GENOMIC CHANGES UNDERLYING ADAPTIVE TRAITS AND REPRODUCTIVE ISOLATION BETWEEN YOUNG SPECIES OF *CYPRINODON* PUPFISHES

Joseph Alan McGirr

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology in the Graduate School.

Chapel Hill
2020

Approved by:

Christopher H. Martin

Daniel R. Matute

Robert J. Duronio

Christopher S. Willett

R. Brian Langerhans

ABSTRACT

Joseph Alan McGirr: Genomic changes underlying adaptive traits and reproductive isolation between young species of *Cyprinodon* pupfishes
(Under the direction of Christopher H. Martin)

Adaptive radiations showcase dramatic instances of biological diversification resulting from ecological speciation, which occurs when reproductive isolation evolves as a by-product of adaptive divergence between populations. While this process seems widespread and may account for much of life's diversity, there is little known about genomic differences between species that influence differences in phenotypes and contribute to reproductive barriers. In my dissertation work, I used a variety of evolutionary genomic methods to study the genetic basis of rapid ecological speciation within an adaptive radiation of *Cyprinodon* pupfish endemic to San Salvador Island, Bahamas, which consists of a dietary generalist species and two trophic specialists – a molluscivore and a scale-eater.

In my first chapter, I combined genome-wide divergence scans, selections scans, and association mapping to discover loci that were highly diverged between species, showed signs of recent selection, and were associated with variation in jaw size – the primary axis of phenotypic divergence in this system. In my second chapter, I found that the scale-eater and molluscivore species showed similar gene expression patterns compared to the generalist species, providing the first evidence of parallel changes in gene expression underling adaptation to divergent niches. These findings indicated convergent adaptation to higher trophic levels through shared genetic pathways. In my third and fourth chapters, I measured gene expression levels in F1 hybrids

generated from crosses between San Salvador species. Intriguingly, many genes that were differentially expressed between sympatric species were also misregulated in their F1 hybrids. These results indicate that divergent ecological selection in sympatry can drive hybrid gene misregulation which may act as a primary reproductive barrier between nascent species. In my fifth chapter, I combined whole-genome resequencing data with total mRNA sequencing to identify candidate *cis*-acting genetic variation influencing rapidly evolving craniofacial phenotypes. I found very few alleles fixed between species – only 157 SNPs and 87 deletions. By measuring allele-specific expression in F1 hybrids, I found strong evidence for *cis*-regulatory alleles affecting expression divergence of genes with putative effects on skeletal development. These results highlight the utility of the San Salvador pupfish system as an evolutionary model for craniofacial development.

To Kimiko, for making these years my happiest, and to my parents, for their enduring support and dedication to my education

ACKNOWLEDGEMENTS

A dissertation is what you get when you let blind curiosity guide major life decisions. When you let an inclination toward research give way to total devotion. When the answer you're looking for is always one experiment away. I feel extremely grateful to have had the opportunity to do the research culminating in these chapters, and there are many colleagues, friends, and family deserving thanks.

First, I would like to thank my advisor Chris Martin, for inspiring me to think deeply about the process of speciation, providing me with the intellectual freedom to pursue my own ideas (the good ones and the terrible ones), showing me how to navigate the alien world of academia, and for laying a rich foundation of research to build upon. I would also like to thank my committee: Daniel Matute for his unwavering commitment to my success as a graduate student and beyond, and pushing me to reach for the absolute upper limits of my potential. Everyone needs a Daniel Matute in their life. Bob Duronio for being my committee chair, introducing me to the world of chromatin biology, and for the hours spent in his office drawing out decision trees. Chris Willett for making the red pod a close community and providing me with a new lab home near the end of my PhD. Brian Langerhans for his thought provoking questions during my committee meetings and valuable discussions about adaptation.

This work was supported by a Graduate Student Fellowship from the Triangle Center for Evolutionary Medicine, a Rosemary Grant Travel Award, and an L.I. Gilbert Travel Award. I thank the Gerace Research Centre on San Salvador Island for logistics and the Bahamian government BEST Commission for permission to conduct this research. I thank Daniel Matute,

Chris Willett, Jennifer Coughlan, Emilie Richards, Michelle St. John, Bryan Reatini, Kimiko Suzuki, and Aaron Comeault for illuminating comments on chapters that have been previously published.

Friendships and family have been essential to completing this work. I thank my lab mates Emilie Richards and Michelle St. John for conversations that have benefitted my research as well as my mental health, and for making lab feel like a home. I thank Dave Turissini for teaching me how to code in R. I thank Bryan Reatini and Kate Gould for all of the family dinners, game nights, and camping trips that enabled a healthy work-life balance. I thank my parents Allison and Kevin, my step-father John, and sister Melanie for their love that has supported me thorough 23 years of schooling. I would have never made it without their guidance and belief in me. Finally, I want to thank my fiancé Kimiko, the most exciting finding I made during my PhD, far more significant than even the tiniest *p*-values in the pages below. She became my main source of support, helped me to understand myself, and made graduate school the happiest years of my life. While the answers to my research questions might always remain one experiment away, I've found the definitive answer to at least one very important question, and I am grateful for her.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: NOVEL CANDIDATE GENES UNDERLYING EXTREME TROPHIC SPECIALIZATION IN CARIBBEAN PUPFISHES[1]

Introduction

Identifying genetic changes underlying phenotypic diversity is necessary to understand how these changes drive adaptation and speciation (Coyne 2004; Moczek 2008; Rausher and Delph 2015; Byers et al. 2016). Adaptive radiations showcase the world's most dramatic instances of rapid ecological divergence (Turner 1976; Schluter 2000; Seehausen 2006; Losos and Ricklefs 2009; Lamichhaney et al. 2016) making them ideal for investigating the genetic basis of traits influencing novel niche use. Characterizing divergent regions underlying adaptation will address several longstanding questions in evolutionary genomics, such as: how many differentiated regions do we find between closely related species? Is novel trophic specialization driven by selective sweeps? Does the effect size of loci contributing to phenotypic divergence depend on the distance between fitness peaks across an adaptive landscape? (Hermisson and Pennings 2005; Orr 2005; Noor and Feder 2006; Barrett and Schluter 2008; Jensen 2014; Dittmar et al. 2016; Hoban et al. 2016). Genomic divergence scans measuring relative genetic differentiation and genome-wide association mapping are two strategies used to detect candidate gene regions responsible for species differences (Gompert et al. 2012; Visscher

[1] This chapter previously appeared as an article in *Molecular Biology and Evolution.* The original citation is as follows: McGirr, J. A., and C. H. Martin. 2017. Novel candidate genes underlying extreme trophic specialization in Caribbean Pupfishes. Mol. Biol. Evol. 34:873–888.

et al. 2012; Comeault et al. 2014; Pallares et al. 2014; Puzey et al. 2015; Chaves et al. 2016; Irwin et al. 2016). Together these powerful tools can be used to discover genomic regions that are both highly diverged between species and associated with ecologically important traits (Li and Durbin 2011; Xia et al. 2013; Byers et al. 2016).

A number of recent genome-wide F_{st} scans comparing closely related species pairs have located small regions (typically < 200kb) that are highly differentiated relative to the rest of the genome (Carneiro et al. 2014; Poelstra et al. 2014; Soria-Carrasco et al. 2014; Lamichhaney et al. 2015; Malinsky et al. 2015), suggesting these regions are responsible for species-specific phenotypes. Recent literature has emphasized the importance of estimating F_{st} alongside within-population nucleotide diversity (π) and between-population divergence (D_{xy}) in order to more accurately interpret the evolutionary significance of genetically differentiated regions (Nachman and Payseur 2012; Cruickshank and Hahn 2014; Irwin et al. 2016). Importantly, any reduction of within-population diversity will necessarily inflate estimates of F_{st} because it is a relative measure of differentiation (Noor and Bennett 2009; Nachman and Payseur 2012; Cruickshank and Hahn 2014). Therefore, F_{st} interpretations are heavily dependent on the interplay of forces acting to reduce within-population diversity, including selective sweeps, purifying selection, background selection, and low recombination rates (Noor and Bennett 2009; Cruickshank and Hahn 2014). Estimating between-population divergence at loci with high F_{st} and low within-population diversity can help distinguish between these possibilities because nucleotide divergence between species increases at loci under different selective regimes (Nachman and Payseur 2012; Cruickshank and Hahn 2014; Irwin et al. 2016). However, between-population divergence can also be influenced by patterns of hitchhiking and background selection (Cruickshank and Hahn 2014). Selection statistics comparing the distribution of allele

frequencies across segregating sites can also help determine if reduced diversity at a locus is due to selective sweeps, in which selection has increased the frequency of a single (hard sweep) or multiple haplotypes (soft sweep) (Maynared Smith and Haigh 1974; Tajima 1989; Hermisson and Pennings 2005; Pavlidis et al. 2013; Jensen 2014). Statistics that rely on the distribution of allele frequencies within and between populations should be interpreted in the context of their demographic history (Galtier et al. 2000; Andolfatto 2001; Nielsen 2005; Hoban et al. 2016). This can be achieved by inferring changes in ancestral population sizes and using these estimates to model a demography-corrected neutral distribution of allele frequencies (Pavlidis et al. 2013; Schiffels and Durbin 2014). Combining F_{st}, π, D_{xy}, and selective sweep statistics can reveal functionally diverged regions of the genome; however, these statistics alone are insufficient to determine how such regions might affect phenotypic differences between species.

Genome-wide association studies expand on divergence scans by identifying regions that are directly associated with phenotypic differences between species. The simplest approach involves estimating associations between SNPs and quantitative traits by fitting a linear regression of phenotype on allele frequency (Purcell et al. 2007; Visscher et al. 2012), while more advanced methods account for population structure and estimate the effect size of SNPs associated with traits (Price et al. 2006; Kang et al. 2010; Zhou and Stephens 2012; Zhou et al. 2013). Accounting for population structure can help filter out false positive associations, but may also filter out true associations (Marchini et al. 2004; Zhao et al. 2011). Thus, we implemented both types of association models alongside genome divergence scans. We used this mixed strategy to identify candidate SNPs affecting novel ecological traits in an excellent system for examining rapid adaptive diversification.

Three sympatric *Cyprinodon* species inhabit the hypersaline lakes of San Salvador Island, Bahamas, and radiated within the past 10,000 years based on the most recent drying of these lakes (Mylroie, J.E, Hagey 1995; Turner et al. 2008). A generalist species, *C. variegatus*, feeds primarily on algae and detritus, a diet representative of all allopatric Cyprinodontidae (Martin and Wainwright 2011). The first of two specialist species, the 'snail-eater' *C. brontotheroides*, expanded its diet to include more gastropods and ostracods (Martin and Wainwright 2013a). Snail-eater oral jaws are smaller with a larger in-lever to out-lever ratio compared to the generalist, increasing mechanical advantage for biting (Martin and Wainwright 2013a). The snail-eater is also defined by a prominent protruding nasal region that may be used for leverage while crushing hard-shelled prey (Martin and Wainwright 2013a,b). The second sympatric specialist, the 'scale-eater' *C. desquamator,* expanded its diet to include scales removed from other species during quick strikes. Scale-eaters have greatly enlarged jaws with a smaller in-lever to out-lever ratio, large adductor muscles, and an elongated body compared to the generalist and snail-eater species (Martin and Wainwright 2013a). Phylogenetic analyses of outgroup *Cyprinodon* species and surveys of pupfish populations on neighboring Bahamian islands confirm that scale-eating and snail-eating niches are entirely unique to *C. desquamator* and *C. brontotheroides,* respectively, and each species is endemic to hypersaline lakes on San Salvador Island, providing strong support that these specialists diverged from a generalist common ancestor during recent adaptive radiation (Martin and Wainwright 2011; Martin 2016b).

Adaptive landscapes describe the relative fitness of various trait (or allelic) combinations given a particular environment – where adaptive peaks represent optimal combinations and adaptive valleys represent unfit combinations (Wright 1988; Schluter 2000). If the scale-eater and snail-eater specialists rapidly ascended to novel adaptive peaks within the past 10,000 years,

then we should expect to see high rates of morphological diversification in traits associated with trophic specialization. Indeed, San Salvador *Cyprinodon* pupfishes exhibit morphological diversification rates up to 51 times faster than other Cyprinodontidae clades, with jaw size undergoing the most rapid diversification (Martin and Wainwright 2011; Martin 2016b). The San Salvador pupfish system is one of the few examples of a multi-peak adaptive landscape measured for three species (Martin and Wainwright 2013b; Martin 2016a), presenting an excellent opportunity to test mathematical models of adaptation. This landscape was estimated using F_2 hybrids generated from F_1 hybrid intercrosses and backcrosses to all three species. This produced a continuum of phenotypes that were used to estimate relationships between fitness and phenotypic resemblance to parental types. The fitness optima for generalist and snail-eater phenotypes were separated by a small fitness valley, while the phenotypic optimum of the scale-eater presumably exists outside of the range of phenotypic variation tested in the F_2 population (Fig. 1.1) (Martin and Wainwright 2013b). Although this landscape did not measure a scale-eater fitness optimum, it does show that the phenotypic distance is greater between the generalist fitness peak and the fitness valley surrounding hybrid phenotypes most resembling the scale-eaters than between the generalist and snail-eater fitness peaks (Fig. A2.1A). This greater phenotypic distance is primarily due to the large jaws of scale-eaters (Fig. A2.1B). Orr's extension of Fisher's geometric model predicts that *de novo* mutations with a large effect on phenotypic variation are more likely to be fixed during adaptation toward distant phenotypic optima than nearby optima (Orr 1998, 2005). Based on this model, we predict more large-effect variants mediated the transition from generalist to scale-eater due to the greater phenotypic distance across the fitness valley separating these species.

Here we focus on identifying loci associated with variation in jaw morphology within this radiation due to the strikingly rapid divergence of this trait that has clear ecological fitness consequences. We identified 12 million SNPs from 37 genomes sequenced to 7× coverage across nine populations of all three species on San Salvador Island. We discovered novel candidate genes associated with jaw size along with evidence supporting the role of large-effect alleles in crossing between distant phenotypic optima.

Materials and Methods

Study system and sample collection

Individuals were caught from hypersaline lakes on San Salvador Island, Bahamas using a hand net or seine net. 14 scale-eaters were sampled from six populations; ten snail-eaters were sampled from four populations; and 11 generalists were sampled from nine populations on San Salvador and a neighboring island. Samples were collected from nine isolated lakes on San Salvador (Great Lake, Stout's Lake, Oyster Lake, Little Lake, Crescent Pond, Moon Rock, Mermaid's Pond, Osprey Lake, Pigeon Creek, and one closely related outgroup *C. variegatus* population from Lake Cunningham, New Providence Island, Bahamas). Fish were euthanized in an overdose of buffered MS-222 (Finquel, Inc.) following approved protocols from the University of California, Davis Institutional Animal Care and Use Committee (#17455) and University of California, Berkeley Animal Care and Use Committee (AUP-2015-01-7053) and stored in 95-100% ethanol.

Morphometrics

Upper jaw lengths were measured using digital calipers from external landmarks on ethanol-preserved tissue specimens from the point of rotation on the quadroarticular joint (lower

jaw joint), to the tip of the most anterior tooth on the dentigerous arm of the premaxilla. Body length was measured from the midline of the posterior margin of the caudal peduncle to the tip of the lower jaw (the nasal protrusion on some preserved *C. brontotheroides* samples obscured the upper jaw). In order to remove the effects of size variation, all measurements were log transformed and regressed against log-transformed body length. We fit a log-transformed trait by log-transformed body length linear regression and used the residuals for association mapping.

Genomic sequencing and bioinformatics

DNA was extracted from muscle tissue using DNeasy Blood and Tissue kits (Qiagen, Inc.) and quantified on a Qubit 3.0 fluorometer (Thermofisher Scientific, Inc.). PCR-free Truseq-type genomic libraries were prepared using the automated Apollo 324 system (WaferGen BioSystems, Inc.) at the Vincent J. Coates Genomic Sequencing Center (QB3). Samples were fragmented using Covaris sonication, barcoded with Illumina indices, and qualitychecked using a Fragment Analyzer (Advanced Analytical Technologies, Inc.). 9-10 samples were pooled in four different libraries for sequencing on four lanes of Illumina 150PE Hiseq4000.

We mapped raw reads from 37 individuals to the *Cyprinodon* reference genome (NCBI, *Cyprinodon variegatus* annotation release 100; total sequence length = 1,035,184,475; number of scaffolds = 9,259; scaffold N50 = 835,301; contig N50 = 20,803) with the Burrows-Wheeler Alignment Tool (v. 0.7.12; (Li and Durbin 2011)). The Picard software package (http://picard.sourceforge.net (v. 2.0.1)) was used to identify duplicate reads (MarkDuplicates) and create BAM indexes (BuildBamIndex). We followed the best practices guide recommended by the Genome Analysis Toolkit (v. 3.5; (DePristo et al. 2011)) in order to call and refine our SNP variant dataset using Haplotype Caller. Filtering SNP variants in GATK for model organisms conventionally requires high-quality known variants to act as a reference. Instead we

called SNPs in our dataset using conservative hard-filtering parameters following GATK guidelines (DePristo et al. 2011; Marsden et al. 2014): Phred-scaled variant confidence divided by the depth of non-reference samples >2.0, Phred-scaled P-value using Fisher's exact test to detect strand bias > 60, Mann-Whitney rank sum test for mapping qualities (z > 12.5), Mann-Whitney rank sum test for distance from the end of a read for those with the alternate allele (z > 8.0). Further filtering was performed using VCFtools (v. 0.1.14; (Danecek et al. 2011)) to only include individuals with a genotyping rate above 90% (no individuals were excluded by this filter) and SNPs with minor allele frequencies higher than 5%. Our final filtered dataset included 12,586,315 variant sites across 37 individuals with a mean aligned read sequencing depth of 7.19 per individual (range: 5.15 – 9.28).

Population genetic analyses

Our filtered dataset was converted from Variant Call Format to PED and MAP files using VCFtools. In order to visualize population structure in our samples (McVean 2009), we performed principal component analyses using eigenvectors output by PLINK's 'pca' function (Purcell et al. 2007 (v. 1.9)). We plotted the first two principal components in R (R Core Team 2016 (v. 3.2.4)).

Genome wide F_{st} for pairwise species comparisons was calculated for each variant site using VCFtools' 'weir-fst-pop' function. Within-population nucleotide diversity (π) was estimated across 10kb windows using VCFtools' 'window-pi' function. We used a custom python script to extract allele frequencies from the VCF files which were then used to estimate between population divergence (D_{xy}) with a separate R script (provided by A. Comeault). We calculated D_{xy} across 10kb windows for ten scaffolds (totaling 9.7Mb) containing candidate SNPs for jaw size variation.

Association Mapping

We first estimated SNP × trait associations for jaw size variation using the PLINK 'assoc' function which fits a standard linear regression of phenotype on allele frequency and subsequently estimates *P*-values for each SNP with an asymptotic Wald test. We set a genome-wide level of significance using Bonferroni correction ($0.05 / 12{,}586{,}315 = 4.0 \times 10^{-9}$). Although this correction is highly conservative (Johnson et al. 2010), we are concerned here with only the most significant outliers. We then used the first two principal components explaining 9.44% of the variance in our dataset to correct for population structure by incorporating them into the model as covariates. We also performed an alternative method of mapping using a Bayesian sparse linear mixed model (BSLMM) implemented in the GEMMA software package (v.0.94.1; (Zhou et al. 2013)). GEMMA's BSLMM combines linear mixed models, which assume every genetic variant has an effect on phenotype, and sparse regression models, which assume few variants will affect the phenotype. Importantly, GEMMA controls for background population structure by estimating and incorporating a kinship relatedness matrix as a covariate in the regression model. The BSLMM uses Markov Chain Monte Carlo (MCMC) to estimate the proportion of phenotypic variation explained by every SNP included in the analysis (PVE), the proportion of phenotypic variation explained by SNPs of large effect (PGE), which are defined as SNPs with a non-zero effect on the phenotype, and the number of large-effect SNPs needed to explain PGE (nSNPs). GEMMA calculates an effect size coefficient (β) and a posterior inclusion probability (PIP) for each SNP. Markers with non-zero values of β are inferred to affect phenotypic variation in one iteration of the MCMC sampler. β can be a positive or negative integer based on the direction of association, so we present estimates of this parameter in terms of its absolute value. PIP reports the proportion of iterations in which a SNP is estimated to have

a non-zero effect on phenotypic variation ($\beta \neq 0$). This estimate might be difficult to interpret for SNPs in high linkage disequilibrium (LD) because tightly linked neutral and causal SNPs could each have a high probability of inclusion in separate iterations. We estimated pairwise LD (r^2) between SNPs on the largest scaffold (4.5 Mb) and found that linkage dropped to background levels between SNPs separated by more than 20kb ($r^2 < 0.1$) (Fig. A2.4). Thus, we summed β and PIP parameters across 20kb windows to account for any unwanted dispersion of these values across SNPs in LD.

We performed 10 independent runs of the BSLMM for all 37 individuals (following (Comeault et al. 2014)) using a step size of 100 million with a burn-in of 50 million steps. We used GEMMA to assess the significance of regions associated with jaw size variation and report the median β and PIP summed across windows for the 10 independent MCMC runs. Independent runs were consistent in reporting the strongest associations for the same 20kb windows. In order to compare the abundance and effect size of candidate loci between specialist species, we plotted the frequency of β estimates for regions with effects on smaller jaws (negative β) and larger jaws (positive β).

Identification of candidate genes

We restricted our search to those regions both fixed between species and associated with jaw size. Accordingly, candidate regions met two rigorous criteria: 1) they must contain one or more SNPs that are fixed in at least one pairwise species comparison and 2) show significant association with jaw size in both association mapping analyses ($P < 4.0 \times 10^{-9}$ and outlier PIP estimates above the 99th percentile). We also took advantage of a recent linkage mapping analysis of phenotypic diversity in San Salvador *Cyprinodon* pupfish by comparing our

candidate regions for overlap with the four scaffolds containing QTL with moderate effects on jaw size in an F_2 intercross between specialists (Martin et al. 2017).

In addition to our candidate regions, we also report association mapping statistics and gene annotations for all 22 SNPs fixed between the generalist and scale-eater species. We used the Phenoscape Knowledgebase (Mabee et al. 2012; Midford et al. 2013) to determine if any of the annotated genes within fixed SNP regions were associated with skeletal system phenotypes across model taxa.

Detecting Selection and Demographic History

We first calculated Tajima's D for each species in 10kb genomic windows using VCFtools' 'TajimaD' function. This statistic compares observed nucleotide diversity to diversity under a null model assuming genetic drift, where negative values indicate a reduction in diversity across segregating sites that may be due to positive selection (Tajima 1989). Second, we used the SweepFinder method first developed by Nielsen et al. 2005 and implemented in the software package SweeD (Pavlidis et al. 2013). SweeD scans across non-overlapping windows to calculate a composite likelihood ratio (CLR) using a comparison between two contrasting models. The first assumes a window has undergone a recent selective sweep, while the second assumes a null model where the site frequency spectrum of the window does not differ from that of the entire scaffold. Windows with high CLR suggest a history of selective sweeps because the site frequency spectrum is shifted toward low and high frequency derived variants (Nielsen et al. 2005a; Pavlidis et al. 2013).

