Fig. 2C,D). These differences between rock- vs sand-irx1b genotypes match expression divergence observed amongst rock- vs sand-species. Finally, when we compared the relative size of the telencephalon among rock-irx1b genotypes, individuals homozygous for rock-alleles exhibited larger telencephala (p < 0.0001, Tukey’s test, Supplementary Fig. S3). We conclude that genetic variants in and around the irx1b gene contribute to divergent specification of the Malawi cichlid forebrain, likely via spatial, temporal and quantitative variation in the expression of irx1b itself.

Our genome sequencing revealed a near-fixed InDel in the 3’ UTR of the Malawi cichlid irx1b gene (Supplementary Fig. S4). Rock-species possess an 85 bp insertion, compared to cichlid species from outside of the Malawi lineage. Sand-dwellers largely lack the insertion and exhibit a 6 bp deletion compared to outgroups. The insertion shows strong genetic similarity to a fragment from the Rex1 family of non-LTR retrotransposons. Given the likelihood that Astatotilapia calliprora populations surrounding Lake Malawi may have seeded the Malawi evolutionary radiation and contributed to rock- and sand-dwelling lineages, we explored the presence/absence of this InDel in Astatotilapia samples. We found that most Astatotilapia individuals and populations had the rock-irx1b allele (the insertion), but that an individual from Chizumulu Island was fixed for the sand-allele and two individuals sampled from Itupi were heterozygous. Because the 85 bp insertion in rock-species is a partial Rex1 fragment, and sand-species carry a 6 bp deletion compared to outgroups, we speculate that the current rock- and sand-divergent alleles were generated by at least two imperfect excision events of an element that invaded the genome of the Malawi + Astatotilapia ancestor. Rex1/Babar retrotransposons have been active
Figure 2. A gastrula-stage map of forebrain diversification. (A) Double in situ hybridization—immunohistochemistry to visualize irx1b expression (red) and BMP [PSMAD] activity (green) across the stages of cichlid gastrulation. DAPI in blue. Three rows represent early, mid and late gastrulation (EG, MG, LG) in embryos of rock- (R) and sand-dwelling (S) species. Sand-dwellers show expanded irx1b expression in the dorsal portion of the embryo at EG, expanded irx1b expression in the anterior domain at MG (arrowheads) and clear PSMAD activity from the developing neural plate (np) earlier in LG (arrowheads). e epidermis. Schematics at left show irx1b expression domains in red, on cartoons of cichlid embryos. (B) Relative expression of rock (red) and sand (blue) irx1b alleles, sampled from 74 heterozygous rock-X sand-F2 embryos, across the stages of gastrulation. F2 embryos were sampled at stage 9 (gastrulation). Because Malawi cichlid species are maternal mouthbrooders and eggs are fertilized in batches per brood, each brood's embryos vary in timing of fertilization by up to 4 h. Embryos within broods can therefore be sub-staged in gastrulation (see “Materials and methods” section). Each bar on the plot represents the relative allelic expression of sand- and rock-irx1b in a heterozygous F2 individual. Quantification of allele-specific expression (ASE) shows that levels are sand-biased, and that this effect is strongest in MG. (C) In situ hybridization of shh and foxg1, during neurula stages, showing development of the telencephalon in rock-X sand-F2 embryos, indexed for irx1b genotype. F2 individuals homozygous for rock-irx1b alleles (R/R) show a more dorsal progression of shh expression (black arrowheads), an earlier and a larger expression domain of the telencephalon marker foxg1. The top two rows are lateral views; bottom row is a dorsal view. Dotted lines demarcate the outline of the embryo in dorsal view. Heterozygous individuals exhibit greater variation in expression domains (middle columns), indicating that genetic factors other than variants in irx1b contribute to this phenotype. (D) In situ hybridization for rx3 and pax6, during neurula stages, chart the development of the eye field in rock-X sand-F2 embryos indexed for irx1b genotype. F2 individuals homozygous for sand-irx1b alleles (S/S) show larger domains of rx3 (black arrowheads) and larger eyes (pax6, also marked by yellow dotted line), but smaller telencephala (white dotted line). All panels are dorsal views. Heterozygous individuals exhibit greater variation in expression domains (middle columns), indicating that genetic factors other than variants in irx1b contribute to this phenotype.
Genomics of divergent social challenge and opportunity. Rock- and sand-dwelling Malawi cichlids live in strikingly different social and physical environments. Rock-dwelling males tend to be more aggressive than sand-dwellers and defend territories year-round as sites for feeding and breeding. By contrast, sand-dwellers are more exploratory than rock-dwellers and only breeding males tend to be territorial, often building sand bowers to attract females and mitigate male-male aggression. Given these observations and genome-wide enrichment for categories related to adult behavior and social interaction, we designed an experimental paradigm to investigate brain gene expression profiles associated with divergent rock- vs. sand-social behaviors.

