Allele-specific expression reveals gene-environment interactions driving ecotype divergence in Malawi cichlid crosses

Joseph Stockert
Georgia Institute of Technology, School of Biological Sciences

2 May 2019

Abstract

During the early stages of adaptive radiation, species diverge through local adaptation to spatially distinct microhabitats (Streelman and Danley 2003). The evolutionary history of spatial niche differentiation is poorly understood in the Lake Malawi cichlid flock because both the rock-dwelling and sand-dwelling ecotypes appear scattered across many clades, meaning that either multiple divergence events or hybridization may have occurred (Hulsey et al. 2017). In order to shed light on the evolutionary mechanisms of the rock/sand divergence, we sequenced neural transcriptomes from experimental F1 hybrids that preferentially restricted courtship behaviors to rocky or sandy territories based on social setting in lab aquariums. We then analyzed these RNA data for signals of allele-specific expression in order to identify interactions between genes and environment (GxE) and cis-regulatory elements that are involved with local adaptation in cichlids.

Introduction

Evolutionary genetics seeks to make comparisons between genetic sequence data and observations of molecular, morphological, behavioral, and ecological features to elucidate the genetic basis of specific biological processes and the evolution history of biological phenomena. (Hofmann et al. 2014). Only recently, computational tools have emerged to integrate the dynamics of DNA-to-RNA transcription into evolutionary genetics research(Gu and Wang 2015). While comparisons between genomic DNA sequences reveal changes in the full set of a lineage's genetic instructions over long time scales, transcriptomic RNA sequences provide a better picture of how the many different parts of these instructions are carried out on a situational basis. Through novel combinations of high-quality reference genomes published for many diverse species and RNAseq data from laboratory animals, researchers gain vastly improved resolution for exploring the genetics of hard to study biological features such as behavior. Here, we integrate the transcriptomes of experimental hybrids with the genomes of their parents to clarify the evolutionary mechanisms of ecotype divergence in the Lake Malawi cichlid radiation.

Field studies have previously characterized the impressive diversity of the Malawi cichlid flock, in which hundreds of closely-related species have emerged over a relatively short time in response to environmental flux (Brawand et al. 2014, Ivory et al. 2016). Given the cichlids' wide phenotypic variation it is unsurprising that many of the mechanisms driving the processes of adaptive radiation in this lineage remain obscure. However, a growing body of research focuses has narrowed in on elucidating the evolution of a few cichlid traits. For example, a recent studie has compared the ecological and morphological differences between two variations on a unique mating behavior in the bower-building family of cichlids (York et al. 2015).

While it is certainly worthwhile to define interesting traits that are specific to a few species, it is equally important to examine foundational traits that lay a groundwork for the evolution of the lineage. One framework for adaptive radiation (Streelman and Danley 2003) organizes these traits into stages of increasing niche specificity. The late stage of radiation is defined by the emergence of complex communication systems – bower-building, which serves to communicate reproductive fitness and evolves by sexual selection, is a strong example of this late-stage radiation. According to this framework, the early stage of adaptive radiation is defined by microhabitat differentiation. In the Malawi cichlids, divergence between two primary ecotypes, rock-dwellers and sand-dwellers, has driven subsequent behavioral diversity (Streelman and Danley 2003). Since bower-building occurs in sand-dwelling species, and sand-dwelling itself is derived from the putatively ancestral rock-dwelling state, it follows that investigating the origins of sand-dwelling can complete our picture of adaptive radiation in cichlids by filling in early evolutionary history where social behavior cannot.