Various demographic histories can shift the distribution of low and high frequency derived variants to falsely resemble signatures of hard selection (Galtier et al. 2000; Nielsen 2005). In order to account for demography, we used the Multiple Sequentially Markovian

Coalescent (MSMC) (Schiffels and Durbin 2014) to infer historical effective population sizes (N_e) in all three species. MSMC is an extension of the Pairwise Sequentially Markovian Coalescent (PSMC) (Li and Durbin 2011), which uses a hidden Markov model to scan genomes analyzing patterns of heterozygosity where long DNA segments with low heterozygosity reflect recent coalescent events. The rate of coalescent events is then used to estimate N_e at a given time. We ran MSMC on unphased GATK-called genotypes from the 100 largest scaffolds for each individual separately, thus using only two haplotypes as in PSMC (the analysis of multiple individuals simultaneously would inform on more recent timescales, but requires phasing). As recommended in the MSMC documentation, we masked out sites with less than half or more than double the mean coverage for that individual, with a genotype quality below 20. We also excluded sites with less than 10 reads as recommended by Nadachowska-Brzyska et al. (2016). Nadachowska-Brzyska et al. (2016) also recommend to only use individuals with a mean coverage of at least 18× (Nadachowska-brzyska et al. 2016). However, all our individuals were sequenced at a lower coverage and we included only the seven individuals with a coverage of at least 7.5×. This means that our MSMC results should be interpreted with caution; however, the consistency among individuals of the same species (see Fig. A2.3) suggest that the general patterns of the analysis are likely to be robust.

To scale the output of MSMC to real time and population sizes, we assumed a six-month generation time (Martin 2016b) and a mutation rate measured for cichlids (6.6×10^{-8} mutations per site per year, (Recknagel et al. 2013)), one of the most closely related fish groups with an available estimate of spontaneous mutation rates.

We used ancestral population sizes determined by MSMC to analytically calculate the expected neutral site frequency spectrum with SweeD. We used the '-eN' flag to model a 100-

fold population decrease around 10,000 years ago (20,000 generations). We used a grid size of 1kb across our folded SNP dataset which defined sites as ancestral or derived variants based on the major and minor allele frequencies. We also ran SweeD without demographic assumptions for comparison. Because the significance of the CLR depends on the background site frequency spectrum of each scaffold, we compared the percentile of each likelihood estimate across unique scaffolds for candidate regions. Windows that showed CLRs above the 95^{th} percentile across their respective scaffolds under the assumptions of a population decrease determined by MSMC were interpreted as regions that recently experienced a hard sweep.

The size of the scaffolds containing jaw size candidate loci should be large enough to discover regions under strong selection. Out of our 31 candidate regions, we excluded one because it fell within a small scaffold that could not be used to sample an adequate background distribution of heterogeneity. Of the 25 scaffolds containing the 31 regions we analyzed with SweeD, the mean scaffold length was 863,416bp. Furthermore, we set a conservative threshold ($>95^{th}$ percentile) to define regions that have experienced hard sweeps. We plot π, D_{xy}, and Tajima's D across 10kb windows using a cubic smoothing spline in R.

Results

Estimating phenotypic distances

Orr's extension of Fisher's geometric model predicts that *de novo* mutations with a large effect on phenotypic variation are more likely to be fixed during adaptation toward distant phenotypic optima than nearby optima (Orr 1998, 2005). In order to test this prediction, we measured the phenotypic distance between hybrids used to estimate the multi-peaked adaptive landscape for San Salvador pupfishes (dataset published in Dryad repository for Martin 2016b

and originally used for Martin and Wainwright 2013b). These hybrids were measured for 16 morphological traits. We measured the distance between fitness peaks in all 16 trait dimensions (standardized and size-corrected to a mean of zero, standard deviation of 1 as in the original studies) and found that the distance between phenotypic optima is greater between the generalist fitness peak and the fitness valley surrounding hybrid phenotypes most resembling the scale-eaters than between the generalist fitness peak and the neighboring higher fitness peak corresponding to hybrids resembling the snail-eater (Fig 1).

Population structure and genome scans

Principal component analysis revealed population structure at the level of species and individual lake population, with the top two principal components together explaining 9.44% of the genetic variation (Fig. 1.2A). The axes show two distinct clusters of scale-eaters: smaller-jawed individuals from Osprey Lake, Great Lake, and Oyster Pond and the largest-jawed individuals from Crescent Pond and Little Lake. Genome-wide mean estimates of within-species diversity (π: generalist = 0.00402, snail-eater = 0.00321, scale-eater = 0.00324) and mean between-population divergence (D_{xy}: generalist × snail-eater = 0.000166, generalist × scale-eater = 0.000169, scale-eater × snail-eater = .000167) were similar for all comparisons, revealing that most variants were shared among species. The similarity between D_{xy} among species suggests that divergence from a generalist ancestor likely occurred near the same time for both specialists.

We used genome-wide F_{st} scans to identify fixed regions associated with each species across nine lake populations on San Salvador and one neighboring island. Very few fixed sites corresponded to the discrete species-specific phenotypes across populations. We found 6,673 sites fixed between specialists, 123 sites fixed between generalist and snail-eater species, and a mere 22 sites fixed between generalist and scale-eater species (Fig. 1.3, Table A1.1). Eight of

these 22 fixed SNPs were also fixed between specialists. Genome-wide mean F_{st} estimates for each comparison (scale-eater/snail-eater = 0.143, generalist/snail-eater = 0.080, generalist/scale-eater = 0.089) were comparable to previous estimates based on microsatellites (Turner et al. 2008) and RADseq derived SNPs (Martin and Feinstein 2014).

Association Mapping

We initially used quantitative trait association mapping in PLINK to identify SNPs associated with jaw length variation among individuals without correcting for population structure, which would remove true positives in addition to false positives. This uncorrected PLINK analysis identified 9,214 variants associated with jaw size variation between the generalist, scale-eater, and snail-eater species ($P < 4.0 \times 10^{-9}$ (Fig. 1.4)). Of these variants, 556 were fixed in at least one pairwise species comparison. 555 of these SNPs were fixed between the two specialists; nine were fixed between the generalist and scale-eater; zero were fixed between the generalist and snail-eater.

Out of the nine PLINK outlier SNPs significantly associated with jaw size and fixed between the generalist and scale-eater, six were located across four different gene regions (*magi3, cabp2*, *lingo1*, and *pigr*) and three unannotated regions (Table A1.1). Out of the top 20 outliers fixed between the snail-eater and scale-eater, 13 were located across five different gene regions (*galr2, gmds, soga3, tmem30a, plxna2*) and seven were located across three unannotated regions (Table 1.1). Combined, PLINK identified 14 divergent regions (nine genic and five unannotated) significantly associated with jaw size and fixed in scale-eaters.

We further assessed the significance of jaw size associations for these top candidate regions containing fixed SNPs by correcting for population structure using two methods. First, we used PLINK to include the top two principal components as covariates in the model (Price et

al. 2006; Hunter et al. 2007). This stringent analysis did not identify any SNPs associated with jaw size at our highly conservative Bonferroni-corrected significance threshold (Table A1.2). However, this likely reflects the fact that the first principal component is significantly correlated with jaw size ($P = 0.0013$, Fig S1). Next, we performed independent association mapping with GEMMA, which corrects for population structure by incorporating a genetic relatedness matrix as a covariate in a Bayesian sparse linear mixed model (Zhou et al. 2013). This is a more reliable correction for population structure because the relatedness matrix accounts for pairwise relatedness between individuals; whereas principal components only capture broad linear axes of population structure (Novembre and Stephens 2008; Kang et al. 2010). Because the uncorrected PLINK analysis likely identified a subset of true associations in addition to false positives, we chose to combine uncorrected PLINK results with our corrected GEMMA results in order to evaluate the significance of regions associated with jaw size (following (Zhou and Stephens 2012)). We identified 31 regions (20kb each) implicated by uncorrected PLINK analyses that also showed association with jaw size after correcting for population structure in GEMMA (Fig. 1.4). We assessed the significance of associations based on PIP (posterior inclusion probability) parameters which report the proportion of iterations in which a SNP is estimated to have a non-zero effect on phenotypic variation (effect size $\beta \neq 0$). These 31 regions showed robust association across 10 independent Markov Chain Monte Carlo (MCMC) runs. We used β effect size parameters to assess whether regions contributed to larger jaw size ($+\beta$) or decreasing jaw size ($-\beta$) and found slightly more candidate regions increased (16) than decreased jaw size (13).

All 31 regions contained variants fixed between specialists and showed outlier median parameter values in the 99th percentile for PIP estimated across all SNPs included in the analysis, indicating an association with jaw size after accounting for population structure (Table 1.1)

(following (Gompert et al. 2012)). These regions span 25 scaffolds and contain 29 genes, 11 of which are annotated for skeletal system functions (NCBI *Cyprinodon* release 100). The top ten regions with the highest PIP implicated three of the same genes identified by PLINK (*galr2, gmds, soga3*) as well as three additional genes (*fam49b, znf664, and pard3*) and one large (60kb) unannotated region. The unannotated region and *galr2* showed the highest β values in the direction of large jaws, while the region containing *gmds* showed the highest β values in the direction of smaller jaws (Fig. 1.5, 1.6). Encouragingly, *galr2* is within a QTL explaining 15% of the variation in jaw size in an F_2 intercross between specialist species (Martin et al. 2017).

History of Selection and Demography

To determine whether candidate regions were potentially subject to hard selective sweeps, we interrogated the site frequency spectrum using SweeD (Pavlidis et al. 2013) and Tajima's D (Tajima 1989). Tajima's D compares observed nucleotide diversity to diversity under a null model assuming genetic drift, where negative values indicate a reduction in diversity across segregating sites (Tajima 1989). SweeD scans across non-overlapping windows to calculate a composite likelihood ratio (CLR), comparing a model assuming selection to a null model calibrated by the observed site frequency spectrum across the entire scaffold. Both of these statistics infer selection based on the shape of the site frequency spectrum, which can also be influenced by changes in effective population size over time (Galtier et al. 2000; Nielsen 2005). We therefore used the Multiple Sequentially Markovian Coalescent (MSMC) (Schiffels and Durbin 2014) to infer historical population size in all three species, and applied these estimates to analytically calculate the expected neutral site frequency spectrum in SweeD. MSMC results suggest that that the population size of all three species has been decreasing across at least the last 10,000 years (~20,000 generations) (Fig. A2.3). This model suggests a

population decrease after a lake colonization event that is consistent with changes in sea level during the last glacial maximum which would have dried out the saline lakes on San Salvador Island (Mylroie, J.E, Hagey 1995; Turner et al. 2008). We first looked for signatures of hard sweeps in both specialist populations by analyzing the site frequency spectrum without demographic assumptions. Next, we calculated the expected neutral site frequency spectrum assuming a population decline as suggested by our demographic model. Windows that showed CLRs above the 95th percentile across their respective scaffolds in this second analysis were interpreted as regions that recently experienced a hard sweep.

Out of our 31 candidate regions affecting jaw size, six were consistent with hard selective sweeps. One candidate region was excluded from these analyses because it fell within a small scaffold that could not be used to sample an adequate background distribution of heterogeneity. All six regions also showed negative estimates of Tajima's D (Fig. 1.5, 1.6). The 60kb unannotated region associated with large jaws showed the strongest signatures of selection, followed by a 40kb region associated with small jaws. This smaller region contains four genes all annotated for skeletal system effects (*hint1, lyrm7, dync2li1, abcg5*) (Fig. 1.6B). Five of the six regions that experienced strong selection also show reduced within-population diversity (π) in the specialist species and increased between-population divergence (D_{xy}) when compared to generalists (Fig. 1.5,1. 6). This pattern may suggest that strong selection on a beneficial allele reduced diversity within specialists across candidate regions. Importantly, low diversity in these regions is not shared between specialists and generalists, possibly suggesting that selection unique to each specialist was responsible for reduced diversity. This combined evidence implicates divergent regions influencing jaw morphology that experienced strong selection within the specialist linages. Finally, we did not find evidence for hard sweeps in 25 of our 31

candidate regions, possibly suggesting that multiple haplotypes were swept to fixation (Hermisson and Pennings 2005; Jensen 2014).

More large-effect alleles were required to evolve large jaws than small jaws

Based on differences in the phenotypic distance across fitness valleys separating each specialist species from its putative generalist ancestor (Fig. 1.1), we predicted to find more large-effect SNPs associated with large jaws than small jaws. There are two lines of evidence supporting this prediction. First, we directly compared positive and negative effect sizes for regions associated with small jaws (-β) and large jaws (+β). Our β outlier threshold included 83 of the regions most strongly associated with jaw size that had the largest effects on jaw size (β > 99.9th percentile). We found more than twice as many outlier SNPs with large effects on increasing jaw size (n = 56) compared to large-effects on decreasing jaw size (n = 27) (Fig. 1.7). Second, we identified five times fewer SNPs fixed between the generalist and scale-eater (n = 22) than SNPs fixed between the generalist and snail-eater species (n = 123) (Fig. 1.3), supporting the prediction that SNPs with larger effect sizes should fix faster than SNPs with smaller effects, especially given short divergence times (Griswold 2006; Yeaman and Whitlock 2011).

Discussion

Genome-wide divergence scans revealed that the evolution of trophic novelty in two ecological specialists involved surprisingly few genetic variants fixed between species. We determined which of these fixed variants influenced the most rapidly diversifying trait in this radiation – jaw size – using quantitative trait association mapping. We uncovered 31 candidate regions fixed between species and associated with jaw size after correcting for population

structure, with six of these regions showing signs of hard selective sweeps. We used these data to test the prediction that more large-effect variants should affect large jawed scale-eaters than small jawed snail-eaters.

Genetic Basis of Jaw Size Divergence

We report 31 divergent candidate regions associated with jaw size among San Salvador *Cyprinodon* pupfish. We identified these regions using 37 genomes sequenced to 7x coverage across nine populations. This is significant because the majority of work on the genetic basis of adaptation has relied on reduced representation strategies (*i.e.* RAD-seq, RNA-seq) that likely overlook loci contributing to adaptation (Hoban et al. 2016). All 31 regions contained SNPs fixed between specialists that were significant in both association mapping approaches. We searched genes listed under the 'skeletal system' ontology in the phenotype database Phenoscape (Mabee et al. 2012; Midford et al. 2013; Manda et al. 2015) finding matches for 11 genes within candidate regions (Table 1.1). The most strongly associated gene annotated for skeletal effects, *galr2*, is interesting for several reasons. The protein product of *galr2* is a transmembrane galanin receptor with a role in numerous physiological functions (Webling et al. 2012). Galanin, the binding substrate of GALR2, has been shown to facilitate bone formation by increasing the size and proliferation of osteoblasts (McDonald et al. 2007; McGowan et al. 2014). Additionally, the scaffold containing *galr2* overlaps with a moderate effect QTL explaining 15% of the variation in jaw size in an independent F_2 mapping cross between the two specialist pupfishes (Martin et al. 2017), increasing confidence in our association mapping strategy. The gene region most associated with smaller jaws was *gmds*, which is important for tagging cell surface proteins involved in many cellular processes such as cell growth, migration, and apoptosis (Moriwaki et al. 2009). This gene represents a novel candidate for craniofacial effects. We identified four

genes annotated for skeletal effects spanning a 40kb region that showed significant association with smaller jaws (*hint1, lyrm7, dync2li1, abcg5*). Mutations in *lyrm7* have been associated with mitochondrial complex III deficiency, a disorder characterized by skeletal muscle weakness and weak muscle tone (hypotonia) (Invernizzi et al. 2013). Mutations in *dync2li1*, a gene involved in skeletogenesis and expressed in the cartilage of growth plates, have been shown to cause short rib polydactyly skeletal disorders (Taylor et al. 2015). Thus, our candidate regions are associated with genes involved in bone and skeletal muscle development – the two tissues most differentiated in the external anatomy of San Salvador pupfishes. Finally, we identified eight SNPs fixed between the generalist and scale-eater that were also fixed between specialists, possibly indicating that these regions affect traits in both specialists. However, none of these overlapping SNPs showed significant association with jaw size after correcting for population structure.

Caveats to our association mapping approach

The significance of our association mapping results should be interpreted with caution. Our principal component analysis revealed significant population structure associated with four different clusters of jaw sizes across species and between two different clusters of large and short-jawed scale-eaters among lake populations (Fig. 1.2A), which likely created a bias toward false positive associations implicated by PLINK. Furthermore, when we accounted for this structure by incorporating the first two principal components as covariates in the model, we did not find any SNPs reaching significance at our conservative Bonferroni-corrected level of significance. However, this analysis almost certainly filtered out true associations because the first PC is highly correlated with jaw size. We reassessed the significance of these associations by using GEMMA – a complementary mapping approach that corrects for population structure

by incorporating a genetic relatedness matrix into a Bayesian sparse linear mixed model (BSLMM) (Zhou et al. 2013). We used the BSLMM to investigate the genetic architecture of jaw size – a complex polygenic trait (Albertson et al. 2003; Helms and Schneider 2003; Pallares et al. 2014; Porto et al. 2016; Martin et al. 2017). Our PIP estimates for regions associated with jaws size variation suggest that jaw shape is controlled by many loci of relatively small effect (see (Comeault et al. 2016) for an example of BSLMMs used for mapping a simple Mendelian color locus). Indeed, a linkage mapping analysis of phenotypic diversity in an F_2 intercross between specialists identified QTL with only moderate effects explaining up to 15% of the variation in jaw size (Martin et al. 2017).

While uncommonly implemented across species, association mapping techniques have proven successful at identifying associations across 'varieties,' 'subspecies,' and 'ecotypes' with greater genetic differentiation (Fournier-Level et al. 2011; Zhao et al. 2011; Pallares et al. 2014) or minimal divergence similar to that of San Salvador pupfishes (Comeault et al. 2014). Association mapping within populations may result in spurious associations due to background population structure (Marchini et al. 2004; Kang et al. 2010), but our sampling of multiple, relatively isolated populations may have provided greater resolution of candidate regions due to sampling a diversity of genetic backgrounds. We do not expect false associations due to sequencing error biases because mean coverage across candidate SNPs mirrored coverage across individuals (range: 4.9x – 6.6x). It is possible that our methods excluded significant SNPs as false negatives. We examined the position of all 22 SNPs fixed between the generalist and scale-eater for gene annotations (Table A1.1), finding four within the gene *col11a1*. None of these four SNPs showed a significant association with jaw size in either mapping approach; however, *col11a1* has been associated with jaw skeleton phenotypes in humans (Hufnagel et al. 2014). It is

unclear whether *col11a1* variants influence jaw divergence in pupfish but escaped detection in both mapping analyses.

Variants with relatively large effects drive divergence across a large fitness valley

Orr's extension of Fisher's geometric model of adaptation predicts that *de novo* mutations with a large effect on phenotypic variation are more likely to be fixed during adaptation toward distant phenotypic optima than nearby optima (Orr 1998, 2005). This distribution of effect sizes for mutations fixed during adaptation has been supported by QTL mapping analyses in multiple systems (Baxter et al. 2009; Rogers et al. 2012; Martin et al. 2017). We show that the phenotypic distance across the fitness valley is larger between the generalist and large-jawed scale-eater species than between the generalist and small-jawed snail-eater species (Figs. 1, S1) (Martin and Wainwright 2013b; Martin 2016a). Based on this adaptive landscape, we predicted more large-effect variants associated with large jaws than small jaws. Adaptive landscapes are not static, and the distance between fitness optima may have fluctuated over the past 10,000 years of divergence in this system (Hansen et al. 2008). However, scale-eater prey has been available since the initial colonization of generalists on San Salvador. Furthermore, the availability of hard-shelled prey (ostracods, gastropods), is likely not substantially depleted in these lakes due to the rarity of snail-eater specialists (<5% of the total pupfish population) and high productivity of eutrophic saline lakes (Martin and Wainwright 2013b).

Although Orr's model assumes a single population and ignores standing genetic variation (Orr 1998; Dittmar et al. 2016) and thus may not apply here, we present two lines of evidence supporting the model in this system. First, we found twice as many outlier regions with the largest effect sizes associated with larger jaws than smaller jaws (Fig. 1.7). Second, there are more than five times as many fixed SNPs between the generalist and snail-eater than between the

generalist and scale-eater (Fig. 1.3). Divergent demographic histories could account for this pattern; however, similar changes in population size over 20,000 generations for each species (Fig. A2.3), combined with evidence for gene flow between species in sympatry (Martin and Feinstein 2014), suggest that this is not the case. Large-effect variants are predicted to become fixed between species more quickly than variants with smaller effects in the presence of gene flow, especially when divergence time is short (Griswold 2006; Yeaman and Whitlock 2011). This difference suggests that more large-effect alleles influencing jaw size were necessary to evolve the specialized scale-eating phenotype, while smaller jaw phenotypes may result from more alleles with small to moderate effect sizes. Further support for this prediction within the San Salvador pupfish system comes from a complementary linkage mapping study which found moderate effect QTL explaining up to 15% of variance in jaw size within an F_2 intercross between both specialists but no significant QTL with effects on nasal protrusion – a trait unique to the snail-eater species (Martin et al. 2017). Overall these data agree with Orr's model, suggesting that large effect loci are used to cross larger distances between fitness optima (Orr 1998, 2005).

Strong selection on candidate regions

We reasoned that strong selection on variants within candidate genes would be necessary for extreme shifts in ecological specialization. This can result in a pattern of hard selective sweeps resulting from a single haplotype rising quickly to fixation in a population derived from *de novo* mutation or standing variation (Orr and Betancourt 2001; Jensen 2014). Alternatively, a soft sweep occurs when selection drives multiple adaptive haplotypes to fixation – a pattern that can only result from selection on standing variation (Hermisson and Pennings 2005; Jensen 2014). Currently there are no theoretical predictions about the likelihood of adaptation from

standing genetic variants versus *de novo* mutation for populations with small values of within-population divergence such as ours (Dittmar et al. 2016), and the relative importance of hard sweeps versus soft sweeps during adaptation is a subject of much debate (Hermisson and Pennings 2005; Pritchard et al. 2010; Jensen 2014; Garud et al. 2015; Schrider et al. 2015). In order to investigate whether regions associated with large jaws experienced hard sweeps, we examined the site frequency spectrum across candidate regions looking for signature shifts in variant frequencies across scaffolds.

Changes in ancestral population size can produce similar signals to hard selective sweeps. To account for this, we first estimated the effective population size of all three species over the past 20,000 generations and observed a 100-fold population decrease occurring within the same time as we predict ancestral populations colonized lakes on San Salvador Island (Mylroie, J.E, Hagey 1995; Turner et al. 2008; Martin and Wainwright 2013a). We next calculated a neutral site frequency spectrum under this bottleneck scenario and still detected hard sweeps in six of our candidate regions (three contributing to smaller jaws and three to larger jaws) (Fig. 1.5, 1.6). Regions containing *hint1*, *lyrm7, dync2li1,* and *abcg5* along with a large unannotated region showed the strongest signs of hard sweeps after accounting for demographic history (Fig. 1.6B). Low estimates of Tajima's D, low nucleotide diversity in specialists, and high divergence between specialists and generalists lend further support for past selection at these loci (Tajima 1989; Nielsen 2005; Cruickshank and Hahn 2014). Alternatively, low recombination rates could account for low nucleotide diversity and high divergence at these loci (Nachman and Payseur 2012). A decrease in population size can also reduce genome-wide nucleotide diversity (Tajima 1989; Galtier et al. 2000). However, our demographic analysis show comparable decreases in population size for the generalist and specialist populations, making nucleotide diversity

comparable across species. Interestingly, 25 of our 31 strongest candidate regions do not show signs of hard selective sweeps. This may support a history of soft selective sweeps, where beneficial standing genetic variants were swept to fixation resulting in multiple haplotypes at candidate loci (Hermisson and Pennings 2005; Jensen 2014).