We evaluated social challenge and opportunity amongst males using a large tank with a ‘rock’ habitat at one end and ‘sand’ at the other, separated by glass bottom (Fig. 3A). When parental rock-species are placed in this tank paradigm, males court females on the rock side of the tank. Males of sand-species court females over sand and construct species-appropriate bowers. When single hybrid rock-x sand-F1 males are placed in this arena with hybrid F1 females, males invariably court females over the ‘rock’ habitat. However, when two rock-x sand-hybrid F1 males (brothers) were allowed to compete for gravid hybrid F1 females in this tank paradigm, we observed something different. One male, typically the larger, courted females over the rock habitat, and the other simultaneously constructed bowers to court females over the sand. We found no difference in gonadal-somatic index (GSI), an established biological metric of reproductive status and maturity, between F1 males behaving as ‘socially rock’ vs. ‘socially sand’ (Supplementary Fig. S5). Our observation of divergent behavior between F1 brothers in the same tank suggests an interaction between the genome and the social environment.

We used RNA-seq to investigate gene expression profiles associated with behavior of rock-x sand-F1 hybrid males that were actively courting females over rock vs. sand. Whole brains of F1 males tested singly (n = 24) as well as F1 brothers assayed in dyads (n = 4 dyads) were collected during courtship, and interrogated by RNA-seq. Strikingly, gene expression profiles clustered not by fraternal relatedness, but rather by behavioral context (Fig. 3B). Males from dyads that courted females over rocks had expression profiles similar to single males (who also courted over rocks) but distinct from their brothers that built bowers and courted females over sand in the same tank. Genes were considered significantly differentially expressed between ‘social rock’ and ‘social sand’ brains if they exhibited both a fold change ≥ 2 and crossed the threshold of \( p_{\text{adj}} < 0.05 \). Based on this criterion, we found 832 genes differentially expressed between rock- vs. sand-behaving males (Fig. 3B, Supplementary Table S3). Among differentially expressed genes, we observed significant functional enrichment for GO Biological Process categories ‘synaptic signaling’ (\( p < 2.3 \times 10^{-21} \)), ‘synaptic plasticity’ (\( p < 3.6 \times 10^{-9} \)), ‘visual behavior’ (\( p < 2.09 \times 10^{-6} \)), mouse phenotypes ‘abnormal learning/memory/conditioning’ (\( p < 5.9 \times 10^{-10} \)), ‘abnormal telencephalon morphology’ (\( p < 3.95 \times 10^{-7} \)), ‘abnormal spatial learning’ (\( p < 9.9 \times 10^{-7} \)), pathways ‘axon guidance’ (\( p < 3.3 \times 10^{-5} \)), ‘oxytocin signaling’ (\( p < 5.2 \times 10^{-5} \)) and ‘estrogen signaling’ (\( p < 1.8 \times 10^{-4} \)) (Supplementary Table S3). Matches against the Allen Brain Atlas database of gene expression yielded enrichment for sub-regions of the telencephalon: CA3, hippocampus (\( p_{\text{adj}} = 4.6 \times 10^{-6} \)), CA2, hippocampus (\( p_{\text{adj}} = 3.7 \times 10^{-5} \)), CA4, hippocampus (\( p_{\text{adj}} = 0.001 \)), claustrum (\( p_{\text{adj}} = 0.001 \)), subiculum (\( p_{\text{adj}} = 0.003 \)), dentate gyrus (\( p_{\text{adj}} = 0.03 \)) and the basomedial nucleus of the amygdala (\( p_{\text{adj}} = 0.03 \)). The hippocampus encodes episodic memory and spatial representations of the environment, and more recently its subregions have been shown to play critical roles in anxiety, social interaction, and social memory formation. Roughly 38% of differentially expressed genes also contained genetically differentiated SNPs/InDels between rock- and sand- species (p-value < 2e-6, Fisher’s exact test), implying considerable cis-acting genetic variation. Enrichment of numerous categories related to brain function and synaptic plasticity showed greater overlap than expected (Fig. 3C). These context-dependent differences suggest concerted changes in brain gene expression as males experienced and responded to different social challenges and opportunities.