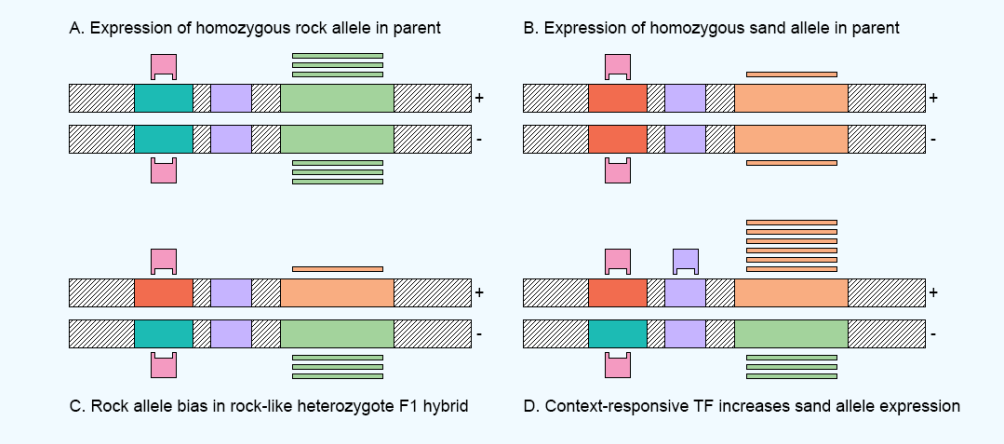

Figure 1. Cartoon depicting differences in relative expression of homologous alleles from the rock ecotype or sand ecotype (A and B) that create an allelic imbalance in hybrid progeny (C and D). This imbalance drives expression of a behavioral phenotype that is more like one parent than the other. Environmentally sensitive cis-regulatory elements can alter the direction of allelic imbalance depending on context (D).

Our understanding of the history of spatial divergence in Malawi cichlids is convoluted by tenuous phylogenetic relationships (Hulsey et al. 2017). These ecotype phylogenies could be vastly improved by identification of specific genetic elements involved in local adaptation. Transcriptomic analysis offers one avenue to the identification of such genes; in an experimental study on F1 hybrid plants, allele-specific expression and differential gene expression were shown to play a role in local adaptation, with microhabitat having an effect on allele expression plasticity (Gould et al. 2018). We aim to use patterns of allele-specific expression (Fig. 1) in F1 hybrids with rock-dwelling and sand-dwelling parents as evidence for possible genetic mechanisms driving the rock-to-sand divergence in Lake Malawi cichlids. In constructed half-rock, half-sand environments, hybrids from the same brood display opposite ecotype behaviors. We hypothesize that environmentally sensitive gene-regulatory control could allow for these two distinct phenotypes to develop over a shared genetic background. To support this hypothesis, we examine the neural RNA transcripts from hybrids reared in two different social conditions for evidence of context-dependent ASE. We identify genes that show regulatory interaction with social context and environment and the molecular pathways responsible for microhabitat

differentiation in Malawi cichlids. This research seeks to expand our knowledge of the processes of adaptive radiation by filling a gap in the mechanisms acting on the early evolutionary stages of this well-known radiation.

Methods

Experimental animals: Ten F1 hybrids were previously reared from crosses of six cichlid species, three rock dwellers (*MC, TI,* and *AB*) and three sand dwellers (*LF, MZ,* and *PC*). Hybrid males were housed alongside several female fish in tanks with substrate divided evenly between speckled sand and artificial rocks. Two tanks housed a single hybrid male (*LFxMC* and *MZxMC*), while four tanks housed pairs of males from the same brood (*LFxMC, MZxTI, PCxAB*, and *PCxMC*). In social tanks, one male displayed a subset of courtship behaviors (display, lead, quiver; York et al. 2015) exclusively over rock while the other displayed the same behaviors exclusively over sand. In the isolated tanks, males displayed these behaviors exclusively over rock. Males interacted with the substrate on their half of the tank, i.e. hiding under rocks or manipulating sand. All males free swam, approached, and chased females in both parts of the tank.

Sample preparation: Males were sacrificed within 90 minutes of a courtship behavior display and decapitated. Brains were stabilized in RNA-Later, frozen in liquid nitrogen, and homogenized. After solubilization in TRIzol and chloroform extraction, RNA was purified via Quiagen RNEasy mini columns and stored at -80°C.

RNAseq: RNA was quantified using Qubit probes, assessed for quality via Agilent 2100 Bioanalyzer, and normalized to 1 µg. Libraries were prepared with the Illumina TruSeq Stranded mRNA Sample Prep Kit and sequenced on the Illumina HiSeq 2500 system.

RNA data preparation: RNA reads were aligned to the *M. zebra* UMD2a reference (Conte et al. 2015) using the Burrow-Wheelers Aligner; SNPs were called using the GATK variant calling pipeline and filtered to include only heterozygous sites. To decrease the effects of mapping bias preferring reference alleles over variant alleles, these sites were then used to mask the *M. zebra* reference, and reads were re-aligned to the masked reference with BWA as before. Reads for either allele were then counted at each variant site using the CountSNPASE script from the Fraser lab (available on GitHub, https://github.com/thefraserlab/aser). Meanwhile, samtools mpileup was used to call variant sites in the parental species genomes, which were then filtered to include only homozygous sites. F1 RNA reads were phased using homozygous parental variants so that each F1 allele could be assigned to either a rock or sand origin. Finally, exonic alleles were labelled with gene annotation IDs.