Conclusions

The San Salvador *Cyprinodon* pupfish radiation has proven itself as an excellent system for investigating the genetic basis of novel trophic specialization. The extensive phenotypic diversity among these species results from low levels of genetic divergence and very few fixed variants. 31 regions with fixed variants showed significant associations with jaw size – the most rapidly diversifying trait in this system. Selection scans across regions associated with jaw size revealed a history of novel adaptation driven in part by hard selective sweeps. Additionally, we identified more variants with larger effects used to adapt to a more distant phenotypic optimum – consistent with Orr's model of adaptation. Our evidence for the evolution of larger jaw size raises an alluring question with broad implications for research on adaptation: why has trophic novelty evolved exclusively on San Salvador Island? It is surrounded by islands with comparable physiochemistry, lake areas, macroalgae communities, and generalist *Cyprinodon* pupfish populations that exhibit similar genetic, phenotypic, and dietary diversity to generalist populations on San Salvador Island. This is consistent with similar levels of ecological opportunity on neighboring islands without specialists (Martin 2016a). Nonetheless, scale-eating and snail-eating species appear to be endemic to a single island. Answering this question will require continued exploration of the ecological and genetic factors shaping this exceptional case of rapid ecological specialization.

Table 1.1. Jaw size association statistics and gene annotations for SNPs fixed between *C. desquamator* and *C. brontotheroides*.

Fixed SNPs fall within 20kb windows showing significant association with jaw size after controlling for population structure (Median PIP > 99th percentile). Asterisks (*) show SNPs in gene regions (bold) annotated for skeletal system effects. A cross (ŧ) indicates overlap with a scaffold within a QTL affecting jaw size.

SNP	Scaffold	Median PIP	PIP percentile	Median β	*P*	Gene region
1	KL652649.1	0.01795	1.0000	7.764633	1.82E-10	-
2	KL652649.1	0.0124	0.9999	4.10637	3.29E-10	-
3	KL653712.1	0.00975	0.9999	1.036102	6.65E-11	FAM49B/ZNF664
4	KL653062.1	0.0076	0.9999	-2.32365	3.82E-13	GMDS
5*	KL652786.1	0.0069	0.9998	2.207843	6.66E-12	**GALR2**
6	KL652758.1	0.0066	0.9998	-1.15222	1.60E-11	SOGA3
7*	KL652786.1	0.00625	0.9998	2.018056	1.41E-10	**GALR2**
8	KL652715.1	0.0058	0.9998	-2.21671	1.62E-09	PARD3
9	KL652649.1	0.0052	0.9998	2.223139	1.05E-10	-
10*	KL653271.1	0.0052	0.9996	0.291561	2.05E-10	**ELN**
11*	KL652666.1	0.0043	0.9995	-0.33468	5.30E-10	DYNC2LI1/**ABCG5**
12 *	KL654513.1	0.00405	0.9994	-0.99172	1.24E-11	**PLAUR**
13*	KL653122.1	0.004	0.9994	1.029314	5.63E-10	**ATP8A1**
14	KL653046.1	0.0039	0.9993	1.189392	3.43E-09	LRP1B
15	KL653805.1	0.0038	0.9991	0.473089	1.23E-09	-
16*	KL652666.1	0.0037	0.9986	0.368651	1.98E-09	**LYRM7/DYNC2LI1/HINT1**
17	KL652617.1	0.0035	0.9983	1.517635	1.48E-12	PLXNA2
18	KL652527.1	0.0034	0.9983	0.140283	8.12E-13	TMEM30A/FILIP1L
19*	KL652983.1	0.0034	0.9981	-0.76248	1.06E-09	**SKI**
20	KL653291.1	0.00335	0.9977	-0.0425	4.74E-11	-
21	KL652991.1	0.0032	0.9967	-0.66796	3.12E-09	DLGAP1
22	KL653356.1	0.003	0.9961	0.979591	3.95E-12	-
23	KL653356.1	0.00295	0.9952	1.580411	4.39E-10	-
24	KL653706.1	0.00285	0.9947	-0.80369	1.60E-09	PLECKHG6
25	KL653420.1	0.0028	0.9940	-0.93815	7.16E-11	-

26	KL652585.1	0.00275	0.9936	1.384922	8.98E-11	FAM172A
27	KL654513.1	0.0027	0.9927	-0.41968	5.95E-12	-
28*	KL653925.1	0.00265	0.9927	-0.50498	1.15E-10	B3BNT3/**B3GNT2**
29	KL652727.1	0.00265	0.9927	0.075912	1.25E-09	RABGAP1
30	KL653654.1	0.00265	0.9919	0.056305	1.96E-09	COL15A1
31*	KL652717.1	0.0026	0.9919	-0.22127	4.65E-10	ASH1L/DAP3/**GBA**

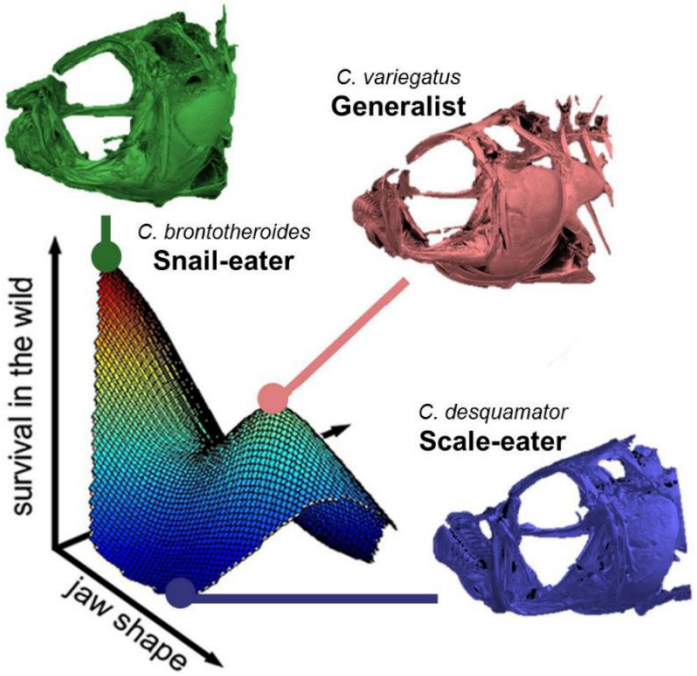

Figure 1.1. Survival Fitness Landscape for San Salvador Cyprinodon Pupfish.

A) *C. variegatus* (red), *C. desquamator* (blue), and *C. brontotheroides* (green) from each lake population were intercrossed in every direction to produce F_2 hybrids which were left for three months in an enclosure on San Salvador. Survival probability is plotted against two axes of the discriminant morphospace, indicating a wide range of jaw phenotypes in the F_2 hybrids (modified from Martin and Wainwright, 2013). Heat colors correspond to survival probability (with blue being low and red being high). MicroCT scans of the cranial skeleton of each species modified from Hernandez et al. (in revision).

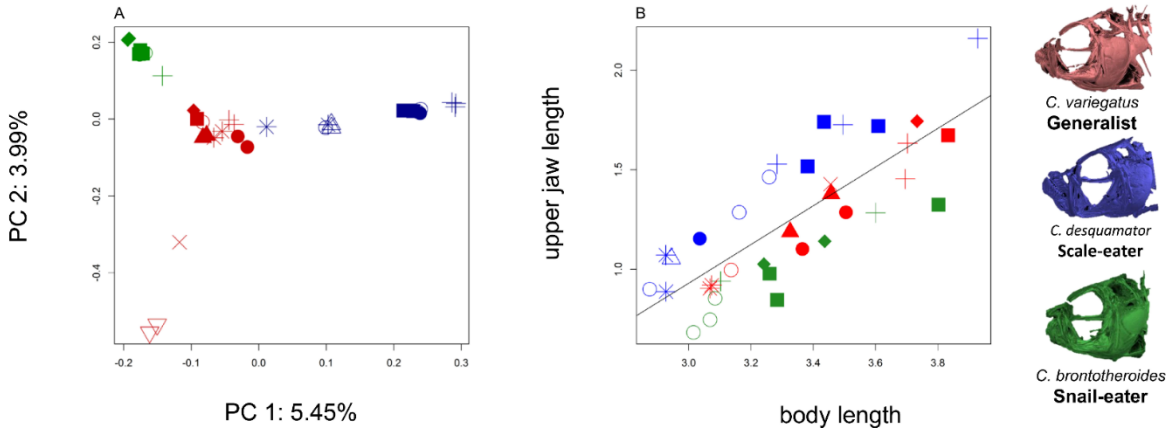

Figure 1.2. Standardized jaw size and population structure.

A) Principal component analysis showing axes accounting for a combined 9.45% of the total genetic variation between samples from 12 million SNPs genotyped from 37 whole-genome sequences. B) Log-transformed upper jaw length (mm) standardized by log-transformed body length for *C. variegatus* (red), *C. desquamator* (blue), and *C. brontotheroides* (green). Symbols represent individual lake of origin. MicroCT scans of the cranial skeleton of each species, modified from Hernandez et al. (in revision). + = Crescent Pond, × = Lake Cunningham, ▲= Mermaid's Pond , ■ = Little Lake, ○ = Osprey Lake , ● = Stout Lake , ∗ = Great Lake, ◆ = Moon Rock, ▽ = Pigeon Creek, △ = Oyster Lake).

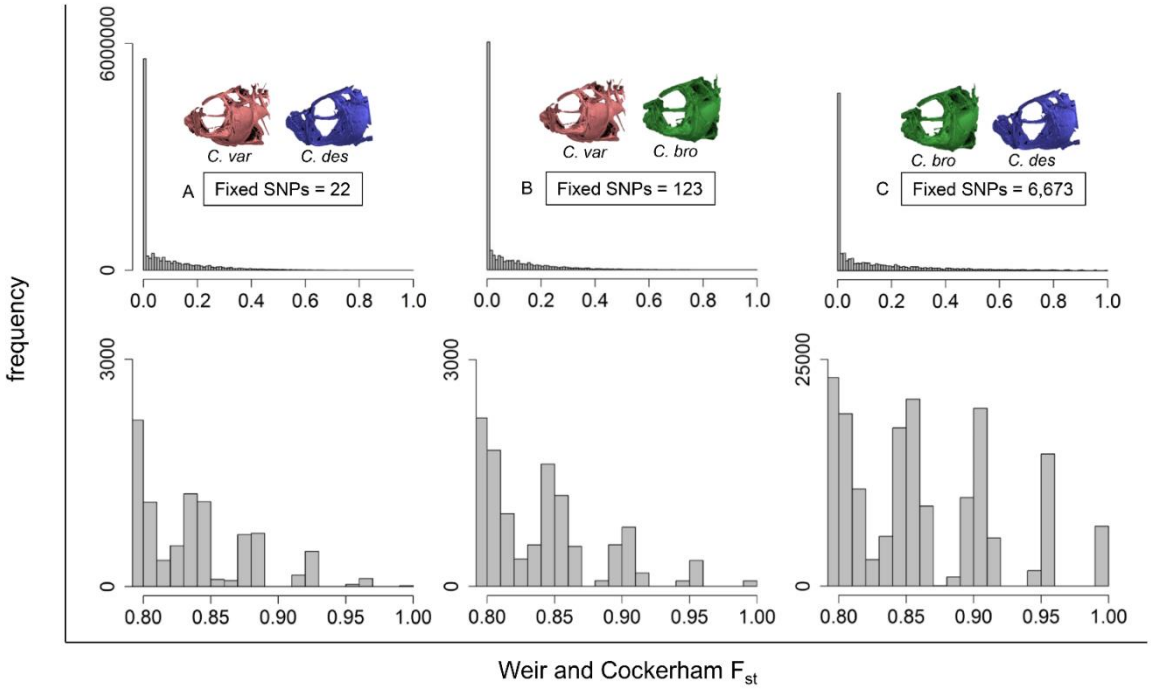

Figure 1.3. Fst distribution across 9,259 scaffolds.

Upper panels show the distribution of genome-wide per-site F_{st} estimates for 12,586,315 SNPs across all *Cyprinodon* scaffolds for A) *C. variegatus* vs. *C. desquamator* (28 individuals from ten lake populations), B) *C. variegatus* vs. *C. brontotheroides* (24 individuals from nine lake populations), and C) *C. brontotheroides* vs. *C. desquamator* (23 individuals from six lake populations). Lower panels show the distribution of SNPs with F_{st} estimates greater than 0.80.

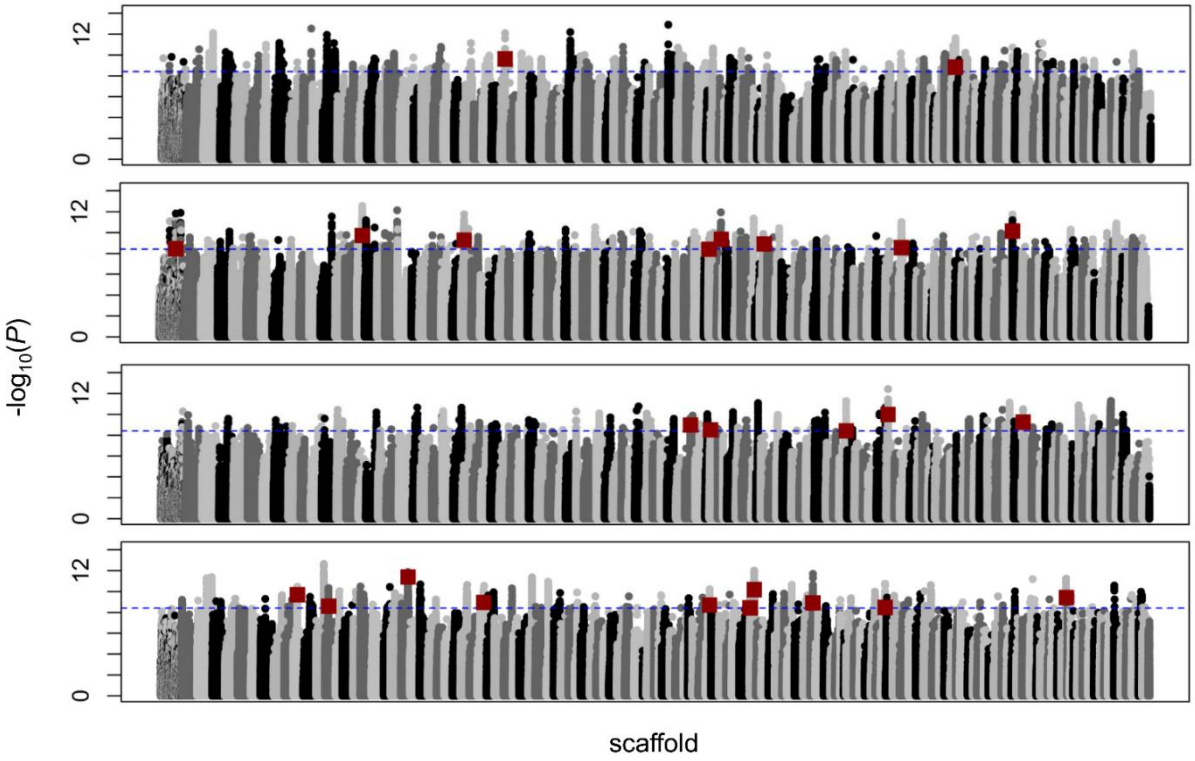

Figure 1.4. Quantitative Trait Association Mapping.

Log-transformed *P* values for 12,586,315 SNP associations with jaw size variation estimated by PLINK (n = 37 individuals). Dotted blue line shows Bonferroni-corrected level of significance ($P < 4.0 \times 10^{-9}$). Red squares show the 31 SNPs spread across 25 scaffolds most strongly associated with jaw size that are also fixed between specialists.

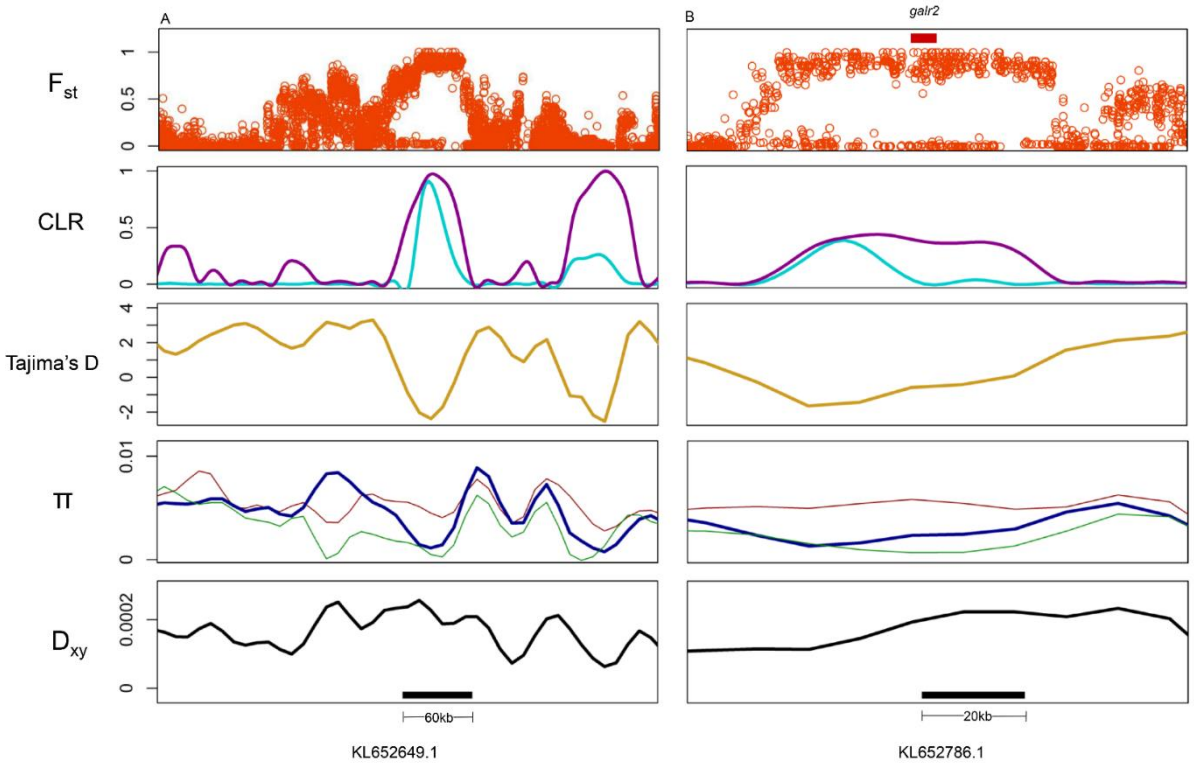

Figure 1.5. Candidate regions associated with large jaw size.

Row 1 shows individual SNP F_{st} values between *C. variegatus*/*C. desquamator*. Row 2 shows composite likelihood ratios estimated by SweeD using an analytical site frequency spectrum assuming a population bottleneck (magenta) and a frequency spectrum calculated without demographic assumptions (cyan) for *C. desquamator*. Row 3 shows Tajima's D (dark yellow) for *C. desquamator*. Row 4 shows within-population diversity (π) (red: *C. variegatus*, green: *C. brontotheroides*, blue: *C. desquamator*). Row 5 shows between-population divergence (D_{xy}, black) for *C. variegatus*/*C. desquamator*. Black bars in row 4 show windows containing fixed SNPs that showed significant association with jaw size in both PLINK and GEMMA association mapping analyses. Red bars in row 1 show exonic and intronic gene regions within windows.

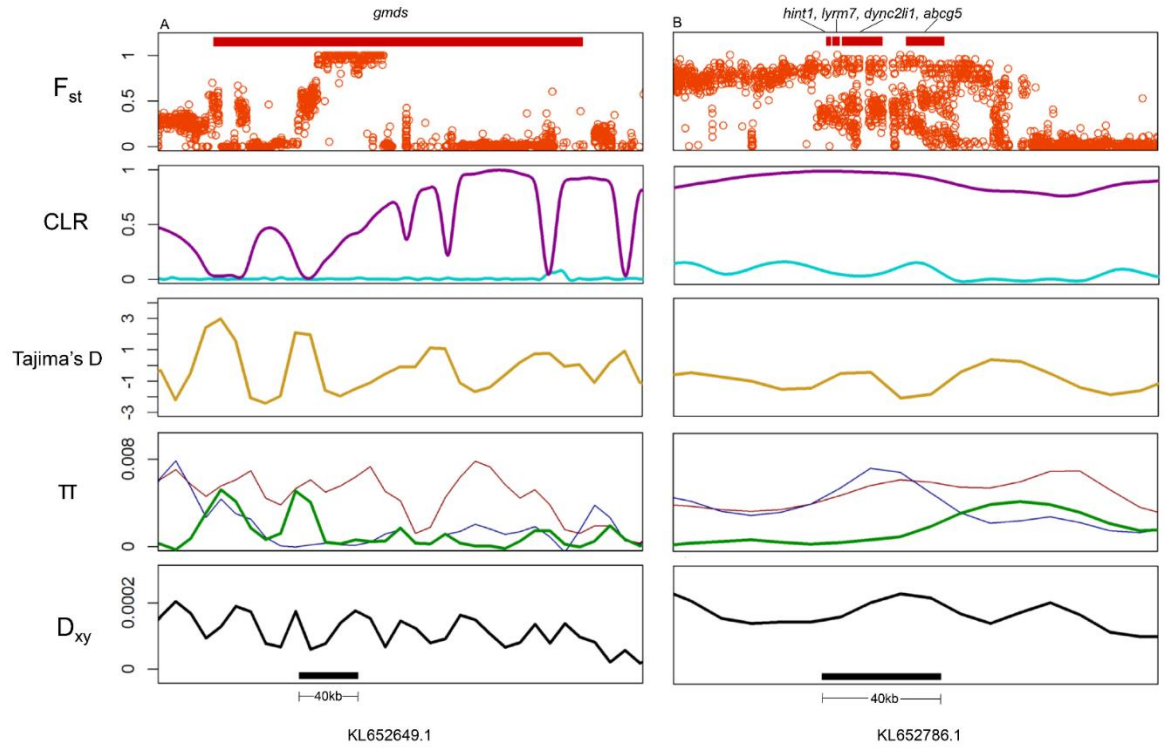

Figure 1.6. Candidate regions associated with large and small jaw size.

Row 1 shows individual SNP F_{st} values between *C. variegatus*/*C. brontotheroides*. Row 2 shows composite likelihood ratios estimated by SweeD using an analytical site frequency spectrum assuming a population bottleneck (magenta) and a frequency spectrum calculated without demographic assumptions (cyan) for *C. brontotheroides*. Row 3 shows Tajima's D (dark yellow) for *C. brontotheroides*. Row 4 shows within-population diversity (π) (red: *C. variegatus*, green: *C. brontotheroides*, blue: *C. desquamator*). Row 5 shows between-population divergence (D_{xy}, black) for *C. variegatus*/*C. brontotheroides*. Black bars in row 4 show windows containing fixed SNPs that showed significant association with jaw size in both PLINK and GEMMA association mapping analyses. Red bars in row 1 show exonic and intronic gene regions within windows.