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

Recently we surveyed genome-wide divergence between sand-dweller sub-groups that construct pit vs. castle bowers, sand-made structures to attract females for mating. Mapping those variants to the same genome reference, we expected distinct patterns of diversification because rock-sand and pit-castle divergence likely occurred at different times, along different trait axes, under the control of different evolutionary forces. Consistent with
expectation, there is clear clustering of genome divergence on chromosome 11 for the pit-castle comparison (Supplementary Fig. S6), while all chromosomes carry the signature of rock-sand diversification (Fig. 1B). However, contrary to our expectation, rock- vs. sand- and pit- vs. castle-radiations have diverged in similar gene sets. Out of 3070 genes identified near 10 kb high FST regions in the rock- vs. sand-comparison, 483 overlap with 1090 genes identified near high FST regions in the pit- vs. castle-comparison (p-value < 2e−9, Fisher’s exact test, Supplementary Table S4). This result may imply that evolutionary diversification in Lake Malawi is limited, or constrained, by chromosomal location.

Overall, genes in proximity to rock-sand divergent variants were enriched for functional categories related to early forebrain and craniofacial development, neuronal function and social behavior. This list of variants, coupled with consistent patterns of functional and pathway enrichment, motivated follow up experiments focused on early brain development and adult social behavior (Fig. 4). It is apparent from our work here and previously25,26.
that Malawi cichlid brains and nervous systems begin to differ during gastrulation in pathways that can be predicted from divergent genome sequences. This is interesting for at least two reasons. First, this observation runs counter to the 'late equals large' textbook example of how brains evolve differences in relative proportions of their parts, and along with other reports suggests that variation in early brain development may be a significant factor in brain evolution. Overall, such early variation in development is not thought to be a driving force in evolution, precisely because early changes can have global and ramifying effects. Collectively, our findings provide a partial description of the conditions wherein variation during the earliest stages of development can contribute to evolutionary diversification. In each case we have examined, variation in gene expression is quantitative, heterochronic and limited to a precise stage or time period.

Sydney Brenner recognized the relationship between the genetic specification of nervous systems and the behavioral output of the brain. However, because these events take place so far apart in the lifespan of a vertebrate, they are rarely studied simultaneously. Here, the genome connects the two phenomena: rock- vs. sand-divergent gene sets and strong functional enrichment indicate that both brain development and social behavior have contributed to the evolutionary diversification of these groups. To evaluate social behavior in rock- vs. sand-dwelling Malawi cichlids, we constructed a social context arena. The presence of sand and simulated rocky caves was sufficient to elicit species-appropriate male behavior when rock- or sand-males were tested with rock- or sand-gravid females. When rock-x sand-F1 hybrid males were tested, one per tank, with F1 hybrid females, males courted females in the rock quadrant of the tank. Notably, when dyads of F1 brothers were tested in this tank paradigm with gravid F1 females, we observed simultaneous 'social rock' and 'social sand' behavior. Brain gene expression profiles from behaving males clustered by social context (social rock vs social sand), and not by fraternal relationships. Differentially expressed genes were enriched for brain regions and pathways implicated in social interaction and overlapped significantly with rock- vs. sand-divergent genetic variants.