Detection of context-specific ASE signals: Significance of allele-specific expression signals were calculated at both SNP- and gene-level using the MBASED package in R 3.5.1 (package version 1.16.0, 1-sample analysis, known haplotype phasing, 1000000 simulations; Mayba and Gilbert 2018). Results were formatted using custom GNU AWK scripts (available upon request) for export to Excel. Patterns of allele bias (Fig. 2) were then compared between rock-like and sand-like groups. First, genes were identified for which allele bias was universal in direction across all samples (concordant ASE). Next, genes were identified that showed

differential allele bias between ecotypes (discordant ASE). Finally, we identified genes for which the level of bias differed between groups but was not necessarily opposite in direction (diffASE): genes included in this group had directionality that either matched or contrasted with ecotype (eg. rock allele up-expressed in rock-type vs. sand allele up-expressed in rock-type).

Gene enrichment testing: Functional enrichment for genes that showed significant context-specific patterns of ASE was compiled using the ToppFun web tool using default settings (Chen et al. 2009).

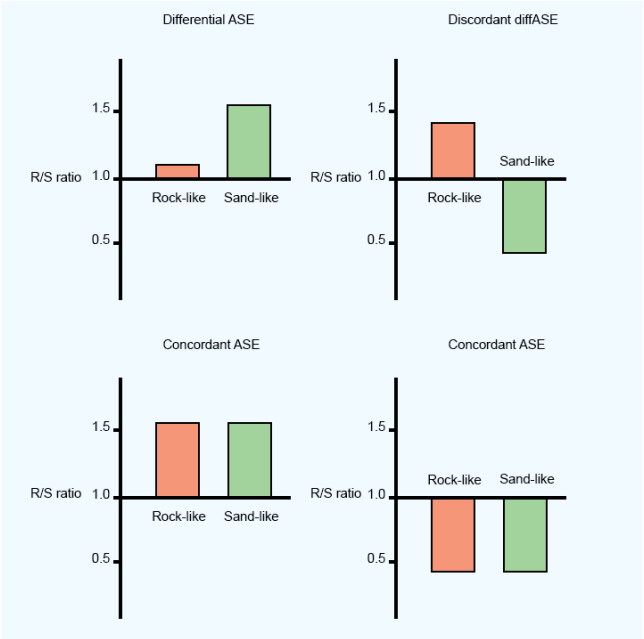

Figure 2. Cartoon depicting patterns of allele-specific expression comparisons between two groups. Differential ASE refers to allelic imbalances that differ in magnitude or direction between groups. Discordant ASE is a specific pattern of diffASE and refers to allelic imbalances that are of similar magnitude and direction in both groups.

Results

We analyzed the neural transcriptomes of social rock-like (n=4), social sand-like (n=4), and isolated (n=2) hybrid cichlids. After alignment and variant calling, we identified 21,473 sites at which phasing could be determined and a minimum of 4 individuals had read data.

SNP-level ASE: We first calculated allelic ratios and determined significance of ASE signals at each SNP. We found 2045 SNPs that showed a significant allele bias ($p<0.05$) for at least one individual from each social ecotype (rock-like or sand-like). For context comparisons, we considered only sites where at least two individuals from each condition had data. Allele ratios at 194 sites showed a pattern of differential ASE; 21 of these sites showed a pattern of discordant ASE, but only 8 of these sites skewed in the expected direction (e.g. rock allele up-expressed in the rock condition and down-expressed in the sand condition). 108 sites showed signals of concordant ASE; a similar number (57 vs 51 sites) skewed universally toward either the rock or sand allele.