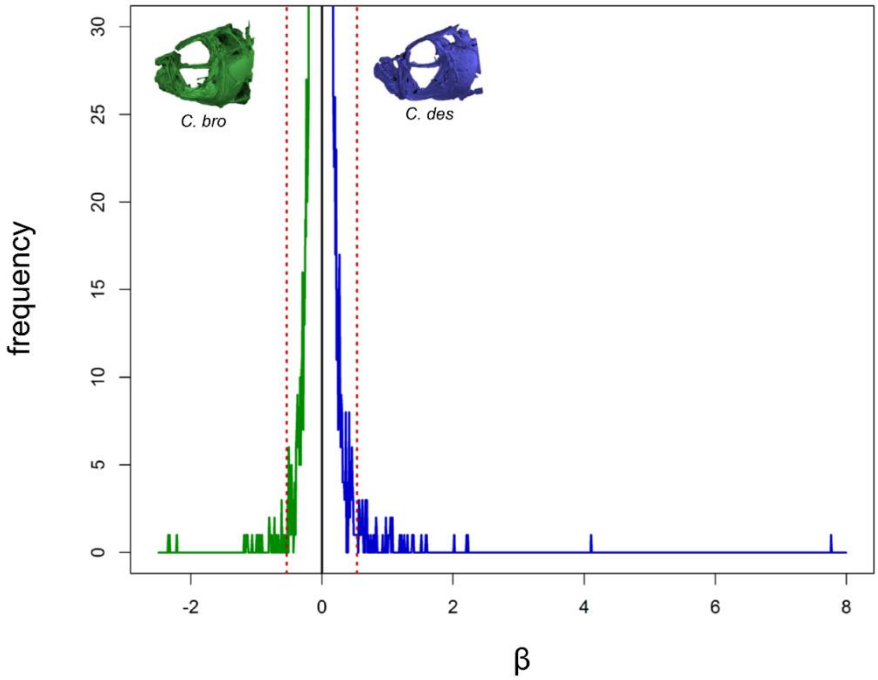

Figure 1.7. More Large-Effect Regions Control Large Jaw Phenotypes.

Distribution of effect size posterior parameters (β) estimated using GEMMA for 20kb regions with a posterior inclusion probability (PIP) greater than zero. We report median β and PIP taken across 10 independent MCMC runs. Association mapping analysis shows twice as many outlier regions with large effects ($\beta > 99^{th}$ percentile (dotted red line)) on increasing jaw size ($n = 56$) compared to large-effects on decreasing jaw size ($n = 27$).

CHAPTER 2: PARALLEL EVOLUTION OF GENE EXPRESSION BETWEEN TROPHIC SPECIALISTS DESPITE DIVERGENT GENOTYPES AND MORPHOLOGIES[1]

Introduction

Abundant research on the genetic basis of adaptive traits has revealed an overarching pattern in nature – when species are faced with similar selective pressures, they often respond with the same adaptive solutions (Conte et al. 2012). For example, parallel changes in gene expression underlying convergent adaptive traits is a well-documented evolutionary phenomenon, with examples from experimental evolution studies imposing uniform selection pressures on replicate populations (Cooper et al. 2003; Riehle et al. 2003), studies in natural systems between closely related taxa (Reid et al. 2016; Derome and Bernatchez 2006; Chan et al. 2010; Nagai et al. 2011; Reed et al. 2011; Manousaki et al. 2013; Zhao et al. 2015), and distantly related taxa (Shapiro et al. 2006; Miller et al. 2007a; Shen et al. 2012). This work has shown that parallelism at the level of gene expression is common in many cases of phenotypic convergence, particularly when divergence time between species is short (Losos 2011; Conte et al. 2012).

However, few studies have investigated the extent of parallel changes in gene expression contributing to species divergence, largely because most expression studies focus on only two

[1] This chapter previously appeared as an article in *Evolution Letters*. The original citation is as follows: McGirr, J. A., and C. H. Martin. 2018. Parallel evolution of gene expression between trophic specialists despite divergent genotypes and morphologies. Evol. Lett. 2:62–75.

species and are either concerned with divergent expression giving rise to divergent phenotypes (Poelstra et al. 2014; Uebbing et al. 2016; Davidson and Balakrishnan 2016) or parallel expression of specific loci (Shapiro et al. 2006; Miller et al. 2007a; Quin et al. 2010) (but see Enard et al. 2002; Ahi et al. 2014). Furthermore, while many genetic and demographic factors are thought to influence the probability of parallel evolution (Rosenblum et al. 2014.; Conte et al. 2012), there are no theoretical expectations for the amount of parallel genetic variation contributing to parallel changes in gene expression during ecological speciation (Schluter et al. 2004; Pavey et al. 2010).

Here we ask whether both parallel and divergent changes in expression underlie novel phenotypes by measuring transcriptomic and genomic divergence between three sympatric species of *Cyprinodon* pupfishes endemic to hypersaline lakes on San Salvador Island, Bahamas. This recent radiation consists of a dietary generalist species (*C. variegatus*) and two novel specialists: a 'snail-eater' (*C. brontotheroides*) and a 'scale-eater' (*C. desquamator*). Scale-eaters have large jaws and elongated bodies, whereas snail-eaters have short, thick jaws and a protruding nasal region that may function in crushing hard-shelled mollusks. These specialists are more morphologically diverged from one another than either is from their sympatric generalist sister species, and occupy higher trophic levels than the generalist species (Martin and Wainwright 2013a; Martin 2016a, Martin et al. 2017, Hernandez et al. 2017). Scale-eating and snail-eating rapidly evolved within saline lakes on San Salvador Island, Bahamas. These lakes filled within the past 10,000 years after the last glacial maximum (Mylroie and Hagey 1995; Turner et al. 2008), suggesting that speciation occurred rapidly. Scale-eaters and snail-eaters have only been found on San Salvador, and likely diverged from a generalist common ancestor based on phylogenetic analyses of outgroup species (Holtmeier 2001; Turner et al. 2008; Martin

and Wainwright 2011; Martin and Wainwright 2016b). Pupfish populations on many neighboring Bahamian islands and throughout the Caribbean are dietary generalists (Martin and Wainwright 2011; Martin 2016a) and these specialist niches appear unique within atherinomorph and cyprinodontiform fishes (Martin and Wainwright 2011). Indeed, the scale-eating pupfish is separated by 168 million years from other scale-eating fishes (Martin and Wainwright 2013b).

We performed total mRNA sequencing to examine gene expression in lab-reared individuals of all three San Salvador pupfish species from different lake populations at two developmental stages. We also searched 42 whole genomes for SNPs unique to each specialist and determined whether fixed variants near differentially expressed genes showed signs of hard selective sweeps (Pavlidis et al. 2013). We found significant parallelism at the level of gene expression in specialists, but did not find evidence for shared fixed variants underlying parallel changes in expression. We tested whether this counterintuitive result of parallel changes in expression between divergent trophic specialists may be due to 1) decreased pleiotropic constraint for genes showing parallelism or that 2) specialists experience parallel selective environments and adapted to higher trophic levels using similar genetic pathways. Finally, we identified genes differentially expressed between generalists and scale-eaters that contain fixed genetic variants within regions that were previously associated with jaw size variation and showed signs of experiencing a recent hard selective sweep (McGirr and Martin 2017). These regions with fixed variants represent promising *cis*-regulatory elements underlying divergent jaw size – the most rapidly diversifying trait in the San Salvador pupfish radiation (Martin and Wainwright 2013c).

Methods

Study system and sample collection

Individuals were caught from hypersaline lakes on San Salvador Island, Bahamas using a hand net or seine net in 2011, 2013, and 2015. Whole genome resequencing was performed for wild-caught individuals from a total of nine isolated lakes on San Salvador (Great Lake, Stout's Lake, Oyster Lake, Little Lake, Crescent Pond, Moon Rock, Mermaid's Pond, Osprey Lake, and Pigeon Creek). 14 scale-eaters were sampled from six populations; 11 snail-eaters were sampled from four populations; and 13 generalists were sampled from eight populations on San Salvador. Outgroup samples included one *C. laciniatus* from Lake Cunningham, New Providence Island, Bahamas, one *C. bondi* from Etang Saumautre lake in the Dominican Republic, one *C. diabolis* from Devil's Hole in California, and captive-bred individuals of *C. simus* and *C. maya* from Laguna Chicancanab, Quintana Roo, Mexico. Sampling is further described in (McGirr and Martin 2017; Richards and Martin 2017). Fish were euthanized in an overdose of buffered MS-222 (Finquel, Inc.) following approved protocols from the University of California, Davis Institutional Animal Care and Use Committee (#17455) and University of California, Berkeley Animal Care and Use Committee (AUP-2015-01-7053) and stored in 95-100% ethanol.

RNA sequencing and alignment

Juvenile pupfish were derived from either F_0 wild caught or F_1 lab raised individuals that were held in a common laboratory environment and fed identical diets (Table B1.1; 25-27^o C, 10-15 ppt salinity, pH 8.3). We collected larvae at two developmental stages: 8-10 and 17-20 days post-fertilization (dpf). The variation in sampling time is due to uncertainty in precise spawning times since eggs were fertilized naturally within breeding tanks and collected on the same day or subsequent day following egg laying. However, we sampled hatched larvae in a

haphazard manner over multiple spawning intervals and it is unlikely that sampling time varied consistently by species. Larvae were euthanized in an overdose of buffered MS-222, and stored in RNA later (Ambion, Inc.) at 4º C for one day, followed by long-term storage at -20º C for up to one year. We extracted whole-larvae RNA using RNeasy kits (Qiagen) from 15 larvae (8-10 dpf) (Three F_2 generalists and F_2 snail-eaters from Crescent Pond, three F_1 generalists and F_2 snail-eaters from Little Lake, and three F_1 scale-eaters from Little Lake; Table B1.1). We also dissected 14 larvae (17-20 dpf) to isolate tissues from the anterior craniofacial region containing the dentary, angular articular, maxilla, premaxilla, palatine, and associated craniofacial connective tissues using fine-tipped tweezers washed with RNase AWAY (Three F_2 generalists and F_2 snail-eaters from Crescent Pond, three F_1 generalists and F_2 snail-eaters from Little Lake, and two F_1 scale-eaters from Little Lake; Table B1.1).

Libraries were prepared using the KAPA stranded mRNA-seq kit (KAPA Biosystems 2016) at the High Throughput Genomic Sequencing Facility at UNC Chapel Hill. Stranded sequencing on one lane of Illumina 150PE Hiseq4000 resulted in 677 million raw reads. We filtered raw reads using Trim Galore (v. 4.4, Babraham Bioinformatics) to remove Illumina adaptors and low-quality reads (mean Phred score < 20). We mapped these reads to the *Cyprinodon* reference genome using the RNA-seq aligner STAR (v. 2.5 (Dobin et al. 2013a)). We used the featureCounts function of the Rsubread package (Liao et al. 2014) requiring paired-end and reverse stranded options to generate read counts across previously annotated features. We assessed mapping and count quality using MultiQC (Ewels et al. 2016).

Differential expression analyses

We quantified differences in gene expression between all three species at two developmental stages. Our raw counts determined by featureCounts were normalized with

DESeq2 (v. 3.5 (Love et al. 2014)) which uses counts to calculate a geometric mean for each gene across samples, divides individual gene counts by this mean, and uses the median of these ratios as a size factor for each sample. Next, we used DESeq2 to perform pairwise tests pooling species across lakes to identify differentially expressed genes between generalists *vs.* snail-eaters and generalists *vs.* scale-eaters at 8-10 dpf and 17-20 dpf (Table B1.1). Genes with fewer than two read counts were discarded from all analyses (n = 1,570), along with genes showing low normalized counts at a threshold determined by DESeq2 (Love et al. 2014). Wald tests determined significant differences in expression between species by comparing normalized posterior log fold change estimates and correcting for multiple testing using the Benjamini–Hochberg procedure with a false discovery rate of 0.05 (Benjamini and Hochberg 1995).

We performed two analyses to test whether specialist species exhibited nonrandom patterns of parallel changes in expression relative to their generalist sister species. We used a Fisher's exact test to determine whether there was a significant overlap between genes that showed differential expression in both comparisons (i.e. genes that were differentially expressed between generalists *vs.* snail-eaters and generalists *vs.* scale-eaters). A gene that was differentially expressed in both comparisons could either show the same direction of expression in specialists or opposite directions of expression. We performed 10,000 permutations sampling from a binomial distribution to estimate the expected number of genes showing shared and opposite directions of expression. Under this null model of gene expression evolution, a strong positive deviation from 50% of genes showing a shared direction of expression in specialists would indicate significant parallel changes in expression.

Our scale-eater sample sizes were lower than generalist and snail-eater samples for each pairwise comparison (see above). We used a down sampling procedure to test whether sample

size affected patterns of parallel changes in expression. We analyzed differential expression for generalists *vs.* snail eaters and generalists *vs.* scale-eaters in 1,000 permutations where generalists and snail-eaters were randomly sampled from our full dataset to match scale-eater sample sizes (n = 3 for 8-10 dpf comparisons; n = 2 for 17-20 dpf). Next, we identified the number of genes differentially expressed between generalists *vs.* snail eaters and generalists *vs.* scale-eaters in each permutation and calculated the proportion of those genes that showed the same direction of expression in specialists relative to generalists. A strong positive deviation from 50% of genes showing a shared direction of expression across permutations would indicate that parallel evolution of expression in specialists is robust to variation in sample size.

Gene ontology enrichment analyses

We performed gene ontology (GO) enrichment analyses for differentially expressed genes using GO Consortium resources available at geneontology.org (Ashburner et al. 2000; GO Consortium 2017). We used BlastP (v. 2.6 (Camacho et al. 2009)) to identify zebrafish protein orthologs with high similarity (*E*-value < 1) to NCBI protein accessions for genes that we identified as differentially expressed between *Cyprinodon* species. Orthology was established using one-way best hits, where a protein sequence in *Cyprinodon* was the best match to a sequence in zebrafish, and reciprocal best blast hits, where a sequence in *Cyprinodon* was the best match to a sequence in zebrafish and vice versa. While reciprocal best hits robustly predict orthology with high precision, it is highly conservative and fails to detect many true orthologs in duplication rich clades such as teleosts (Altenhoff and Dessimoz 2009; Salichos and Rokas 2011; Dalquen and Dessimoz 2013). Thus, we performed GO enrichment analyses using orthologs defined as one-way best hits, and compare these results to enrichment analyses using more conservative orthologs defined as reciprocal best hits.

Genes were either differentially expressed between generalists and snail-eaters, generalists and scale-eaters, or in both comparisons. Thus, we performed two GO enrichment analyses for: 1) genes that were differentially expressed in both companions, and 2) genes differentially expressed in one comparison. We grouped enriched GO categories into similar representative terms using the REVIGO clustering algorithm (Tomislav 2011). REVIGO groups semantically similar terms to reduce the size and redundancy of lists from GO enrichment analyses, where grouping is guided by *P*-values corrected for multiple comparisons (Tomislav 2011). When similar terms show similar enrichment, they are assigned to a single representative term. We measured differences in the proportion of representative terms describing metabolic and developmental processes between genes showing parallel and divergent changes in expression between specialists.

Measuring pleiotropy for differentially expressed genes

The probability of parallel evolution of gene expression may be higher for genes that are less constrained by negative pleiotropy (Cooper and Lenski 2000; Manceau et al. 2010; Rosenblum et al. 2014). High gene pleiotropy is correlated with participation in more protein-protein interactions (PPIs), which in turn effects multiple biological processes (He and Zhang 2006; Safari-alighiarloo et al. 2014). Genes that act across multiple developmental stages are also more pleiotropic (Stern and Orgogozo 2008). We used one-way best hits zebrafish orthologs to estimate pleiotropy for differentially expressed genes based on their number of associated GO biological processes, PPIs, and developmental stages when they are known to be expressed (Papakostas et al. 2014). We again used GO Consortium resources (Ashburner et al. 2000; GO Consortium 2017) to determine the number of biological processes associated with each gene. We examined biological process annotations only for genes from ZFIN (zfin.org) with

experimental evidence (GO evidence code EXP). The String protein database (v. 10; (Szklarczyk et al. 2015)) calculates a combined score measuring confidence in protein interactions by considering known interactions (experimentally determined and from manually curated databases) and predicted interactions. We used the String database to quantify PPIs for protein products of differentially expressed genes, focusing only on interactions with experimental evidence (i.e. non-zero experimental evidence scores). Next, we determined the number of developmental stages where a gene is known to be expressed using the Bgee expression call database for zebrafish (v. 14.0 (Bastian et al. 2008)). We considered eight developmental stages from larval day five to juvenile day 89 from the Zebrafish Stage Ontology (ZFS) that were deemed 'gold quality,' meaning there was no contradicting call of absence of expression for the same gene, in the same developmental stage (Bastian et al. 2008).

We tested whether genes showing parallel changes in expression between specialists showed lower levels of pleiotropy than genes showing divergent changes in expression by fitting a generalized linear model on count data for pleiotropy estimates (negative binomial family; *glm.nb* function in the R library "MASS"). We did not measure pleiotropy for genes expressed at 17-20 dpf due to the low number of zebrafish orthologs matched for genes with parallel expression in craniofacial tissues (11 out of 23).

Genomic variant discovery and population genetic analyses

SNP variants were called using previously outlined methods (McGirr and Martin 2017; Richards and Martin 2017). Briefly, 42 individual DNA samples extracted from muscle tissue were fragmented, barcoded with Illumina indices, and quality checked using a Fragment Analyzer (Advanced Analytical Technologies, Inc.). Sequencing on four lanes of Illumina 150PE Hiseq4000 resulted in 2.8 billion raw reads that were mapped from 42 individuals to the

Cyprinodon reference genome (NCBI, *C. variegatus* Annotation Release 100, total sequence length = 1,035,184,475; number of scaffold = 9,259, scaffold N50, = 835,301; contig N50 = 20,803; (Lencer et al. 2017)). We followed Genome Analysis Toolkit (v 3.5) best practices and hard filter criteria to call and refine our SNP variant dataset (QD < 2.0; FS < 60; MQRankSum < -12.5; ReadPosRankSum < -8 (DePristo et al. 2011)). We filtered our final SNP dataset to include individuals with a genotyping rate above 90% (no individuals were excluded by this filter) and SNPs with minor allele frequencies higher than 5%, resulting in 16 million variants with a mean sequencing coverage of 7× per individual (range: 5.2–9.3×).

We identified SNPs that were fixed in each specialist species. We calculated genome wide F_{st} using VCFtools' 'weir-fst-pop' function for two different population comparisons involving samples collected from San Salvador: generalists (n = 13) *vs.* snail-eaters (n = 11) and generalists (n = 13) *vs.* scale-eaters (n = 9). Differences in sample sizes made our analyses biased to detect more fixed variation between generalists vs. scale-eaters (n = 13 vs. 9) than between generalists vs. snail-eaters (n = 13 vs. 11). We also performed 1,000 permutations calculating genome wide F_{st} between randomly subsampled groups in order to identify non-randomly differentiated genomic regions between species. We calculated the 99th percentile estimates of F_{st} across all SNPs between randomly sampled generalists and snail-eaters (n = 13 vs. n = 11) and between randomly sampled generalists and scale-eaters (n = 13 vs. n = 9). We took the 99th percentile of these distributions to set a threshold defining significantly divergent outliers (Fig. B2.7).

Our SNP dataset included 14 scale-eaters, however, we split our scale-eater population into two groups (large-jawed scale-eaters, n = 9 and small-jawed scale-eaters, n = 5) based on previous evidence that these two populations are genetically distinct (McGirr and Martin 2017;

Richards and Martin 2017). This allowed us to identify SNPs unique to large-jawed scale-eaters (i.e. *C. desquamator* (Martin and Wainwright 2013a)), which were the only type of scale-eater we sampled for RNA-seq. We identified which of these SNPs resided in gene regions (either exonic, intronic, or within 10kb of the first or last exon) for genes showing differential expression. We determined whether these regions showed signatures of hard selective sweeps using SweeD ((Pavlidis et al. 2013); methods previously described in (McGirr and Martin 2017)). Briefly, SweeD sections scaffolds into 1,000 windows of equal size and calculates a composite likelihood ratio (CLR) using a null model where the site frequency spectrum of each window does not differ from that of the entire scaffold. We previously estimated ancestral effective population sizes of San Salvador pupfishes using MSMC (Schiffels and Durbin 2014; McGirr and Martin 2017) and used these estimates to correct the expected neutral site frequency spectrum for the inferred recent population bottleneck in Caribbean pupfishes using SweeD. Windows with fixed SNPs that showed CLRs above the 95th percentile across their respective scaffolds (>10,000bp) under the assumptions of a recent population bottleneck were interpreted as regions that recently experienced a hard sweep.

Results

Differential expression between generalists and each specialist

Total mRNA sequencing across all 29 samples resulted in 677 million raw reads, which was reduced to 674 million reads after quality control and filtering. 81.2% of these reads successfully aligned to the reference genome and 75.5% of aligned reads mapped to annotated features with an average read depth of 309× per sample. The number of reads mapping to

annotated features was comparable across generalists, snail-eaters, and scale-eaters (ANOVA; 8-10 dpf $P = 0.47$; 17-20 dpf $P = 0.33$; Fig. B2.1).

Snail-eaters and scale-eaters occupy novel niches among over 2,000 species of atherinomorph fishes (Martin and Wainwright 2011), and these trophic specialist species likely evolved from a generalist common ancestor within the past 10,000 years (Mylroie, J.E, Hagey 1995; Turner et al. 2008). We analyzed transcriptomic changes underlying rapid trophic divergence by comparing specialist species gene expression against their sympatric generalist sister species. We used DESeq2 to identify genes that were differentially expressed between generalists *vs.* snail-eaters and generalists *vs.* scale-eaters at 8-10 days post-fertilization (whole body tissue) and 17-20 dpf (craniofacial tissue).We measured expression across 22,183 genes with greater than two read counts out of 24,383 total genes annotated for the *Cyprinodon variegatus* assembly (NCBI, *C. variegatus* Annotation Release 100, (Lencer et al. 2017)).

At 8-10 dpf, we found 1,014 genes differentially expressed between generalists *vs.* snail-eaters and 5,982 genes differentially expressed between generalists *vs.* scale-eaters (Fig. 2.1A and C; Fig. 2.2A) (Benjamini and Hochberg adjusted $P \leq 0.05$). 818 genes were differentially expressed in both comparisons, which is a significantly larger amount of overlap than expected by chance (Fisher's exact test, $P < 1.0 \times 10^{-16}$). Remarkably, 815 of these 818 genes showed the same direction of expression in specialists relative to generalists (Fig. 2.2B). Specifically, 441 differentially expressed genes showed lower expression in both specialist species compared to generalists, while 374 showed higher expression in specialists. Only three genes showed opposite directions of expression (Fig. 2.2B). Two genes showed higher expression in snail-eaters and lower expression in scale-eaters while one gene showed higher expression in scale-eaters (Table B1.2). This is significantly more parallel change in expression between specialists than would be

expected under a null model of gene expression evolution, where a gene has an equal chance of showing a shared or opposite direction of expression in specialists relative to generalists (10,000 permutations, $P < 1.0 \times 10^{-4}$; Fig. B2.2). Parallel evolution of expression in specialists was consistent at both the gene and isoform level (Fig. B2.2, B2.3).