Social context is known to influence the brain. For instance, our clustering results are similar to those of Whitfield and colleagues (Whitfield et al. 2003) who showed that brain gene expression in honey bees was predictive of behavior. Likewise, changes in brain morphology and gene expression predictably accompany the ascent to dominance in the cichlid fish Astatotilapia burtoni. Our data seem not to fit the model of dominant-subordinate however. In our experiments, the gonado-somatic index (GSI) did not differ between social rock vs. social sand males within dyads. Both males exhibited nuptial coloration, courted females and in cases with multiple gravid females, both brothers reproduced. Body size was associated with divergent social rock vs. social sand behavior of F1 males; the social rock brother was always larger (mean mass was 26.96 g ± 3.4 [SE] compared to 19.45 g ± 2.2).

Our experiments demonstrate that F1 hybrid male brains can express both social rock- and social sand-behavioral programs, and that social context determines which program is executed. This observation is similar to, but also different than, pit-digging × castle-building F1 male Malawi cichlids who carry out parental bower behaviors in a specific sequence. Notably, in both cases, the hybrid males exhibit one of the two different parental behaviors at any one time—there is no intermediate behavior. In the pit- vs. castle-case, we think that the bower structure itself and/or a threshold signal from females might lock the hybrid male brain into a behavioral state. In the rock- vs sand-case here, it appears that other social cues (i.e., the presence and size of a rival male) lock the hybrid male brain into a behavioral state. These context-dependent behaviors, accompanied by changes in brain gene expression, are compelling examples of interaction between the genome and the social environment. The cellular and genetic basis of these behaviors and their plasticity deserves further attention. Here, our
Materials and methods

Genome sequencing. We extracted genomic DNA, from fin clips of 22 male individuals (Qiagen DNeasy, Cat #69504), from rock dwelling and 14 sand dwelling Lake Malawi species (Supplementary Table S2) representing broad diversity across the rock and sand lineages in Lake Malawi. We made libraries using the Illumina Nextera Library prep kit and performed paired-end sequencing on the Illumina Hi-Seq 2500 at Georgia Tech. The Metriaclima zebra reference genome version MZ_UMD2a was used for genome alignment, variant discovery and annotation using standard BWA and GATK practices. The maximum likelihood tree in Fig. 1A was constructed using SNPhylo, from variant data.

Genetically divergent regions. Vcftools was used to calculate F_{ST} between the 8 rock and 14 sand species excluding SNPs with less than half variants called across species (max-missing 0.5). Variants with F_{ST} = 1 were noted to be alternately fixed between rock and sand lineages in our dataset. F_{ST} was also measured across 10 kb windows (fist-window-size). Significance thresholds were marked using the fdrtool package in R. All variants were annotated using Snpeff. We tested the genes within 25 kb of significantly differentiated variants for enrichment of functional categories. The cichlid gene names were converted to human analogs using Treefam based mapping and functional enrichment was determined using the TOPPFUN web-browser interface.

PhastCons analysis. Pairwise alignments were generated using lastz 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 = 15,000”. This was followed by using UCSC genome utilities (https://genome.ucsc.edu/utl.html) axtChain tool with -minScore = 5000. 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 alignments were generated from pairwise alignments with the multiz v1.2 program, using default parameters and the following pre-determined phylogenetic tree: (((M. zebra, P nyererei), A. burtoni), N. brichardi), O. niloticus) in agreement with Brawand et al. Sequence conservation scores were then obtained using PhastCons with a phylogenetic model estimated by the phyloFit program, both from the PHAST software package (v.1.3). The model fitting was done using default parameters. 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 ran PhastCons again to attain the conservation scores with —target-coverage 0.3—expected-length 100.