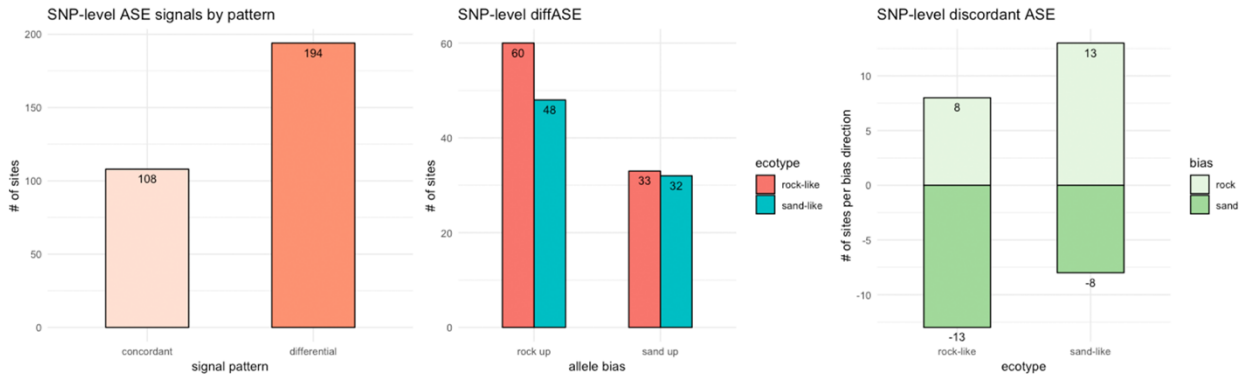

Figure 3. (a) More sites showed a context-dependent pattern of allelic imbalance than a universal parental bias, evidence of environmentally sensitive regulatory differences between hybrid ecotypes. (b) Among cases of differential ASE, the allelic imbalances preferred the paternal allele in both hybrid ecotypes. (c) Bias directions at discordant ASE sites generally mismatched the expressed ecotype.

Gene-level ASE

Magnitude of read counts varied drastically from site to site. In an attempt to reduce variance and noise in the data, we added gene annotations and summed total allele counts across whole genes. After re-running MBASED on gene-level counts, we identified 48 genes that showed significant GxE interactions (diffASE). There was no difference in the number of discordant ASE genes that matched or mis-matched ecotype (Figure 4).

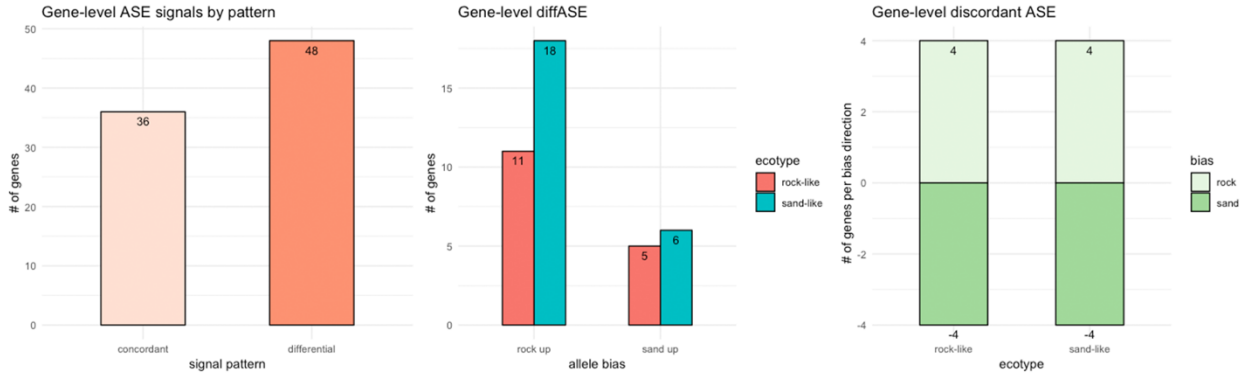

Figure 4. (a) After calculating ASE at the whole-gene level, the proportion of differential to concordant signals decreased. (b) While gene-level allelic imbalances again favored the parental allele, more significant ASE signals appeared in sand-like hybrids than rock-like, opposite of the SNP-level analysis.

Gene enrichment:

The set of discordant ASE signals was not particularly enriched for many notable gene classes associated with behavior; many of these genes, such as *pik3r4* (sand allele upexpressed 81% in PCxAB_SUB, p=0.011, rock allele up-expressed 25% in PCxAB_DOM, p=0.044) are expressed in many fish tissues. However, a significant discordant ASE gene identified in the SNP-level analysis, *gabbr1* (rock allele up-expressed 48% in LFxMC_DOM, p=0.017; sand allele up-expressed 44% in PCxMC_SUB, p=0.040) is expressed primarily in the cerebellum and eye of adult zebrafish (Cocco et al. 2017).

Discussion

General findings: Overall, results of our analysis show definitive presence of genes interacting with the social environment to regulate divergent behavioral regimes. In experimental animals expressing two distinct microhabitat phenotypes atop nearly identical genotypes, we identified widespread cases of allelic imbalances, including a number of genes where the allelic imbalance favored one parental allele in one context and favored the other in a different context. Such cases of discordant ASE represent a transcriptomic divergence that mirrors the overlaying ecotype divergence. Differences in allele-specific expression are often attributed to the action of *cis*-regulatory elements. In *cis*-regulation, variant sequences bind with different affinities to environmentally-sensitive transcription factors; thus, one allele may be expressed at a different level than the other (Witkopp and Kalay 2012). This type of regulation may increase the ability of an animal to deal with environmental stress through phenotypic trade-offs (Gould et al. 2018). For example, in a fish faced with competition for mating territory, regulatory responses could induce a preference for a potentially less suitable microhabitat type where there is no competition. Our finding of significant context-dependent ASE points to *cis*-regulation as being a putative molecular driver of divergence in Malawi cichlids. Further, our identification of a gene that is expressed in the brain of adult fish (*gabbr1*) provides promising evidence that allele-specific expression analysis can uncover sources of variation that specifically act on the pathways controlling complex behaviors such as local adaptation. Genes that show clear patterns of context-dependent ASE may be good candidates for further ontology and transcription factor enrichment studies to clarify the exact molecular pathways forming such behaviors.