Craniofacial morphology is the most rapidly diversifying trait in the San Salvador radiation (Martin and Wainwright 2013c). In order to detect genes expressed during jaw development, we compared expression within craniofacial tissue at the 17-20 dpf stage. We found a similar pattern of parallel changes in gene expression at this developmental stage (Fig. B2.4). 120 genes were differentially expressed between generalists *vs.* snail-eaters and 1,903 genes differentially expressed between generalists *vs.* scale-eaters (Fig. 2.1B and D). Again, we saw a significant amount of overlap between comparisons with 23 genes differentially expressed in both comparisons (Fisher's exact test, $P < 1.0 \times 10^{-5}$). 22 of these 23 genes showed the same direction of expression in specialists relative to generalists (Fig. B2.4). Specifically, 10 genes showed lower expression in both specialist species compared to generalists, while 12 showed higher expression in specialists (Fig. B2.4). Only one gene (*mybpc2*) showed opposite directions of expression, with higher expression in snail-eaters and lower expression in scale-eaters (Table B1.2).

Our sample sizes for scale-eater species were lower for comparisons at 8-10 dpf (n = 3) and 17-20 dpf (n = 2) relative to snail-eaters and generalists (n = 6 for both stages). We measured differential expression for generalists *vs.* snail eaters and generalists *vs.* scale-eaters in 1,000 permutations where we randomly down-sampled generalists and snail-eaters from our full dataset to match scale-eater sample sizes. Figure 2.2C shows the proportion of genes differentially expressed at 8-10 dpf in both comparisons that showed the same direction of

expression in specialists relative to generalists across 1,000 permutations. The total number of differentially expressed genes in each permutation was variable (Fig. B2.5 A and C, median number of genes common to both comparisons = 61). Despite this variability, we found that the parallel evolution of expression in specialists was robust to smaller sample size, with greater than 90% of genes showing parallel evolution of expression in 90% of permutations (Fig. 2.2C). However, at 17-20 dpf parallel changes in expression were not as consistent across permutations (Fig. B2.4 C and B2.5 F).

Genes showing parallel changes in expression are enriched for metabolic processes

We performed GO enrichment analyses with one-way blast hit zebrafish orthologs for genes showing parallel changes in expression between specialists (n = 620) and genes showing divergent expression patterns in snail-eaters (n = 102) and scale-eaters (n = 3,349). We restricted these analyses to genes expressed at 8-10 dpf because the number of genes showing parallel expression in specialists at 17-20 dpf (n = 23) was low and did not show enrichment for any biological process.

We grouped enriched GO categories into similar representative terms using the REVIGO clustering algorithm (Tomislav 2011). Genes showing parallel changes in expression between specialists were enriched for metabolic processes (20% of representative terms; Fig. 2.3A; Table B1.3). In contrast, genes with divergent expression patterns in specialists were enriched for cranial skeletal development and pigment biosynthesis (7% and 3% of terms, respectively) while only 11% of enriched categories described metabolic processes (Table B1.4).

We also performed GO enrichment analyses using orthologs that were established using a more conservative reciprocal best hit approach, where a sequence in *Cyprinodon* was the best match to a sequence in zebrafish and vice versa. As expected, we identified fewer reciprocal

best hits than one-way hits (615 genes showing parallel changes in expression between specialists, 95 genes showing divergent expression unique to snail-eaters, and 2,150 genes showing divergent expression unique to scale-eaters). Encouragingly, we still found that genes showing parallel changes in expression were enriched for metabolic processes (26% of representative terms), whereas genes showing divergent expression showed less enrichment for metabolic processes (20% of representative terms). However, we did not see any enrichment for cranial development or pigment biosynthesis for genes showing divergent expression using reciprocal best hit orthologs.

We tested whether genes showing parallel changes in expression were less constrained by pleiotropy than genes showing divergent expression between specialists. We estimated pleiotropy for orthologs of differentially expressed genes based on their number of protein-protein interactions (PPIs), associated GO biological processes, and developmental stages when they are known to be expressed. However, we did not find any difference in pleiotropy for genes showing parallel changes in expression compared to genes showing divergent expression using any of these three metrics (GLM; biological processes: $P = 0.67$; PPIs: $P = 0.09$; developmental stages: $P = 0.89$) (Fig. B2.6).

Genetic variation underlying parallel changes in expression

We identified 79 SNPs fixed between generalists *vs.* snail-eaters and 1,543 SNPs fixed between generalists *vs.* scale-eaters (also see our previous study on genome-wide association mapping jaw length in these species). None of these fixed variants were shared between specialists. Next, we determined which of these fixed SNPs fell within gene regions (either exonic, intronic, or within 10kb of the first or last exon; Table 2.1). 26 SNPs fixed in snail-eaters

overlapped with 17 gene regions, whereas 1,276 SNPs fixed in scale-eaters overlapped with 245 gene regions.

Next, we identified fixed variants near genes that showed differential expression. We found 319 SNPs fixed in scale-eaters within 71 gene regions that showed differential expression between generalists and scale-eaters at 8-10 dpf and 118 SNPs within 26 gene regions differentially expressed between generalists and scale-eaters at 17-20 dpf. We suspect that some of these fixed variants are within *cis*-regulatory elements responsible for species-specific expression patterns that ultimately give rise to phenotypic differences in scale-eaters. Conversely, we only identified a single SNP fixed in snail-eaters within a gene (*tmprss2*) that was differentially expressed between generalists and snail-eaters at 8-10 dpf. We did not find any fixed variants near genes differentially expressed between generalists and snail-eaters at 17-20 dpf, possibly suggesting that fixed variants regulate expression divergence at an earlier developmental stage.

Since we did not find any variants that were fixed between snail-eaters and generalists that were also fixed between scale-eaters and generalists, we searched for shared variation at a lower threshold of genetic divergence. We calculated the 99th percentile outlier F_{st} estimates between randomly subsampled groups of each species across 1,000 permutations to create two null distributions of genome-wide divergence. We took the 99th percentile of these distributions as an estimate of significantly high divergence (Fst > 0.36 for generalists *vs.* snail-eaters; Fst > 0.42 for generalists *vs.* scale-eaters; Fig. B2.7). We found 4,410 SNPs above this lower threshold of divergence near 134 genes showing parallel changes in expression between specialists at 8-10 dpf. The most differentiated SNPs near genes showing parallel changes in expression show F_{st} < 0.8 between generalists *vs.* snail-eaters and generalists *vs.* scale-eaters. Overall, these results

suggest it is unlikely that the parallel evolution of gene expression in specialists is controlled by shared variation that is fixed or nearly fixed in specialist populations.

The genetic basis of extreme craniofacial divergence

We previously described 30 candidate gene regions containing variants fixed between trophic specialist species associated with variation in jaw length. These candidates also showed signatures of a recent hard selective sweep (McGirr and Martin 2017). Encouragingly, we found ten of these genes differentially expressed between generalists and scale-eaters (eight at 8-10 dpf and two at 17-20 dpf) and one between generalists and snail-eaters (8-10 dpf; Table B1.5).

We searched for signatures of hard selective sweeps across the 84 gene regions containing fixed variation in specialists (Table 2.1). Interestingly, 80% of these gene regions showed signs of a hard sweep (estimated by SweeD; CLR > 95^{th} percentile across their respective scaffolds; Table B1.6). All of these gene regions contained SNPs that were either fixed between generalists *vs.* snail-eaters or generalists *vs.* scale-eaters and showed differential expression at 8-10 dpf, 17-20 dpf, or both. Finally, we compared this list of genes experiencing selection to those annotated for cranial skeletal system development (GO:1904888) and muscle organ development (GO:0007517). While this search was limited to zebrafish orthologs identified as one-way best hits, we were able to identify three genes containing fixed variation in scale-eaters that likely influence craniofacial divergence through cis-acting regulatory mechanisms (*loxl3b* (annotated for cranial effects)*; fbxo32* and *klhl40a* (annotated for muscle effects)).

Discussion

We combined RNA sequencing with genome-wide divergence scans to study the molecular evolution of two trophic specialist species that rapidly diverged from a generalist common ancestor within the last 10,000 years. We examined how gene expression and SNP variation influence snail-eater and scale-eater niche adaptations using comparisons between each specialist and their generalist sister species. We found a significant amount of parallelism at the level of gene expression yet no parallelism at the level of fixed genetic variation within specialists. Specifically, 80% of genes that were differentially expressed between snail-eaters and generalists were up or downregulated in the same direction when comparing expression between scale-eaters and generalists (Fig. 2.2A). We explored two possible explanations for this pattern: 1) reduced pleiotropic constraints made these genes likely targets for parallelism or 2) convergent processes drove parallel gene expression evolution in this highly divergent pair of specialist species due to shared adaptations to a higher trophic level.

Pleiotropic constraints do not explain parallel changes in gene expression

Genes that effect one or a few traits are less constrained than genes with many phenotypic effects, perhaps making them simpler shared targets for expression divergence during adaptive evolution between independently evolving lineages. Indeed, theory predicts that the probability of parallel evolution of gene expression should be higher for genes with minimal pleiotropic effects (Manceau et al. 2010, Rosenblum et al. 2014). We predicted that genes showing parallel changes in expression between specialists would show lower degrees of pleiotropy than divergently expressed genes. We estimated three measures of gene pleiotropy (number of associated GO biological processes, protein-protein interactions (PPIs), and developmental stages when they are known to be expressed) and found no significant difference

in any measure for genes showing parallel versus divergent changes in expression patterns (Fig. B2.6). This finding is consistent with some empirical evidence and theoretical models of gene expression evolution that found pleiotropy constrains the variability of gene expression within species, but does not hinder divergence between species (Tulchinsky et al. 2014; Uebbing et al. 2016).

Parallel changes in gene expression underlie convergent metabolic adaptations to a higher trophic level in each specialist

While the specialists are more morphologically diverged from one another than either is from the generalist species, particularly in their craniofacial phenotype and male reproductive coloration (Martin and Wainwright 2013a; Martin et al. 2017) (Fig. 2.3B and C), dietary isotope analyses show that they occupy a higher trophic level than generalists (Martin 2016b). Fish scales and mollusks contribute to more nitrogen-rich diets in specialists compared to generalist species that primarily consume algae and detritus (Martin 2016b). Perhaps the same metabolic processes required for this type of diet are adaptive at higher trophic levels for both scale-eaters and snail-eaters, which might explain patterns of parallel changes in expression. Thus, we predicted that genes showing parallel changes in expression would affect metabolic processes that may be similar between specialists, whereas genes showing divergent expression between specialists would affect morphological development.

GO enrichment analyses using one-way best-hit zebrafish orthologs support both hypotheses. We found that 20% of GO terms enriched for genes showing parallel changes in expression described metabolic processes, and zero described cranial skeletal development or pigment biosynthesis (Fig. 2.3A; Table B2.3). In contrast, 10% of terms showing enrichment in the divergently expressed gene set described developmental processes (cranial skeletal

development and pigment biosynthesis) and only 11% described metabolic processes (Fig 3A, Table B1.4). GO enrichment analyses using more conservatively defined reciprocal best hit orthologs confirmed that genes showing parallel changes in expression were highly enriched for metabolic processes (26% of representative terms). These results suggest that the parallel evolution of expression in specialists confers adaptation to a higher trophic level. Snail-eating and scale-eating may present similar metabolic requirements relative to the lower trophic level of algivorous generalists. This is consistent with the high macroalgae content of generalist diets relative to both specialist species (Martin and Wainwright 2013c) and the shorter intestinal lengths observed in both specialists relative to the generalist (CHM and JAM personal observation).

Enrichment analyses using one-way best hit orthologs indicate that genes showing divergent expression in specialists are responsible for shaping divergent cranial and pigmentation phenotypes between species (Fig. 2.3), but we did not find enrichment for these processes using reciprocal best hit orthologs. This may be because up to 60% of orthologous relationships are missed by the reciprocal best-hit criterion in lineages with genome duplications, including teleosts (Dalquen and Dessimoz 2013). Finally, both approaches we used to establish orthology indicated that genes showing divergent expression in specialists were moderately enriched for metabolic processes (Fig. 2.3A; Table B2.4). While parallel changes in expression may broadly influence adaptation to a higher trophic level, these divergently expressed metabolic genes likely play a role in dietary specialization unique to each species.

Parallel changes in gene expression despite unshared genetic variation

We find significant parallel evolution of gene expression across genes that are annotated for effects on metabolism, yet shared expression patterns do not seem to be driven by the same

fixed variants. This is surprising in this young radiation given that the probability of shared genetic variation underlying phenotypic convergence increases with decreasing divergence time (Schluter et al. 2004; Conte et al. 2012; Martin and Orgogozo 2013). Although 80% of differentially expressed gene regions containing fixed SNPs show signs of experiencing a selective sweep, and almost none of these variants were in exons, it is still possible that fixed alleles do not regulate parallel changes in expression for metabolic genes. Indeed, we found 4,410 SNPs that showed significant differentiation between generalists *vs.* snail-eaters and generalists *vs.* scale-eaters near 134 genes showing parallel changes in expression. These shared variants all showed $F_{st} < 0.8$, suggesting that parallel expression is not controlled by shared variation that is fixed or nearly fixed in specialist populations. However, our results do not rule out a role for fixed variation influencing the parallel evolution of expression through long-range chromosome interactions or during earlier critical developmental stages, such as neural crest cell migration at approximately 48 hpf.

It is surprising that we do not find fixed variation shared between specialists near genes showing parallel changes in expression given that the probability of parallel genetic variation underlying phenotypic convergence is higher when divergence time between species is short (Schluter et al. 2004; Conte et al. 2012; Martin and Orgogozo 2013). Many studies that show parallel adaptation at the gene level describe convergence within pigmentation and skeletal development pathways (Miller et al. 2007b; Reed et al. 2011; Conte et al. 2012; Kronforst et al. 2012). Perhaps the architecture of metabolic adaptation is more flexible, having more mutational targets or employing more late-acting developmental regulatory networks that are less constrained than early-acting networks (Kalinka et al. 2010; Garfield et al. 2013; Martin and Orgogozo 2013; Reddiex et al. 2013; Ferna et al. 2014; Comeault et al. 2017). Our findings

highlight the importance of understanding convergence across different biological levels of organization.

Candidate genes influencing trophic adaptations

We found many genes affecting metabolism that were differentially expressed in the same direction in specialists relative to their generalist sister species. While the metabolism ontology includes a broad class of proteins with a variety of biological functions, we find many with distinct effects on dietary metabolism. For example, the gene *asl* (argininosuccinate lyase) is important for nitrogen excretion. Variants of *asl* are associated with argininosuccinic aciduria and citrullinemia, conditions involving an accumulation of ammonia in the blood (Saheki et al. 1987; Hu et al. 2015). This gene, along with some of 274 other genes we found annotated for nitrogen metabolism, may show parallel changes in expression between specialists as an adaptation to nitrogen-rich diets (Martin 2016b).

We also identified candidate genes influencing cranial divergence that were differentially expressed between scale-eaters and generalists, contain SNPs fixed in scale-eaters, and showed signs of a hard selective sweep. *loxl3b* is highly expressed in scale-eaters at 8-10 dpf and annotated for cranial effects (Table B1.6). The protein encoded by this gene (lysyl oxidase 3b) controls the formation of crosslinks in collagens, and is vital to cartilage maturation during zebrafish craniofacial development (Van Boxtel et al. 2011). Mutations in *loxl3b* are associated with Stickler Syndrome, which is characterized by cranial anomalies and cleft palate (Alzahrani et al. 2015). *fbxo32* and *klhl40a* are both expressed at lower levels in scale-eaters at 8-10 dpf relative to generalists and may influence skeletal muscle divergence between species (Table B1.6). High expression of *fbxo32* is associated with muscle atrophy, while mutations in *klhl40a* cause nemaline myopathy (muscle weakness) (Ravenscroft et al. 2013; Mei et al. 2015). Variants

fixed in scale-eaters near these genes, along with fixed variation near differentially expressed genes previously associated with large jaw size (McGirr and Martin 2017; Table B1.5) represent strong *cis*-acting regulatory candidates potentially influencing scale-eater cranial traits.

Caveats to gene expression analyses and the robustness of parallel evolution

We compared the transcriptomes of derived trophic specialists to a contemporary generalist sister species to identify gene expression divergence important for the evolution of trophic traits. However, the generalist transcriptome represents an approximation of the putative ancestral state, and has evolved independently over the past 10,000 years (Holtmeier 2001; Turner et al. 2008; Martin and Wainwright 2011; Martin 2016a). We chose to sample RNA at 8-10 dpf and 17-20 dpf to identify transcriptional variation that influences larval development, however, some activation of parallel gene networks is likely specified at pre-hatching developmental stages (Garfield et al. 2013; Ferna et al. 2014). It is also possible that we did not have the power to identify subtle differences in expression for genes that showed high divergence between specialists and generalists. Detecting differential expression of transcripts is notoriously difficult when read counts are low and variance within treatment groups is high (Conesa et al. 2016; Lin et al. 2016). We were able to detect differential expression for genes with a mean normalized count as low as 1.6 (median = 150) and $\log_2$ fold change as low as 0.2 (median = 1.11). Furthermore, our scale-eater sample sizes (8-10 dpf n = 3; 17-20 dpf n = 2) were lower than that of generalists and snail-eaters (n = 6 at both stages; Table B1.1). Nonetheless, down sampling analyses suggest that patterns of parallel expression are robust to smaller sample sizes for 8-10 dpf tissue (Fig. 2.2C), but less so for 17-20 dpf tissue (Fig. B2.4 C).

Finally, our novel results are consistent with a recently published independent analysis of gene expression in San Salvador pupfishes that identified many of the same genes we found divergently expressed between specialists (Lencer et al. 2017). We examined this dataset using the same significance thresholds for differentially expressed genes as described in Lencer *et al.* for mRNA extracted from all three species at 8 dpf and 15 dpf ($P < 0.1$ and $|\text{Log}_2 \text{ fold change}| > 0.2$). We found that 40% of genes divergently expressed between specialists in this dataset were divergently expressed in our own dataset. Importantly, Lencer *et al.* only examined cranial tissues at both of these developmental stages and they did not choose to examine parallel evolution of expression. We also searched for evidence of parallel change in expression for mRNA extracted from all three species at 8 dpf in the Lencer et al. dataset. 28.8% of genes that were differentially expressed between snail-eaters and generalists were up or downregulated in the same direction between scale-eaters and generalists. This is a lower proportion of parallel change in expression than we identified (Fig. 2.2), but this is most likely because Lencer *et al.* only sampled RNA from cranial tissues at 8 dpf, unlike our sampling of whole larvae. Thus, the majority of parallel changes in expression between specialists likely occurs in non-cranial tissues, consistent with our shared metabolic hypothesis.

Conclusion

Here we find significant parallel evolution of gene expression between two highly divergent specialist species relative to their generalist sister species. While there are many cases of parallel changes in expression underlying parallel specialization, to our knowledge, this represents the first case of parallel expression underlying divergent specialization. Numerous studies have shown that shared genetic variation underlying phenotypic convergence is more likely when divergence times between species are short (Schluter et al. 2004; Conte et al. 2012;

Martin and Orgogozo 2013). Scale-eating and snail-eating species have evolved rapidly within the last 10,000 years, yet we do not find the same variants fixed in both species underlying parallel changes in expression. We show that parallel evolution of expression likely reveals convergent adaptation to a higher trophic level in each specialist, despite their highly divergent resource use and morphology.

Table 2.1. Genomic distribution of fixed variants.

The first five columns show the total number of fixed SNPs in each species comparison and how many fall within exons, introns, 10kb of the first or last exon of a gene, and outside of 10kb from the first or last exon of a gene. Final two columns show the number of genes with fixed SNPs within the gene and/or within 10kb of the first or last exon. The last column shows the number of differentially expressed (DE) genes near fixed SNPs that includes DE genes from 8-10 dpf and 17-20 dpf comparisons.

comparison	fixed SNPs	exonic	intronic	within 10kb of coding region	>10kb away from coding region	number of genes near fixed SNPs	number of DE genes near fixed SNPs
generalist vs. snail-eater	79	0	16	10	53	17	1
generalist vs. scale-eater	1,543	140	512	624	267	245	83

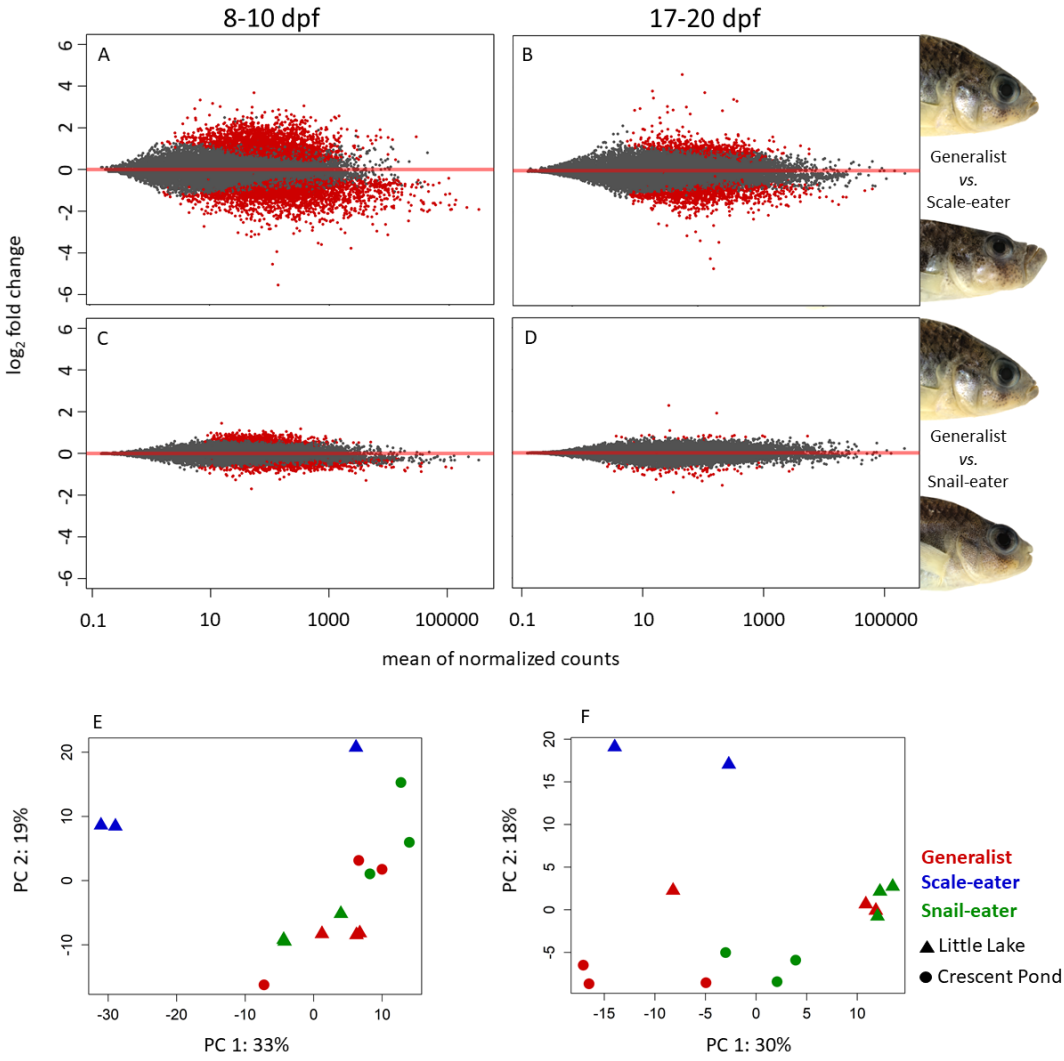

Figure 2.1. Differential gene expression between generalists and trophic specialists.