Vertebrate-conserved enhancer elements. A comparative genomic approach was used to identify putative craniofacial and neural crest CNEs in mammals that segregate SNPs between rock-sand cichlid species. 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 from published literature. We used the liftOver tool, which maps orthologous genomic regions between species to convert genomic coordinates from one species to another. Using a Human to Oreochoemis niloticus to Metriaclima zebra mapping and a Mouse to Oreochoemis niloticus to Metriaclima zebra mapping, we identified the orthologous genomic locations of the published craniofacial and neural crest enhancers in cichlids. We designated any alternately fixed variant (variant with F_{ST} = 1) that was also within an orthologous CNE as putatively involved in the rock-sand divergence (Supplementary Table S2).

Brain region enrichment analysis. We identified 10,391 cichlid genes with human homologues and generated an expression matrix for each gene across 250 human brain structures spanning telencephalon, diencephalon, mesencephalon, and metencephalon using adult human brain microarray data collected by the Allen Brain Institute. Cortical regions and gyri for which fish do not have putative homologues were excluded from the analysis. The expression matrix was generated using the get_expression function in the ABAEnrichment Bioconductor package in R. We then calculated the specificity of expression for each gene in each of these brain regions using the specificity_index function in the pSI package for R. This function calculates a matrix of gene expression specificity indices, and corresponding p-values, as described previously. We then tested whether (1) genes within 25 kb of rock vs. sand significantly differentiated variants (described above under “Genetically Divergent Regions”), and (2) genes that were differentially expressed between rock- vs. sand-behaving F1 hybrid males, were enriched for transcriptional markers of specific brain regions using the fisher.iteration function with Benjamini–Hochberg correction, again using the pSI package for R. For enrichment testing of differentially expressed genes, we restricted analysis to genes that met the following criteria: (1) transcripts for the gene were detected in all eight paired behaving males, and (2) at least 6 transcripts were detected in each subject.

Staging during gastrulation. Cichlid gastrulation was split into three sub stages within the gastrula stage. Gastrulation lasts 8 to 12 h, depending on the species, and is defined as after the shield (as described in
Embryos from each species, as well as the F2 individuals, were measured starting from the earliest the telen-sonar hunter), **Copadichromis borleyi** and **Metriaclima zebra**. At that point, embryos were 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.

**Immunohistochemical staining.** Embryos were harvested at 24 h post fertilization (hpf) from each of the rock-dwelling cichlids **Metriaclima patricki**, **Metriaclima zebra** and **Labeotropheus fuelleborni** as well as the sand-dwellers **Aulonocara jacobsfreibergi**, **Copadichromis virginalis**, **Copadichromis borleyi** and **Tramitichromis intermedius**. 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 (1.55 mL 5 M NaCl, 250 ul Tris–HCl, pH 7.5, and 95 mg NaBH4) overnight, and 10% 2-mercaptoethanol for wavelengths (500–650 nm). After in situ hybridization, embryos were immunostained for pSMAD 1,5,8 protein, using published protocols. 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 (Zeiss) and Image J. For experiments in Fig. 2A, multiple embryos of multiple rock- and sand-species were examined (n = 20 embryos total); and in Fig. 2C,D, multiple embryos per gene, genotype and stage were used (n = 110 embryos total).