Caveats of this study: Our analysis was limited by a few key considerations that may have affected the number of genes identified as showing significant ASE. The general upshot was that data missing for some individuals reduced the power of context comparisons for a large number of variant sites. In several cases, sites that showed interesting ASE signals in a few individuals were considered insignificant overall because read data was missing for the other individuals. We used five different crosses in our behavioral experiment, yet only had one of each cross represented for any given context (i.e. only one LFxMC_DOM, LFxMC_SUB, etc.). Without replicates for each cross in each context, it is difficult to determine whether a site was excluded for a given individual because that site was non-variant or simply because of RNAseq quality. This may have also introduced species effects that confounded our comparisons across contexts: in some cases, sites were missing only for a single cross. We were unable to rule out the possibility of allelic imbalances that are tied to species-specific regulation and not to environmental context. Additionally, the number of samples in each comparison added a level of complexity to our significance threshold that merits some reevaluation. To call a gene significant in differential ASE comparisons, we required all individuals of either ecotype to have a significant signal. This resulted in a minimum overall p-value of 0.05^2 for a significant gene. However, various combinations of p-values above 0.05 for individuals could still result in an overall p-value that is less than 0.05. Thus, we believe that our results are extremely conservative and that an additional p-value correction could increase then number of significant ASE genes.

Informing future work: This work serves as a starting point for the application of allele-specific expression analysis to understanding the divergence of complex behaviors in adaptive radiations. The successes and failures of our experiment can inform future analyses on this and

other systems. We aligned our RNA data to a masked reference genome and successfully avoided a universally-paternal bias that plagued a previous attempt to analyze this dataset. This protocol is sufficient to negate reference bias even when the reference species is included as an experimental animal. We also successfully phased RNA alleles by incorporating whole genomes from the parent species. As mentioned above, the statistical methodology for calling ASE signals significant needs improvement and will serve as the immediate next step in this research going forward.

References

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* (2009).

Cocco, A. et al. Characterisation of the γ-aminobutyric acid signalling system in the zebrafish (Danio rerio Hamilton) central nervous system by reverse transcription quantitative polymerase chain reaction. *Neuroscience* 343, 300-321 (2017).

Conte, M. A. et al. An improved genome reference for the African cichlid, Metriaclima zebra. *BMC Genomics* 16, 724. (2015).

Brawand, D. et al. The genomic substrate for adaptive radiation in African cichlid fish. *Nature* 513, 375-381 (2014).

Gould, B.A., Chen, Y. and Lowry, D.B. Gene Regulatory Divergence Between Locally Adapted Ecotypes in Their Native Habitats. *Mol Ecol.* Accepted Author Manuscript (2018).

Gu, F. and Wang, X. Analysis of allele specific expression – a survey. *Tsinghua Sci & Tech* 20(5), 513-529 (2015).

Hofmann, H.A. et al. An evolutionary framework for studying mechanisms of social behavior. *Trends in Ecology & Evolution* 29, 581-589 (2014).

Hulsey, C. D., Zheng, J., Faircloth, B. C., Meyer, A., and Alfaro, M. E. Phylogenomic analysis of Lake Malawi cichlid fishes: Further evidence that the three-stage model of diversification does not fit. *Mol. Phylogenet. Evol.* 114, 40-48 (2017).

Mayba, O. and Gilbert, H. MBASED: Package containing functions for ASE analysis using Meta-analysis Based Allele-Specific Expression Detection. (2018).

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

Wittkopp, P. J., and Kalay, G. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genetics* 13, 59-69 (2012).

York, R. A., et al. Evolution of bower building in Lake Malawi cichlid fish: phylogeny, morphology, and behavior. *Frontiers Eco & Evo* 3, 1-13 (2015).

York, R. A., et al. Behavior-dependent *cis* regulation reveals genes and pathways associated with bower building in cichlid fishes. *PNAS* 115(47), E11081-E11090 (2018).