Red points represent genes that are differentially expressed in 8-10 dpf whole-larvae tissue (A, C) and 17-20 dpf craniofacial tissue (B, C) between generalists *vs.* scale-eaters (A, B) and generalist *vs.* snail-eaters (C, D). Bottom panels show the top two principal components accounting for a combined 52% (8-10 dpf; E) and 48% (17-20 dpf; F) of the total variation between samples across 413 million reads mapped to annotated features. Triangles represent samples from Little Lake and circles represent samples from Crescent Pond on San Salvador Island.

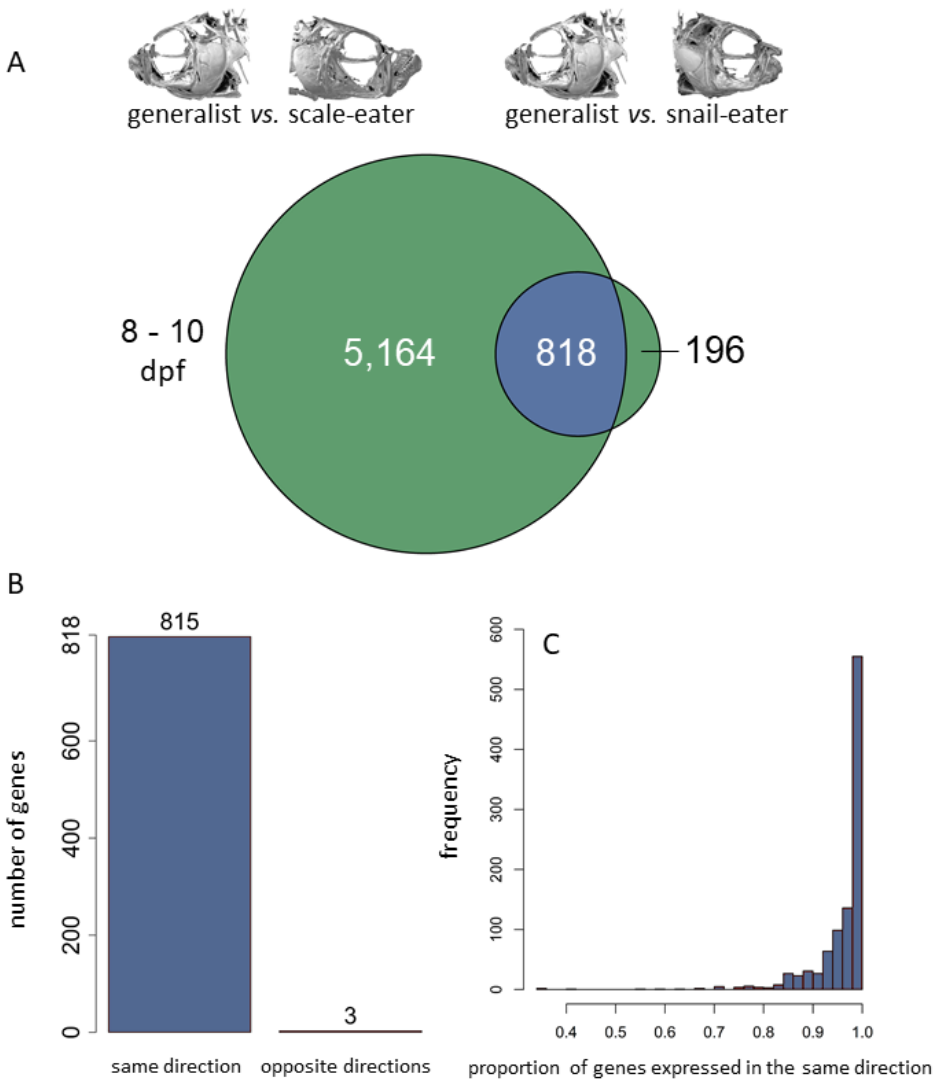

Figure 2.2. Parallel evolution of gene expression between specialists despite divergent trophic adaptation.

A) Circles illustrate genes differentially expressed in 8-10 dpf whole-larvae tissue for generalists *vs.* scale-eaters (left) and generalists *vs.* snail-eaters (right). Genes showing differential expression in both comparisons are shown in blue, and those showing divergent expression patterns unique to each specialist are green. Significantly more genes show differential expression in both comparisons than expected by chance (Fisher's exact test, $P < 1.0 \times 10^{-16}$). B) Significantly more genes show the same direction of expression in specialists relative to generalists than expected by chance (10,000 permutations; $P < 1.0 \times 10^{-4}$; Fig. B2.2). C) Distribution of the proportion of genes differentially expressed in the same direction between specialists relative to generalists after 1,000 down sampling permutations show that parallel expression is robust to variation in sample size (median number of genes common to both comparisons = 61).

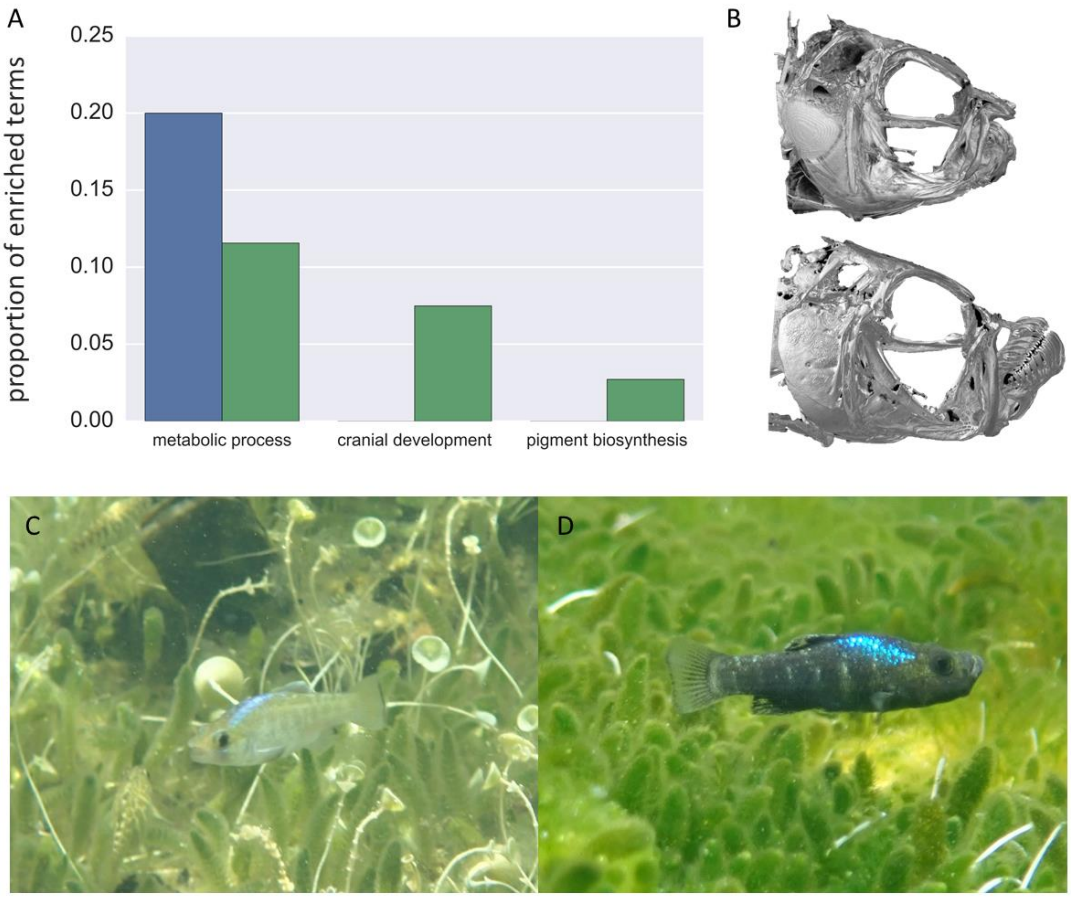

Figure 2.3. Parallel gene expression underlies metabolic adaptations while divergent expression underlies trophic morphology.

A) Genes showing parallel changes in expression between specialists (blue) and genes showing divergent expression (green) are contrastingly enriched for terms describing metabolic processes (parallel: 20% of enriched terms; divergent: 11% of terms). Genes showing divergent expression are enriched for cranial skeleton development (7% of terms) and pigment biosynthesis (3% of terms). B) µCT scans show drastic craniofacial divergence between snail-eaters (top) and scale-eaters (bottom) (modified from Hernandez et al. 2017). Bottom panels show male breeding coloration characteristic of light snail-eaters (C) and dark scale-eaters (D).

CHAPTER 3: HYBRID GENE MISREGULATION IN MULTIPLE DEVELOPING TISSUES WITHIN A RECENT ADAPTIVE RADIATION OF CYPRINODON PUPFISHES[1]

Introduction

Changes in gene expression are an important source of variation in adaptive morphological traits (Carroll 2008; Wolf et al. 2010; Indjeian et al. 2016). As genetic variation accumulates in regulatory and coding sequences, stabilizing selection on gene expression results in coevolution such that molecular functions are largely maintained (Coolon et al. 2014; Hodgins-Davis et al. 2015). Crossing divergent species to form F1 hybrids can break up such coadapted variation, resulting in genetic incompatibilities within developing tissues that give rise to adaptive traits (Michalak and Noor 2004; Landry et al. 2007; Mack and Nachman 2017a). Genetic incompatibilities that reduce hybrid fitness can drive reproductive isolation either intrinsically – causing sterility or increased embryonic mortality – or extrinsically whereby incompatibilities reduce hybrid performance in a particular environment (Schluter 2000; Coyne 2004).

Of particular importance to the process of speciation are genetic incompatibilities caused by hybrid misregulation – transgressive expression levels in hybrids that are higher or lower than both parental species (Michalak and Noor 2004; Ranz et al. 2004; Haerty and Singh 2006;

[1] This chapter previously appeared as an article in *PLoS One.* The original citation is as follows: McGirr, J. A., and C. H. Martin. 2019. Hybrid gene misregulation in multiple developing tissues within a recent adaptive radiation of *Cyprinodon* pupfishes. PLoS One. 14(7): e0218899.

Rockman and Kruglyak 2006; Malone and Michalak 2008; Renaut et al. 2009). This pattern of expression causes Dobzhansky-Muller incompatibilities (DMIs) if incompatible alleles in hybrids cause misregulation that results in reduced hybrid fitness and thus increased postzygotic reproductive isolation (Presgraves 2003; Coyne 2004; Sweigart et al. 2006; Ortíz-Barrientos et al. 2007; Malone and Michalak 2008; Renaut et al. 2009; Davidson and Balakrishnan 2016). Laboratory studies searching for genes that cause DMIs often identify genes causing sterility or embryonic lethality in hybrids. This approach ignores the fitness consequences of misregulation occurring at later developmental stages within diverse tissue types, thus underestimating the actual number of genetic incompatibilities distinguishing species (Fang et al. 2012; Schumer et al. 2014). Combining findings from these studies with analyses of hybrid misregulation in tissues that give rise to adaptive morphological traits can reveal a broader view of incompatibilities that arise during speciation.

Studies of gene expression in hybrids can also implicate regulatory mechanisms underlying expression divergence between parental species, which is important for understanding how expression levels are inherited and how they shape adaptive traits (Wittkopp et al. 2004; McManus et al. 2010; Mack et al. 2016). Research on hybrid gene expression thus far has shown mixed results regarding patterns of inheritance (Signor and Nuzhdin 2018). Some studies found evidence for ubiquitous transgressive expression inherited in F1 hybrids (i.e. over- or under-dominance) (Ranz et al. 2004; Rockman and Kruglyak 2006; Roberge et al. 2008), while others found predominately additive patterns (Hughes et al. 2006; Rottscheidt and Harr 2007; Davidson and Balakrishnan 2016). Mechanisms of gene expression divergence in F1 hybrids are characterized as interactions between *cis*-regulatory elements and *trans*-regulatory factors. *Cis* elements are often non-coding regions of DNA proximal to genes that are bound by

trans-acting proteins and RNAs to regulate mRNA abundance. It is possible to identify mechanisms of gene expression divergence between parental species by bringing *cis* elements from both parents together in the same *trans* environment in F1 hybrids and quantifying allele specific expression (ASE) of parental alleles at heterozygous sites (Cowles et al. 2002; Wittkopp et al. 2004). *Cis* and *trans* regulatory variants can compensate for one another if stabilizing selection favors an optimal level of gene expression. Hybrid misregulation is expected when different compensatory variants have accumulated in diverging lineages (Denver et al. 2005; Landry et al. 2005; Bedford and Hartl 2009; Goncalves et al. 2012).

Here we investigated F1 hybrids from crosses between two closely related species of *Cyprinodon* pupfishes to understand regulatory mechanisms that led to the evolution of novel craniofacial adaptations in this group (Fig 1A). *Cyprinodon brontotheroides* – hereafter referred to as the molluscivore – is a trophic specialist species endemic to San Salvador Island, Bahamas that has adapted to eat hard shelled prey including mollusks and ostracods (Martin and Wainwright 2013a,c). This species likely diverged from a generalist common ancestor within the past 10,000 years to occupy this novel niche (Mylroie, J.E, Hagey 1995; Holtmeier 2001; Turner et al. 2008; Martin and Wainwright 2011; Martin 2016a). Adapting to this niche involved extreme morphological divergence in craniofacial traits compared to its sympatric generalist sister species *Cyprinodon variegatus* (Martin and Wainwright 2013c; Lencer et al. 2016). This species consumes mainly algae and detritus and is hereafter referred to as the 'generalist.' Almost all other Caribbean pupfish species are generalists, with the exception of a novel scale-eating pupfish that is also a member of the San Salvador pupfish radiation (Martin and Wainwright 2011, 2013c) and a second sympatric radiation of trophic specialists in Laguna Chichancanab, Mexico (Strecker 2006; Humphries et al. 2019). Molluscivores exhibit a novel

skeletal protrusion on the anteriodorsal head of the maxilla not found in generalist populations that may be used to stabilize prey items held within its oral jaws, which are shorter and more robust relative to generalist species (Fig 1A). This jaw morphology provides higher mechanical advantage for crushing mollusks and other hard-shelled prey (Wainwright and Richard 1995; Martin and Wainwright 2011).

Molluscivores and generalists readily hybridize in the laboratory to produce fertile F1 offspring with approximately intermediate craniofacial morphologies and no obvious sex ratio distortion (Martin and Wainwright 2011, 2013b; Martin and Feinstein 2014). These species remain largely reproductively isolated in sympatry across multiple lake populations (genome-wide average F_{st} = 0.08; (Martin and Feinstein 2014; West and Kodric-Brown 2015; Mcgirr and Martin 2016)). Therefore, unlike most studies of hybrid misregulation, we are not solely concerned with identifying gene expression patterns underlying hybrid sterility or lethality. Rather, we also aim to characterize misregulation in developing tissues that gives rise to novel craniofacial phenotypes within a young species pair with ongoing gene flow. We dissected craniofacial tissue from 17-20 day old F1 hybrids and extracted total mRNA to quantify gene expression levels. We also extracted whole-larvae mRNA from 8 day old generalists, molluscivores, and their F1 hybrids. We found genes misregulated in hybrids at both stages. Finally, we quantified allele specific expression (ASE) across exome-wide heterozygous sites to uncover mechanisms of regulatory divergence and found evidence for putative compensatory variation influencing patterns of hybrid misregulation.

Materials and Methods

Study system and sample collection

Our methods for raising larvae and extracting RNA were identical to previously outlined methods (McGirr and Martin 2018). We collected fishes for breeding from three hypersaline lakes on San Salvador Island, Bahamas (Little Lake, Osprey Lake, and Crescent Pond) using a hand net or seine net between 2011 and 2017. These fishes were reared at 25–27°C, 10–15 ppt salinity, pH 8.3, and fed a mix of commercial pellet foods and frozen foods. All lab bred larvae were raised exclusively on newly hatched brine shrimp after hatching and before euthanasia. Individuals were euthanized in an overdose of buffered MS-222 and stored in RNA later (Ambion, Inc.) at -20°C for up to 18 months. We used RNeasy Mini Kits (Qiagen catalog #74104) to extract RNA from all samples.

We previously generated 24 transcriptomes belonging to generalists and molluscivores collected at two early developmental stages: 8-10 days post fertilization (dpf) and 17-20 dpf (McGirr and Martin 2018). RNA was extracted from whole-larvae tissue at 8-10 dpf. We dissected all 17-20 dpf samples to extract RNA from anterior craniofacial tissues containing the dentary, angular, articular, maxilla, premaxilla, palatine, and associated craniofacial connective tissues (Fig. C2.1). Dissections were performed using fine-tipped tweezers washed with RNase AWAY (Molecular BioProducts). These 24 samples were generated by breeding populations of lab-raised fishes that resulted from either one or two generations of full-sib breeding between wild caught fishes from Little Lake and Crescent Pond on San Salvador Island, Bahamas (Table 3.1). There was variation in sampling time because eggs were fertilized naturally within breeding tanks and collected on the same day or subsequent day following egg laying. We collected larvae

in a haphazard manner over multiple spawning intervals and it is unlikely that sampling time varied consistently by species.

Here we analyze an additional 19 transcriptomes from generalists, molluscivores, and their F1 hybrids (Table 3.1). First, we crossed a generalist female with a molluscivore male to generate four F1 hybrids that were collected at 17-20 dpf and extracted RNA from dissected craniofacial tissues. A lab-reared female generalist was used to generate hybrids that was derived from wild caught generalists from Little Lake following one generation of full-sib mating. A lab-reared male molluscivore was used to generate hybrids that was derived from wild caught molluscivores from Little Lake following two generations of full-sib mating.

We performed separate crosses to collect larvae at exactly 8 dpf (190-194 hours after fertilization rather than 8-10 days). We crossed a generalist female with a molluscivore male to generate three F1 hybrids for whole-larvae RNA extractions. The parents of these hybrids were wild-caught from Osprey Lake. Finally, we extracted whole-larvae RNA from six generalists and six molluscivores collected at 8 dpf. These samples were generated from wild-caught individuals from Osprey Lake and Crescent Pond. In total, we analyzed transcriptomes from 43 individuals that involved four separate rounds of sequencing (Table 3.1 and C1.1).

RNA sequencing and alignment

The previously reported 24 transcriptomes were sequenced at the High Throughput Genomic Sequencing Facility at UNC Chapel Hill in April 2017 (McGirr and Martin 2018). The 24 libraries were prepared at the facility using the KAPA stranded mRNAseq kit (KAPA Biosystems 2016) followed by sequencing on one lane of Illumina 150 paired-end Hiseq4000 (Table 3.1 and C1.2).

19 additional transcriptomes were sequenced at The Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley. All 19 libraries were prepared at the facility using the Illumina stranded Truseq RNA kit (Illumina RS-122-2001) and all sequencing was performed on Illumina 150 paired-end Hiseq4000. Four libraries for RNA extracted from 17-20 dpf hybrid craniofacial tissues were pooled on a single lane and sequenced in June 2017. 15 libraries for whole-larvae RNA samples collected at exactly 8 dpf were pooled across one and three lanes and sequenced in May (n = 9) and July (n = 6) 2018, respectively (Table 3.1 and C1.1).

We filtered all raw reads using Trim Galore (v. 4.4, Babraham Bioinformatics) to remove Illumina adaptors and low-quality reads (mean Phred score < 20) and mapped filtered reads to the scaffolds of the *Cyprinodon* reference genome (NCBI, *C. variegatus* annotation release 100, total sequence length = 1,035,184,475; number of scaffolds = 9259, scaffold N50 = 835,301; contig N50 = 20,803; (Lencer et al. 2017)) using the RNAseq aligner STAR with default parameters (v. 2.5 (Dobin et al. 2013a)). We used the featureCounts function of the Rsubread package (Liao et al. 2014) requiring paired-end and reverse stranded options to generate read counts across 24,952 previously annotated features (Lencer et al. 2017) with an average coverage depth of 136 reads (Table C1.2 and C1.3). We assessed mapping and count quality using MultiQC (Ewels et al. 2016). We previously showed that there was no difference between generalists and molluscivores in the proportion of reads that map to annotated features of the *Cyprinodon* reference genome (McGirr and Martin 2018). Similarly, here we found no difference in the proportion of reads mapping to features between generalists, molluscivores, and hybrids (Fig. C2.2; ANOVA, $P = 0.6$), but we did find that fewer reads mapped to features in 17-20 dpf samples than 8 dpf samples (ANOVA, $P = 2.38 \times 10^{-10}$).

Since we analyzed RNA from 43 individuals that were sequenced across four different dates and their libraries were prepared with either KAPA or TruSeq stranded mRNAseq kits, we tested whether a significant amount of between-sample variance in read counts was explained by sequencing date or library preparation kit. We fit linear models (using the lm() function in R) to determine whether normalized counts across individuals were influenced by 1) the number of duplicate reads, 2) the uniformity of coverage across a transcript, or 3) the depth of coverage across GC-rich transcripts. All of these measures could have been influenced by different library preparation methods (Alberti et al. 2014; Biosystems 2014; Van Dijk et al. 2014). RseQC identified duplicates as paired reads that mapped to the exact same locations. These can be natural duplicates (and informative for differential expression comparisons) or result from differences in fragmenting and PCR amplification methods used by different library preparation kits (Parekh et al. 2016). We quantified the number of duplicate reads and the median percent GC content of mapped reads for each sample using RSeQC (Wang et al. 2012). We also used RSeQC to estimate transcript integrity numbers (TINs) which is a measure of potential *in vitro* RNA degradation within a sample. TIN is calculated by analyzing the uniformity of coverage across transcripts (Wang et al. 2012, 2016). We performed ANOVA to determine whether the proportion of duplicate reads, GC content of reads, TINs, the number of normalized read counts, number of raw read counts, or number of raw fastq reads differed between samples grouped by library preparation method and by sequencing date.

Differential expression analyses and hybrid inheritance of expression patterns

We performed differential expression analyses with DESeq2 (v. 3.5 (Love et al. 2014)). This program fits negative binomial generalized linear models for each gene across samples to test the null hypothesis that the fold change in gene expression between two groups is zero.

DESeq2 uses an empirical Bayes shrinkage method to determine gene dispersion parameters, which models within-group variability in gene expression, and logarithmic fold changes in gene expression. DESeq2 normalizes raw read counts by calculating a geometric mean of counts for each gene across samples, dividing individual gene counts by this mean, and using the median of these ratios as a size factor for each sample. These sample-specific size factors account for differences in library size and sequencing depth between samples. Gene features showing less than 10 normalized counts in every sample in each comparison were discarded from analyses. These filtering criteria would exclude genes that are only expressed in one group. However, this conservative threshold discarded those genes that showed low coverage across all samples, which would have low power to detect differences in expression between groups. Differential expression between groups was determined with Wald tests by comparing normalized posterior log fold change estimates and correcting for multiple testing using the Benjamini–Hochberg procedure with a false discovery rate of 0.05 (Society 2017). We also used DESeq2 to perform clustering and principal component analyses (Fig. C2.3).