**Rock–sand hybridization and genotyping.** Two rock-sand crosses, one between **Copadichromis borleyi** (CB, sand-dweller sire) and **Metriaclima zebra** (MZ, rock-dweller dam) and another between **Mchenga conophoros** (MC, sand-sire) and **Petrotilapia** sp. ‘chick 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 F1 were grown in tanks and allowed to spawn normally to generate F2. Several F2 broods were taken from multiple F1 females for each cross, a total of 355 individuals for the CB × MZ cross and 608 for the MC × PT cross. The embryos were fixed at every stage starting at gastrula (stage 9) until early pharyngula (stage 14). The F2 embryos were RNA-extracted at stage 9. DNA extraction was performed by fixing the embryos (stage 11–14) in 70% ethanol, then removing the tail from each individual and extracting the DNA using an extraction kit (Qiagen). Following extraction, the F2 embryos were genotyped using custom probes (CAAAATCTCC|CT/CCGCCGG, Taqman custom probes, Invitrogen) designed to identify a SNP in **irx1b** using RT-PCR. A subset of the embryos was also sequenced at a 900 bp interval around the **irx1b** SNP to verify the custom probes.

**Quantitative F2 analysis.** We quantified **irx1b** in F2 at stage 9 and separated by genotypic class. The 74 heterozygous rock X sand F2 embryos were dissected to remove most of the yolk and the total RNA was extracted from each individual using an RNA Extraction Kit (Qiagen).

The amount of mRNA specific to each allele of **irx1b** was quantified by using the RNA-to-Ct kit (Invitrogen) and the custom probes. The delta Ct for each heterozygote was generated with the equation, 2(allele from dam − allele from sire). We used a Student’s t-test to compare the expression of rock- vs sand-alleles in heterozygous F2.

**Forebrain and eye measurements.** The telencephalon and forebrain were measured by integrating the area of transverse sections in embryos of rock- and sand-dweller cichlid species, using previously published methods. The rock-dweller species included **Cynotilapia afra** (CA, planktivore), **Labeotropheus fuelleborni** (LF, algivore) and **Metriaclima zebra** (MZ, generalist); sand-dweller species included **Aulonocara jacobsfreibergi** (AJ, ‘sonar’ hunter), **Copadichromis borleyi** (CB, planktivore) and **Mchenga conophoros** (MC, insectivore/generalist). Embryos from each species, as well as the F1 individuals, were measured starting from the earliest the telencephalon can be differentiated from the forebrain (mid-somitogenesis, stage 12) and at each subsequent stage until the forebrain has defined prosomeres (early pharyngula, stage 14). To keep measurements standardized across stages, all measurements were defined by forebrain morphology at the earliest timepoint (stage 12). The ‘eye’ measurement remains consistent at all stages, the ‘anterior’ measurement includes the telencephalon and presumptive olfactory bulb, and the ‘posterior’ measurement includes the diencephalon and each of its constitutive prosomeres (dorsal and ventral thalamus and hypothalamus). To facilitate measurements, we used gene expression of **rx3** (for stage 12 embryos) and **pax6** (stage 13 and 14) to identify the different structures of the forebrain and eye.

**RNA extraction and sequencing, adult social behavior.** Two adult F1 hybrid males from rock-x sand-interspecific crosses (Supplementary Table S5) were introduced to an assay tank containing 3–5 hybrid females of the same cross with simulated rock habitat on one side and simulated sand habitat on the other side separated by empty tank space (Fig. 3A). Males were observed over a period of four days for courtship behavior, either similar to a rock parent (courting females over the rocky parts of the tank), here designated “social rock” – or to a sand parent (bower building in the sand side of the tank while courting females), here designated “social sand”. When courtship behavior of both males was observed, we primed experiments the next morning by resetting the
Expression Omnibus database under the accession number GSE122500.