We conducted pairwise comparisons to identify genes differentially expressed between hybrids vs. parental species, hybrids vs. generalists, hybrids vs. molluscivores, and generalists vs. molluscivores. “Parental species” refers to generalists and molluscivores derived from the same populations as the parents of the hybrid samples. We did not sequence any of the parents crossed to generate hybrids. We defined genes as misregulated in hybrids if they were significantly differentially expressed between hybrids and the parental species samples. First, we compared whole-larvae gene expression between samples collected at 8 dpf (six generalists, six molluscivores, and three hybrids). All of the 8 dpf samples were sequenced at the Vincent J. Coates Genomic Sequencing Laboratory, University of California Berkeley (VJCGSL UCB) and

their libraries were all prepared using the TruSeq stranded mRNAseq kit. Second, we compared craniofacial tissue gene expression between samples collected at 17-20 dpf (six generalists, six molluscivores, and four hybrids). The generalist and molluscivore samples were sequenced at the High-Throughout Sequencing Facility, University of North Carolina Chapel Hill (HTSF UNC) and their libraries were prepared using the KAPA stranded mRNA-seq kit, while the hybrids were sequenced at the VJCGSL UCB and their libraries were prepared using the TruSeq kit. In order to understand how sequencing at different facilities and using different library prep methods affected the proportion of genes misregulated between hybrids and parental species at 17-20 dpf, we performed a third set of comparisons between hybrids collected at 8 dpf (sequenced at VJCGSL UCB with TruSeq) and generalists and molluscivores from a previous study collected at 8-10 dpf (sequenced at HTSF UNC with KAPA; (McGirr and Martin 2018)). We measured how many genes were differentially expressed between 8 dpf hybrids vs. 8-10 dpf parental species than there were differentially expressed between 8 dpf hybrids vs. 8 dpf parental species. This allowed us to estimate an upper-limit on the proportion of genes falsely identified as differentially expressed between 17-20 dpf hybrids and 17-20 dpf parental species due to samples being sequenced at different facilities with different library preparation kits.

To determine whether genes showed additive, dominant, or transgressive patterns of inheritance, we quantified differences in gene expression between hybrids vs. parental species and compared them to genes differentially expressed between generalists vs. molluscivores (Fig. 3.2). Hybrid inheritance was considered additive if hybrid gene expression was intermediate between generalists and molluscivores with significant differential expression between generalists and molluscivores, respectively. Inheritance was dominant if hybrid expression was significantly different from one parent species but not the other. Genes showing misregulation in

hybrids showed transgressive inheritance, meaning hybrid gene expression was significantly higher (overdominant) or lower (underdominant) than both parental species.

Gene ontology enrichment analyses

The *Cyprinodon* reference genome is annotated for genomic features (NCBI, *C. variegatus* Annotation Release 100, (Lencer et al. 2017)), and many annotated genes share the same name as their zebrafish orthologs. Of the 26,522 protein coding genes annotated for the *Danio rerio* GRCz11 genome annotation release 106 and the 23,373 protein coding genes annotated for the *Cyprinodon* reference genome, 7,222 genes share the same name. We performed gene ontology (GO) enrichment analyses for genes differentially expressed between species and misregulated in hybrids that shared the same name as zebrafish orthologs using GO Consortium resources available at geneontology.org (Gene Ontology Consortium 2000). We searched for enrichment across biological process ontologies curated for zebrafish.

Allele specific expression and mechanisms of regulatory divergence

We followed the best practices guide recommended by the Genome Analysis Toolkit (v. 3.5 (DePristo et al. 2011)) in order to call and refine SNP variants within coding gene regions using the Haplotype Caller function. We called SNPs across all filtered reads mapped to annotated features for 17-20 dpf samples and 8 dpf samples using conservative hard-filtering parameters (DePristo et al. 2011): Phred-scaled variant confidence divided by the depth of nonreference samples > 2.0, Phred-scaled *P*-value using Fisher's exact test to detect strand bias > 60, Mann–Whitney rank-sum test for mapping qualities (z > 12.5), Mann–Whitney rank-sum test for distance from the end of a read for those with the alternate allele (z > 8.0). We used the VariantsToTable function (with genotypeFilterExpression "isHet == 1") to output heterozygous variants for each individual. We counted the number of reads covering

heterozygous sites using the ASEReadCounter (with -U ALLOW_N_CIGAR_READS -minDepth 20 --minMappingQuality 10 --minBaseQuality 20 -drf DuplicateRead). In total we identified 15,429 heterozygous sites across all 32 individuals with sequencing coverage ≥ 20× that fell within 3,974 genes used for differential expression analyses. At the 8 dpf stage, we found 2,909 of the 3,974 genes that contained heterozygous sites common to all samples. At the 17-20 dpf stage, we found 2,403 genes containing heterozygous sites common to all samples.

We assigned each heterozygous allele as the reference allele, alternate allele, or second alternate allele and matched each allele to its corresponding read depth. This allowed us to identify allele specific expression (ASE) by measuring expression variation between the two sites. We only measured ASE at sites that were heterozygous in all samples in each stage in order to account for differences in nucleotide diversity within populations (Mcgirr and Martin 2016). We used a binomial test in R (binom.test) to determine if a heterozygous site showed significantly biased expression of one allele over another ($P < 0.05$;(McManus et al. 2010; Mack et al. 2016)). We measured ASE across 2,909 genes that contained heterozygous sites common to all 8 dpf samples and 2,403 genes that contained heterozygous sites common to all 17-20 dpf samples. A gene was considered to show ASE in hybrids if a heterozygous SNP within that gene showed consistent biased expression in all hybrid samples (17-20 dpf n = 4; 8 dpf n = 3) and did not show ASE in the parental samples (n = 12 for both developmental stages). We also estimated a more conservative measure of ASE at the gene level using MBASED (Mayba et al. 2014), which uses a pseudo-phasing approach that assigns an allele with a larger read count at each SNP to the 'major' haplotype, assuming that ASE is consistent in one direction along the length of the gene. This program calculates ASE using a beta-binomial test comparing the counts of alternate alleles across each gene. For each sample, we performed a 1-sample analysis with unphased gene

counts using default parameters run for 1,000,000 simulations to identify genes showing significant ASE ($P < 0.05$).

A common approach to identify regulatory mechanisms underlying expression divergence is to measure ASE at sites that are heterozygous in hybrids and alternately homozygous in parental species (Wittkopp et al. 2004; Signor and Nuzhdin 2018). However, generalists and molluscivores diverged recently and there are no fixed SNPs within coding regions out of a total of over 12 million variants examined in 42 resequenced genomes (McGirr and Martin 2018). We measured ASE across heterozygous sites in parental populations to exclude genes which already showed ASE in a pure species background and then determined which genes showed ASE unique to hybrids to make inferences about putative compensatory divergence underlying hybrid misregulation. Gene expression controlled by compensatory variation in parental species is often associated with misregulation in their hybrids (Landry et al. 2005, 2007; Bedford and Hartl 2009; Goncalves et al. 2012). Regulatory elements that have opposite effects on the expression level of a particular gene can compensate for one another to produce an optimal level of gene expression favored by stabilizing selection (Denver et al. 2005; Goncalves et al. 2012). Diverging species can evolve alternate compensatory mechanisms while maintaining similar expression levels (True and Haag 2001). Hybrids of such species would have a mismatched combination of regulatory elements that no longer compensate one another, which is expected to result in biased expression of parental alleles (Wittkopp et al. 2004; Landry et al. 2005). Thus, we identified gene expression controlled by putative compensatory regulatory variation if a gene 1) did not show differential expression between generalists and molluscivores, 2) showed significant ASE at one or more heterozygous sites in F1 hybrids, and 3) did not show

ASE at any site in purebred generalists or molluscivores. Finally, we looked for overlap between genes showing compensatory regulation and genes showing misregulation in hybrids.

Results

Differential expression between generalists and molluscivores

We previously found 1,014 genes differentially expressed in whole-larvae tissue between six generalists and six molluscivores collected 8-10 days post fertilization (dpf; (McGirr and Martin 2018)). Here we compared gene expression in whole-larvae tissue collected at exactly 8 dpf (190-194 hours after fertilization rather than 8-10 dpf) between six generalists and six molluscivores. We found 700 out of 17,723 (3.9%) genes differentially expressed between species (Fig 1C). 235 of the 700 genes were annotated as zebrafish orthologs and used for gene ontology enrichment analyses. Encouragingly, the only significantly overrepresented ontology was skeletal system morphogenesis (GO:0048705) which matched 11 differentially expressed genes (Table C1.4).

We previously found 120 genes differentially expressed in craniofacial tissue between species at 17-20 dpf (McGirr and Martin 2018). Here we reexamined gene expression in those same individuals using a more conservative threshold for genes to be included in differential expression analyses (where a gene must show >= 10 normalized counts in every sample included in the comparison). As expected, we found fewer genes differentially expressed using this more conservative threshold (81 out of 13,901 (0.6%); Fig 1E). These 81 genes did not show enrichment for any biological process ontologies.

Hybrid misregulation in whole-larvae tissue

We compared gene expression in whole-larvae tissue collected at 8 dpf from generalist and molluscivore populations (n = 12) with expression in their F1 hybrids (n = 3) and found that 370 out of 17,705 genes (2.1%) were misregulated in hybrids (Fig. 3.1D). Slightly more genes showed underdominant inheritance (209; 1.2%) than overdominant inheritance (154; 0.89%; Fig. 3.3A and C). The magnitude of differential expression was higher for genes showing underdominant inheritance than overdominant inheritance (Fig. C2.4; Wilcoxon rank sum test, $P = 8.5 \times 10^{-5}$). Of the 370 genes showing misregulation, 138 were annotated as zebrafish orthologs used for gene ontology enrichment analyses. The only significantly overrepresented term was cellular lipid metabolic process (GO:0044255).

The majority of genes showed conserved levels of expression with no significant difference between hybrids and parental species (84.9%). In line with other hybrid expression studies (Hughes et al. 2006; Rottscheidt and Harr 2007; Davidson and Balakrishnan 2016), most genes that did not show conserved inheritance showed additive inheritance (399; 2.3%). We found some genes showing evidence for dominance, with 89 (0.51%) showing 'generalist-like' expression patterns and 168 (0.97%) showing 'molluscivore-like' patterns of inheritance (Fig 3A and C).

Hybrid misregulation in craniofacial tissue

We compared gene expression in craniofacial tissue collected at 17-20 dpf from generalist and molluscivore populations (n = 12) with expression in their F1 hybrids (n = 4) and found extensive hybrid misregulation. More than half of genes (6,590 out of 12,769 (51.6%)) were differentially expressed in hybrids compared to parental species expression (Fig 1F). There was an approximately equal number of genes showing overdominant and underdominant

expression in hybrids, with 3,299 (25.83%) genes showing higher expression in hybrids relative to parental species and 3,291 (25.77%) showing lower expression in hybrids (Fig 1F, Fig 3B and D). While there was a similar number of genes showing over- and underdominance, the magnitude of differential expression was higher for genes showing underdominance (Fig. C2.4; Wilcoxon rank sum test, $P < 2.2 \times 10^{-16}$). Of the 6,590 genes showing misregulation, 2,876 were annotated as zebrafish orthologs used for gene ontology enrichment analyses. Misregulated genes were enriched for 210 ontologies, including embryonic cranial skeleton morphogenesis (GO:0048701; Table C1.5 and A3.6).

Hybrid misregulation is influenced by library preparation and sequencing conditions

All of the 8 dpf samples were sequenced at the same facility using the same library preparation kit. However, the 17-20 dpf generalist and molluscivore samples were sequenced at a different facility than the 17-20 dpf hybrid samples and used a different library preparation kit. We took two approaches toward understanding how sequencing at different facilities and using different library kits may have affected the proportion of genes misregulated between hybrids and parental species at 17-20 dpf.

First, we performed another differential expression comparison between whole-larvae hybrids collected at 8 dpf and whole-larvae parental species that we collected for a previous study between 8-10 dpf (McGirr and Martin 2018). The 8 dpf hybrids were sequenced at the same facility with the same library kit as the 17-20 dpf hybrids, while the 8-10 dpf parental species were sequenced at the same facility with the same library kit as the 17-20 dpf parental species. This design mirrored the comparison we used to estimate 17-20 dpf hybrid craniofacial misregulation, but at an earlier developmental stage (Fig. C2.5). Whereas comparisons between 8 dpf hybrids and parental species sequenced under the same conditions revealed 370 genes (2.1%)

misregulated, comparisons between hybrids and parental species sequenced at different sequencing centers with different library preparation kits suggested that 997 (6%) genes were misregulated – a 37% increase (Fig. C2.5). This presents a major caveat to our findings, but does not suggest that all genes showing hybrid misregulation in 17-20 dpf craniofacial tissues are false-positives. Using this estimate of bias to correct for different library preparation methods for our 17-20 dpf samples, we estimate that 19.1% genes were misregulated in hybrid craniofacial tissue rather than the raw estimate of 51.6%.

We also investigated whether a significant amount of between-sample variance in read counts was explained by library preparation method or sequencing date. For each sample we quantified the number of normalized read counts, raw read counts, and raw fastq reads. We also estimated the proportion of duplicate reads out of total mapped reads, the median percent GC content across mapped reads, and the uniformity of coverage across mapped reads (median transcript integrity numbers (TINs)). All of these measures could be influenced by different library preparation methods (Alberti et al. 2014; Biosystems 2014; Van Dijk et al. 2014). However, library preparation method was not associated with differences in the number of normalized read counts or median TINs (Fig. 3.4A and B; Welch two sample t-test, $P > 0.05$). When we grouped samples by sequencing date rather than library preparation method, we found that the 17-20 dpf hybrid craniofacial samples (sequenced 6/17) did not show any difference in median GC content, raw read counts, or raw fastq reads compared to samples sequenced on different dates (Fig S6). However, these samples did show lower proportions of duplicate reads, fewer normalized read counts, and lower TINs compared to samples sequenced on all other dates (Fig. 3.4C-E; ANOVA; $P < 0.01$). TINs quantify the uniformity of coverage across transcripts and are informative as a measure of *in vitro* RNA degradation, which likely suggests that hybrid

craniofacial samples experienced more degradation than other samples prior to sequencing. Indeed, lower TIN was significantly correlated with a lower number of normalized counts across samples (Fig. 3.4F; linear regression; $P = 2.0 \times 10^{-5}$). We found approximately the same number of genes overexpressed in hybrids (25.83%) as there were genes underexpressed (25.77%), suggesting that many genes were overexpressed in hybrids despite potential RNA degradation.

Overall, we found that our estimate of the proportion of genes misregulated in 17-20 dpf hybrid craniofacial tissue (51.6%) was biased due to differences in the number of duplicate reads produced by two different library preparation methods (Fig. 3.4E). We quantified this bias by measuring hybrid misregulation between samples collected at an earlier developmental stage and found that 19.1% of genes were misregulated in 17-20 dpf hybrid craniofacial tissues after correcting for library preparation biases (Fig. C2.5). We found that 17-20 dpf hybrid craniofacial tissues likely experienced more *in vitro* RNA degradation than other samples, but this did not produce a bias toward more genes showing underdominant expression in hybrids (Fig. 3.3B).

Putative compensatory variation underlies misregulation in hybrids

If a gene shows similar gene expression levels between parental species but shows biased allelic expression only in hybrids, it may be regulated by compensatory variation, and such genes are likely to be misregulated in F1 hybrids (Landry et al. 2005; Goncalves et al. 2012). We identified 15,429 heterozygous sites across all 8 dpf and 17-20 dpf individuals with sequencing coverage $\geq 20\times$ that fell within 2,909 (8 dpf) and 2,403 (17-20 dpf) genes used for differential expression analyses. We estimated allele specific expression (ASE) for these genes and paired these data with patterns of differential expression between parental species to identify genes controlled by putative compensatory variation.

We measured ASE across sites within 2,770 genes that showed no difference in expression between generalists and molluscivores at 8 dpf. We found 157 genes (5.4%) that were likely regulated by compensatory mechanisms, which showed ASE only in hybrids and were not differentially expressed between generalists and molluscivores. Of these, nine genes (0.33%) also showed misregulation in hybrids (Fig. 3.5A and C). We also measured ASE across sites within 2,387 genes that showed no difference in expression between generalists and molluscivores at 17-20 dpf. We found 1080 genes (44.81%) that were likely regulated by compensatory mechanisms. In support of this wide-spread compensatory regulation, 581 of these 1080 genes (53.8%) also showed misregulation in hybrids (Fig. 3.5B and D). These 581 genes showed enrichment for protein maturation, mRNA splicing, macromolecule catabolic process, and intracellular catabolic process.

We also found more genes showing compensatory regulation in 17 dpf tissues than 8 dpf tissues using a more conservative method to identify genes showing ASE with MBASED (Mayba et al. 2014). At 8 dpf, 61 genes (2.2%) showed expression patterns consistent with compensatory regulation, and 18 (0.65%) were misregulated in hybrids. At 17 dpf, 95 genes (3.98%) showed expression patterns consistent with compensatory regulation, and 55 (2.30%) were misregulated in hybrids.

We found many more genes showing ASE (using binomial tests) in 17-20 dpf hybrid craniofacial tissue than any other samples (Fig. 3.6A; ANOVA, $P = 2.81 \times 10^{-5}$). Since misregulation is expected in hybrids when gene expression is controlled by compensatory variation between parental species (Landry et al. 2005; Bedford and Hartl 2009; Goncalves et al. 2012), the high number of genes showing putative compensatory regulation and high number of genes showing ASE in hybrids supports the pattern of extensive misregulation in 17-20 dpf

hybrid craniofacial tissue. We likely overestimated the amount of misregulation in this tissue because hybrids were sequenced using a different library preparation kit than parental species (see above). However, ASE was estimated by examining allelic ratios in individual samples.. 17-20 dpf hybrid craniofacial tissue was sequenced at the same facility using the same library preparation kit as all of the 8 dpf samples (Table 3.1 and C1.1), yet we only found a high number of genes showing ASE in the 17-20 dpf hybrids (Fig 6A).

We tested whether this pattern might be due to higher rates of *in vitro* degradation in hybrid samples (reflected by low TINs), which could increase variance in the abundance of reads at heterozygous sites and bias ASE estimates. Lower TIN was correlated with higher ASE (Fig. 3.6D; linear regression; $P = 9.04 \times 10^{-14}$). This correlation persisted when 17-20 dpf hybrid craniofacial samples were excluded from the model (Fig. 3.6E; linear regression; $P = 0.034$), suggesting that rates of mRNA degradation may differ depending on genotypes at heterozygous sites. While this explains some of the elevated ASE in 17-20 dpf hybrid craniofacial samples, the proportion of genes showing ASE was much higher in these samples than predicted by the latter linear model. Even the lowest TIN for a 17-20 dpf hybrid sample (32.68) predicted a much lower range of genes showing ASE (8.2% -14.1%) compared to the observed range (32.8% - 51.6%). Finally, we also estimated ASE again with a higher coverage threshold (>=100 counts supporting each heterozygous allele) to reduce the chances of increased variance affecting binomial tests and still found that hybrid craniofacial samples showed more ASE than other samples (Fig. 3.6B; ANOVA, $P = 3.85 \times 10^{-4}$).

Discussion

Molluscivores show extreme craniofacial divergence relative to their generalist sister species, exhibiting a novel maxillary protrusion and short robust jaws (Fig 1A; (Martin and Wainwright 2013a; Hernandez et al. 2018)). Given the extreme craniofacial divergence observed between molluscivores and their generalist sister-species, we might expect to find genes expressed in hybrids outside the range of either parent species as a result of discordance between alternatively coadapted genes in regulatory networks shaping divergent craniofacial morphologies. However, genetic divergence between generalists and molluscivores is low, with only 79 SNPs fixed between species (genome-wide average F_{st} = 0.08, D_{xy} = 0.00166; (Mcgirr and Martin 2016; McGirr and Martin 2018)). Despite this low genetic divergence and ongoing gene flow between species, we found gene misregulation in F1 hybrids at two developmental stages and tissue types. We also measured allele specific expression (ASE) for genes expressed in hybrids and parental species and found evidence for putative compensatory divergence influencing hybrid misregulation at both developmental stages.

Hybrid misregulation during juvenile development

While many studies on hybrid misregulation search for regulatory divergence in 'speciation genes' associated with sterility and inviability (Malone and Michalak 2008; Coolon et al. 2014; Davidson and Balakrishnan 2016; Mack et al. 2016), our results highlight the importance of considering misregultion over multiple early developmental stages and in the context of adaptive morphological traits. We found evidence of misregulation in whole-larvae hybrid tissues sampled eight days post fertilization (dpf; 2.1% of genes) and in 17-20 dpf hybrid craniofacial tissues (19.1% of genes after correcting for bias due to library preparation method).

There are several reasons why we might expect to find a higher proportion of genes misregulated in 17-20 dpf hybrid craniofacial tissues relative to 8 dpf whole-larvae tissues. The molluscivore shows exceptional rates of morphological diversification, particularly in craniofacial traits (Martin and Wainwright 2011). Perhaps 17-20 dpf is a crucial developmental window when gene networks shaping these traits become extensively misregulated in hybrids. It is just after this stage that the relative length of the premaxilla, maxilla, palatine, and lower jaw tend to increase more for generalists than molluscivores (Lencer et al. 2016). It is also possible that regulatory changes are compounded throughout development and have cascading effects, resulting in higher rates of misregulation in later stages. Finally, some of the increased misregulation in hybrid craniofacial tissue can likely be attributed to our sampling design. We found that hybrid craniofacial samples showed lower TINs and lower normalized counts (Fig. 3.4A and D), suggesting that these samples may have experienced more *in vitro* RNA degradation than other samples (Wang et al. 2016). While it is difficult to predict how much overdominance we would expect in these samples given that misregulation has not been previously studied in isolated craniofacial tissues, we found approximately the same number of genes overexpressed in hybrids (25.83%) as there were genes underexpressed (25.77%), suggesting that many genes were overexpressed in hybrids despite potential RNA degradation.

We found roughly twice the amount of bias-corrected misregulation in hybrid craniofacial tissues compared to a study of misregulation in whole-larvae tissue that measured gene expression in F1 hybrids generated between benthic and limnetic lake whitefish (Renaut et al. 2009; Renaut and Bernatchez 2011). These populations also diverged within the past 10 kya and occupy different habitats within lakes (Whiteley et al. 2010). We also found that genes showing underdominance in hybrids showed a higher magnitude of differential expression

compared to those showing overdominance in 8 dpf and 17-20 dpf tissues (Fig. C2.4), a pattern that has also been observed in lake whitefish (Renaut and Bernatchez 2011) and a generalist/specialist *Drosophila* species pair (McManus et al. 2010).

The consequences of hybrid misregulation

It is unclear whether such extensive gene misregulation in hybrid craniofacial tissues might contribute to intrinsic postzygotic isolation between generalists and molluscivores. F2 hybrids exhibiting intermediate and transgressive craniofacial phenotypes showed reduced survival and growth rates in the wild relative to F2 hybrids resembling parental species (Martin and Wainwright 2013b; Martin 2016a), but short-term experiments measuring F2 hybrid survival in the lab did not find any evidence of reduced survival for hybrids with intermediate phenotypes (Martin and Wainwright 2013b). This was interpreted as evidence that complex fitness landscapes measured in field enclosures on San Salvador with multiple peaks corresponding to the generalist and molluscivore phenotypes were due to competition and foraging ability in the wild (i.e. extrinsic reproductive isolation). However, additional analyses of these data suggest that absolute performance of hybrids may also play a role in their survival. The most transgressive hybrid phenotypes exhibited the lowest fitness, contrary to expectations from negative frequency-dependent disruptive selection (Martin 2016a). It is still possible that intrinsic and extrinsic incompatibilities interact such that gene misregulation weakens performance more in the wild than in the lab. However, note that F1 hybrids used in this study exhibit approximately intermediate trophic morphology relative to parental trophic morphology whereas field experiments used F2 and later generation hybrid intercrosses and backcrosses.