**Differential gene expression analysis.** Raw sequence reads from whole brain transcriptomes were quality controlled using the NGS QC Toolkit\(^8\). Raw reads with an average PHRED quality score below 20 were filtered out. Filtered reads were also trimmed of low-quality bases at the 3’ end. High quality sequence reads were aligned to the *M. zebra* reference genome MZ_UMD2a\(^4\) using TopHat v2.0.9\(^8\). 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\(^8\). Sorted sam files were used as input for the HTSeq-count v0.6.1 program to obtain fragment counts for each locus\(^8\). Fragment counts were scale-normalized across all samples using the calcNormFactors function in the edgeR package v3.6.8\(^7\). Relative consistency among replicates and samples was determined via the Multidimensional scaling (MDS) feature within the edgeR package in R. The native R function hclust(dist) used to cluster samples. Scale-normalized fragment counts were converted into log2 counts per million reads mapped (cpm) with precision weights using voom and fit to a linear model using limma v3.20.9\(^8\). Pairwise contrasts were constructed between socially rock and socially sand samples. After correcting for multiple comparisons using the Benjamini–Hochberg method\(^8\), genes were considered differentially expressed between socially rock and socially sand samples if they exhibited both a fold change ≥ 2 and Padj < 0.05. Using Treefam based mapping\(^6\) the cichlid gene names were converted to human analogs and functional enrichment was determined using the TOPPFUN web-browser\(^6\).

**Ethics statement.** All of the animals were handled according to methods approved by the Georgia Tech Institutional Animal Care and Use Committee (IACUC) protocols (permit A100029) with every effort made to minimize suffering and is in compliance with the ARRIVE guidelines.

Received: 24 February 2021; Accepted: 8 June 2021
Published online: 21 June 2021

**References**

1. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: From polygenic to omniomic. *Cell* 169, 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038 (2017).
2. Fields, C., Adams, M. D., White, O. & Venter, J. C. How many genes in the human genome? *Nat. Genet.* 7, 345–346. https://doi.org/10.1038/ng0794-345 (1994).
3. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 9718–9723 (2012).
4. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390 (2012).
5. King, M.-C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* 188, 107–116 (1975).
6. Stern, D. L. & Orgogozo, V. The loci of evolution: How predictable is genetic evolution? *Evolution* 62, 2155–2177. https://doi.org/10.1111/j.1558-5646.2008.00410.x (2008).
7. Streelman, J., Peichel, C. L. & Parichy, D. Developmental genetics of adaptation in fishes: The case for novelty. *Annu. Rev. Ecol. Evol. Syst.* 38, 655–681 (2007).
8. Colosimo, P. F. *et al.* The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biol.* 2, E109. https://doi.org/10.1371/journal.pbio.0020109 (2004).
9. Kratochwil, C. F. *et al.* Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations. *Science* 362, 457–460. https://doi.org/10.1126/science.aao6809 (2018).
10. Albertson, R. C., Streelman, J. T., Kocher, T. D. & Yelick, P. C. Integration and evolution of the cichlid mandible: The molecular basis of alternate feeding strategies. *Proc. Natl. Acad. Sci. U.S.A.* 102, 16287–16292. https://doi.org/10.1073/pnas.0506649102 (2005).
11. Shapiro, M. D. *et al.* Genomic diversity and evolution of the head crest in the rock pigeon. *Science* 339, 1063–1067. https://doi.org/10.1126/science.1230423 (2013).
12. Lamichhaney, S. *et al.* Evolution of Darwin’s finches and their beaks revealed by genome sequencing. *Nature* 518, 371–375 (2015).
13. Okhovat, M., Berrio, A., Wallace, G., Ophir, A. G. & Phelps, S. M. Sexual fidelity trade-offs promote regulatory variation in the prairie vole brain. *Science* 350, 1371–1374 (2015).
14. Bendesky, A. *et al.* The genetic basis of parental care evolution in monogamous mice. *Nature* 544, 434 (2017).
15. Pfennig, A. R. *et al.* Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* 346, 1256848 (2014).
16. Greenwood, A. K., Wark, A. R., Yoshida, K. & Peichel, C. L. Genetic and neural modularity underlie the evolution of schooling behavior in threespine sticklebacks. *Carr. Biol.* 23, 1884–1888. https://doi.org/10.1007/s12228-013-0705-8 (2013).
17. Marcovitz, A., Jia, R. & Bejerano, G. “Reverse genomics” predicts function of human conserved noncoding elements. *Mol. Biol. Evol.* 33, 1358–1369. https://doi.org/10.1093/molbev/msw001 (2016).
64. Van der Auwera, G. A. et al. From fastq data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* 43, 11–33. https://doi.org/10.1002/0471250953.tb11043 (2013).

65. Lee, T. H., Guo, H., Wang, X. Y., Kim, C. & Paterson, A. H. SNPPhylo: A pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* 15, 162 (2014).

66. Danecek, P. et al. The variant call format and VCF tools. *Bioinformatics* 27, 2156–2158 (2011).

67. Strimmer, K. fdrtool: Versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* 24, 1461–1462. https://doi.org/10.1093/bioinformatics/btn209 (2008).

68. 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* 6, 80–92 (2012).

69. Ramakrishnan Varadarajan, A., Mopuri, R., Streelman, J. T. & McGrath, P. T. Genome-wide protein phylogenies for four African cichlid species. *BMC Evol. Biol.* 18, 1. https://doi.org/10.1186/s12862-017-1072-2 (2018).

70. Chen, J., Bardez, E. E., Aronow, B. J. & Jegga, A. G. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311. https://doi.org/10.1093/nar/gkp427 (2009).

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

72. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14, 708–715 (2004).

73. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050. https://doi.org/10.1101/gr.3715005 (2005).

74. Siepel, A. & Haussler, D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* 21, 468–488 (2004).

75. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* 12, 996–1006 (2002).

76. Grote, S., Prufer, K., Kelso, J. & Dannemann, M. ABAEnrichment: An R package to test for gene set expression enrichment in the adult and developing human brain. *Bioinformatics* 32, 3201–3203. https://doi.org/10.1093/bioinformatics/btw392 (2016).

77. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Heintz, N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res.* 38, 4218–4230. https://doi.org/10.1093/nar/gkq130 (2010).

78. Xu, X., Wells, A. B., O’Brien, D. K., Nehorai, A. & Dougherty, J. D. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *J. Neurosci.* 34, 1420–1431. https://doi.org/10.1523/JNEUROSCI.4488-13.2014 (2014).

79. Murata, Y. et al. Allocentric growth of the trunk leads to the rostral shift of the pelvic fin in teleost fishes. *Dev. Biol.* 347, 236–245. https://doi.org/10.1016/j.ydbio.2010.07.034 (2010).

80. Fraser, G. J., Bloomquist, R. F. & Streelman, J. T. A periodic pattern generator for dental diversity. *BMC Biol.* https://doi.org/10.1186/1741-7007-6-32 (2008).

81. Tucker, J. A., Mintzer, K. A. & Mullins, M. C. The BMP signaling gradient patterns dorsoventral tissues in a temporally progressive manner along the anteroposterior axis. *Dev. Cell* 14, 108–119. https://doi.org/10.1016/j.devcel.2007.11.004 (2008).

82. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* 9, 671–675. https://doi.org/10.1038/nmeth.2089 (2012).

83. Patel, R. K. & Jain, M. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE* 7, 7 (2012).

84. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1110 (2009).

85. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).

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

87. Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat. Med.* 9, 811–818 (1990).

Acknowledgements

This work was supported by Grants from the NIH (R01GM101095, 2R01DE019637-10 to JTS and F32GM128346-01A1 to ZVJ) and the Human Frontiers Science Program (RGPF0052/2019 to JTS). We would like to thank Shweta Bilya and the Genomics Core at Georgia Tech for invaluable assistance with the NGS sequencing. We also thank the members of the Streelman lab for comments on this manuscript.

Author contributions

C.P., J.B.S., K.A. and J.T.S. designed research; C.P., J.B.S., K.A., M.W.N., K.P. and R.F.B. performed research; C.P., K.A., M.M., Z.V.J. and P.T.M. analyzed data; C.P., J.B.S., Z.V.J. and J.T.S. wrote the paper. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-92385-8.

Correspondence and requests for materials should be addressed to C.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