Hybrid misregulation is controlled by putative compensatory divergence

When an optimal level of gene expression is favored by stabilizing selection, compensatory variation can accumulate between species and cause misregulation in hybrids (Landry et al. 2005; Bedford and Hartl 2009). We combined results from differential expression analyses with allele specific expression (ASE) results to identify genes controlled by putative compensatory regulatory divergence between generalists and molluscivores. In 8 dpf whole-larvae tissue, we found 5.4% of genes controlled by compensatory regulation (Fig. 3.5B). The low number of genes controlled by compensatory regulation was reflected by the low number of genes misregulated in 8 dpf hybrids (2.1%). In 17-20 dpf hybrid craniofacial tissues, we found 44.81% of genes controlled by compensatory regulation (Fig. 3.5B). The large number of genes controlled by compensatory regulation is consistent with the extensive misregulation observed in hybrid craniofacial tissue, and the majority of genes showing signs of compensatory regulation were also misregulated in hybrids (53.8%). 17-20 dpf hybrid craniofacial tissue was sequenced at the same facility using the same library preparation kit as the 8 dpf samples, yet we only found a high number of genes showing ASE in the 17-20 dpf hybrids (Fig. 3.6A and B). One caveat to this result is that the high levels of ASE estimated using binomial tests were influenced to some extent by RNA degradation (Fig. 3.6C and D). To our knowledge, this is the first evidence showing a positive correlation between degradation and ASE, and potential mechanisms underlying differential degradation dependent on heterozygous genotype are unclear. The GC content of mRNAs have been shown to positively correlate with decay rate (Romero et al. 2014). Perhaps mRNAs with G and C genotypes are more likely to degrade before their A and T counterparts at heterozygous sites, causing increased ASE in degraded samples. Despite this caveat, linear models showed that degradation did not predict the extremely high levels of ASE

found in 17-20 dpf hybrid tissues consistent with high misregulation (Fig. 3.6), although it is unknown whether ASE should increase linearly with degradation over time.

Conclusion

We found hybrid misregulation in both whole-larvae tissues and craniofacial tissues sampled at early developmental stages. This points to divergent evolution of developmental networks shaping novel traits in the molluscivore. It is unclear whether such misregulation causes intrinsic incompatibilities in hybrids within this recent adaptive radiation. Our results are in line with studies finding widespread compensatory evolution in other systems with greater divergence times between species (Landry et al. 2007; Takahasi et al. 2011; Goncalves et al. 2012; Bell et al. 2013; Mack and Nachman 2017a). Investigating mechanisms regulating gene expression between generalists and molluscivores that result in hybrid misregulation will shed light on whether the variants shaping novel traits may also contribute to reproductive isolation. Examining misregulation across multiple early developmental stages in the context of developing tissues that give rise to adaptive traits can paint a more complete picture of genetic incompatibilities that distinguish species.

Table 3.1. Sampling design for mRNA sequencing.

Parental fishes crossed to produce larvae for sequencing were either wild-caught (F0) or lab-raised over *n* generations (indicated by F*n*). Individuals were sampled eight days post fertilization (dpf), 8-10 dpf, or 17-20 dpf.

mothers	fathers	offspring sampled	lake population	sequencing round	stage (dpf)
F0 generalist	F0 molluscivore	3 hybrids	osprey lake	4	8
F0 generalists	F0 generalists	3 generalists	osprey lake	3	8
F0 molluscivores	F0 molluscivores	3 molluscivores	osprey lake	3	8
F0 generalists	F0 generalists	3 generalists	crescent pond	3	8
F0 molluscivores	F0 molluscivores	3 molluscivore	crescent pond	4	8
F1 generalists	F1 generalists	3 generalists	little lake	1	8-10
F2 molluscivores	F2 molluscivores	3 molluscivores	little lake	1	8-10
F2 generalists	F2 generalists	3 generalists	crescent pond	1	8-10
F2 molluscivores	F2 molluscivores	3 molluscivores	crescent pond	1	8-10
F2 generalist	F3 molluscivore	4 hybrids	little lake	2	17-20
F1 generalists	F1 generalists	3 generalists	little lake	1	17-20
F2 molluscivores	F2 molluscivores	3 molluscivores	little lake	1	17-20
F2 generalists	F2 generalists	3 generalists	crescent pond	1	17-20
F2 molluscivores	F2 molluscivores	3 molluscivores	crescent pond	1	17-20

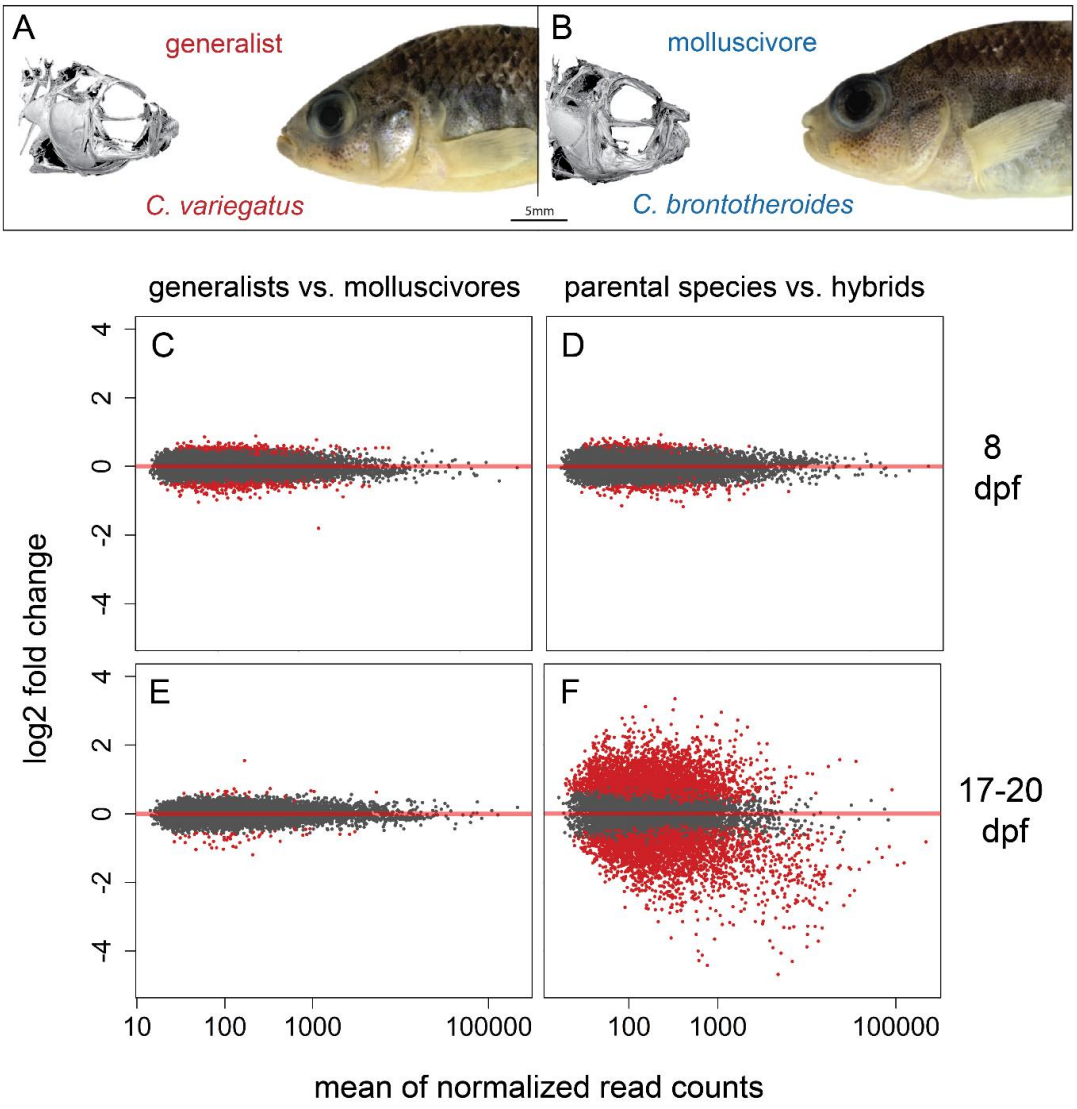

Figure 3.1. Extensive misregulation in F1 hybrid craniofacial tissues.

A) *Cyprinodon variegatus* – the generalist. B) *C. brontotheroides* – the molluscivore (µCT scans of the cranial skeleton of each species modified from (66). Variation in gene expression between generalists vs. molluscivores 8 days post fertilization (dpf), D) parental species vs. hybrids at 8 dpf, E) generalists vs. molluscivores at 17-20 dpf, and F) parental species vs. hybrids at 17-20 dpf. Red points indicate genes detected as differentially expressed at 5% false discovery rate with Benjamini-Hochberg multiple testing adjustment. Grey points indicate genes showing no significant difference in expression between groups. Red line indicates a log2 fold change of zero between groups. Points above/below the line are upregulated/downregulated in molluscivores relative to generalists (C and E) or hybrids relative to parental species (D and F).

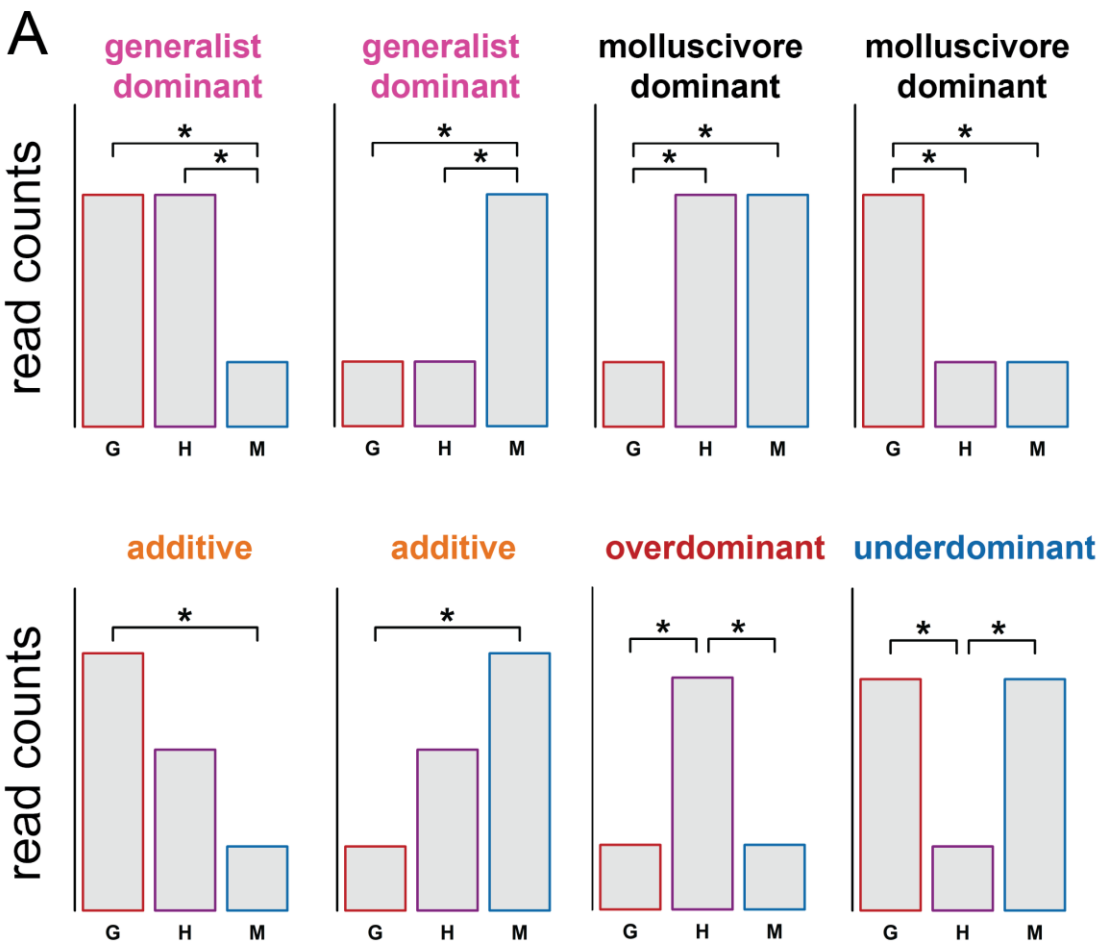

Figure 3.2. Classifying gene expression inheritance in hybrids.

Schematic showing how gene expression inheritance in hybrids was classified. Asterisks indicate significant differential expression between groups. G = generalists, H = F1 hybrids, M = molluscivores.

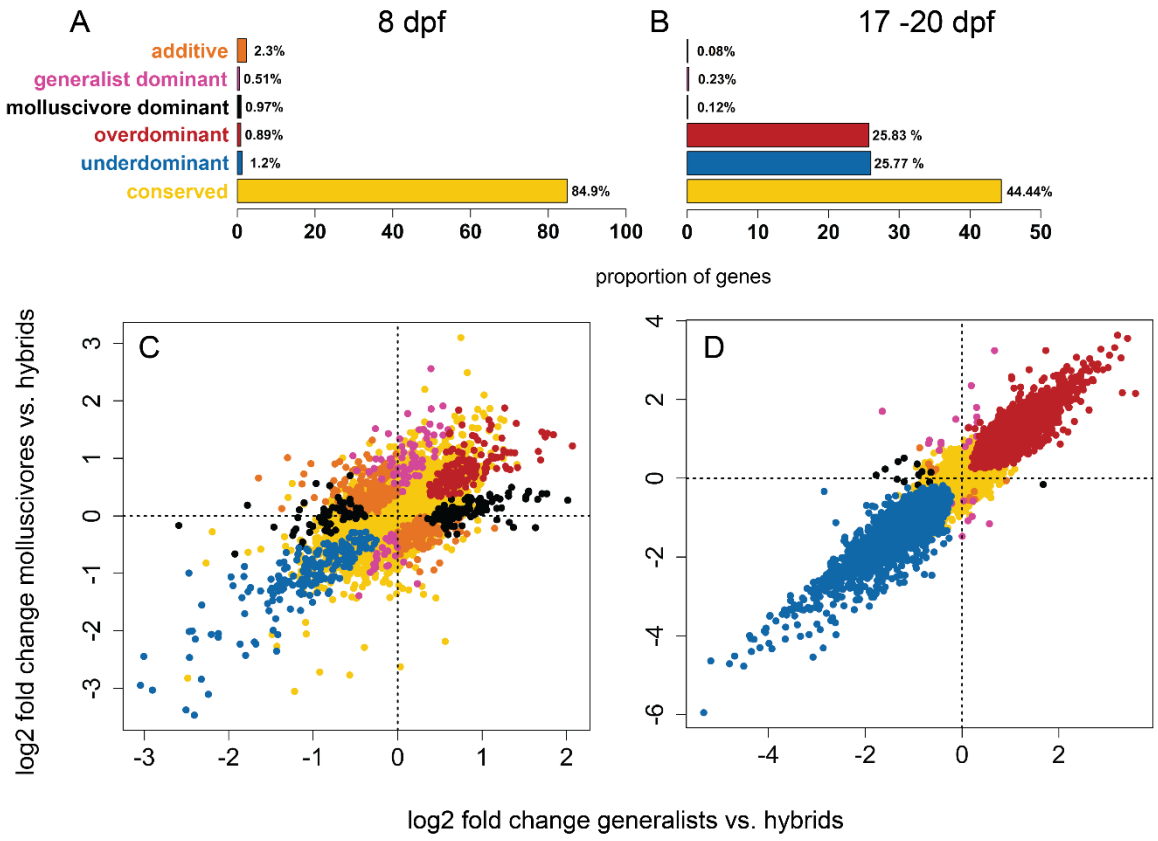

Figure 3.3. Gene expression inheritance in hybrids.

The proportion of A) 17,705 and B) 12,769 genes showing each class of hybrid gene expression inheritance. Log2 fold changes in gene expression between molluscivores vs. hybrids on the y-axis and between generalists vs. hybrids on the x-axis for C) whole-larvae sampled 8 days post fertilization (dpf) and D) craniofacial tissues dissected from 17-20 dpf samples.

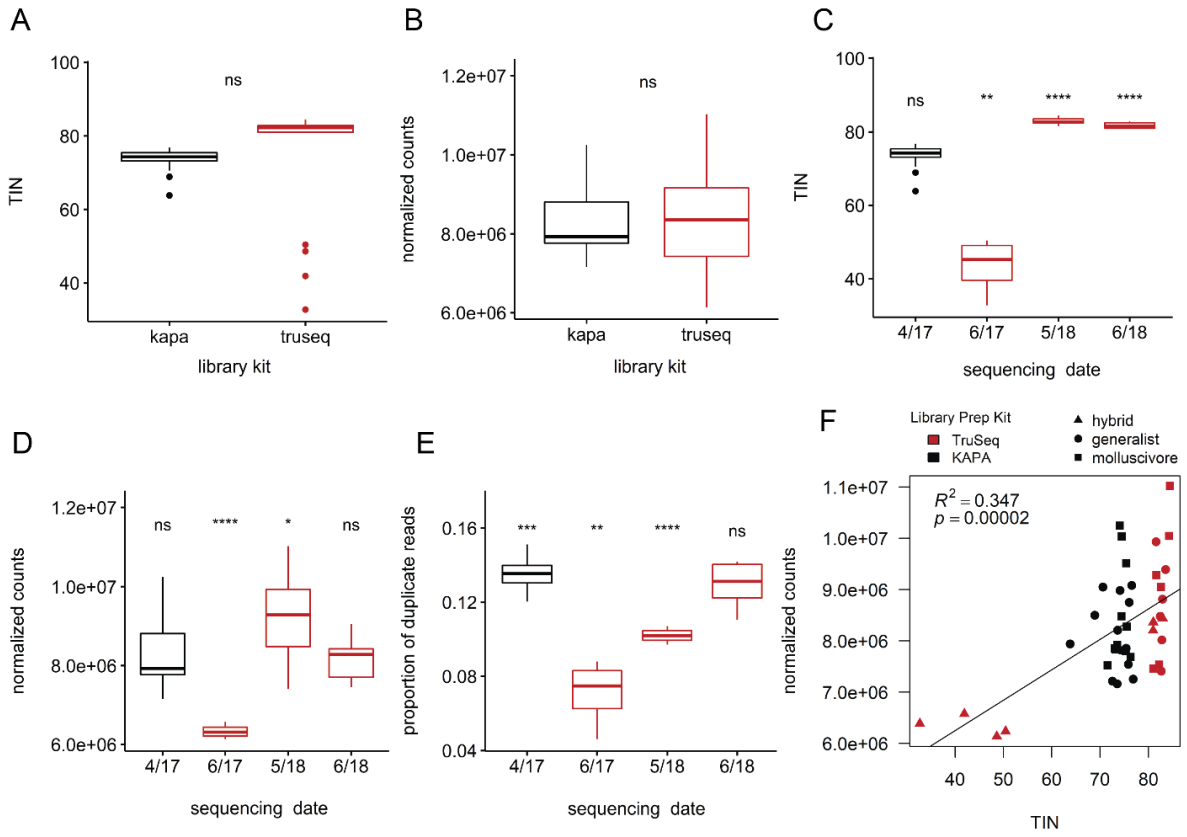

Figure 3.4. Effects of sequencing facility and library preparation kit.

Boxplots show samples grouped by library preparation method (A and B) or by the date they were sequenced (C-E) and whether samples were prepared using Truseq stranded mRNA library kits (red) or KAPA stranded mRNA library kits (black). There was no difference in A) median transcript integrity numbers (TIN) or B) number of normalized counts between groups prepared with different library kits (Welch two sample t-test, $P > 0.05$). Hybrid craniofacial sampled 17-20 days post fertilization (sequenced 6/17) showed significantly lower C) TIN, D) normalized read counts, and D) proportion of duplicate reads compared to samples sequenced on other dates (Pairwise Welch two sample t-test; $P < 0.0001$ = ****, *** = 0.001, ** = 0.01, * = 0.05). F) Lower TIN was correlated with lower normalized read count.

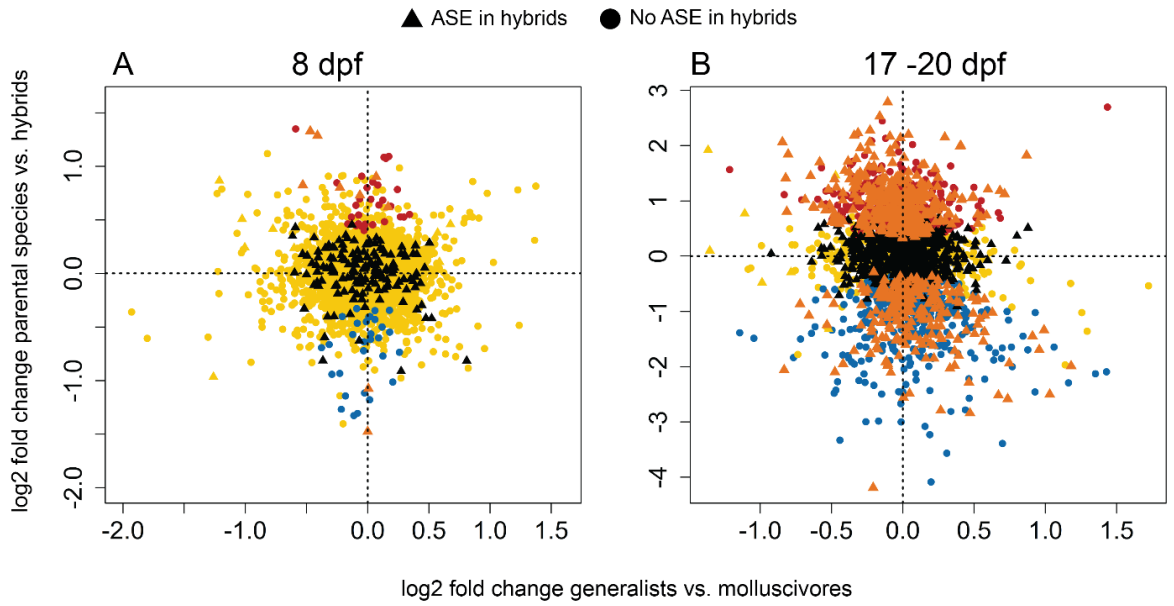

Figure 3.5. Putative compensatory regulation underlying expression divergence between generalists and molluscivores.

Log2 fold changes in gene expression between parental species vs. hybrids on the y-axis and between generalists vs. molluscivores on the x-axis for A) 2,909 genes containing heterozygous sites used for allele specific expression (ASE) analyses in whole-larvae sampled 8 days post fertilization (dpf) and B) 2,403 genes containing heterozygous sites in 17-20 dpf craniofacial samples. Triangle points indicate genes showing significant ASE in all hybrids that did not show ASE in generalists or molluscivores. Circle points indicate genes that did not show significant ASE in hybrids or did not show ASE unique to hybrids. Orange = compensatory regulation and hybrid misregulation (genes showing ASE in hybrids, no difference in expression between generalists and molluscivores, and misregulation in hybrids). Black = compensatory regulation (genes showing ASE in hybrids, no difference in expression between generalists and molluscivores). Blue = overdominant (upregulated in hybrids). Red = underdominant (downregulated in hybrids). Yellow = conserved/ambiguous (No difference in expression between parental populations and hybrids).

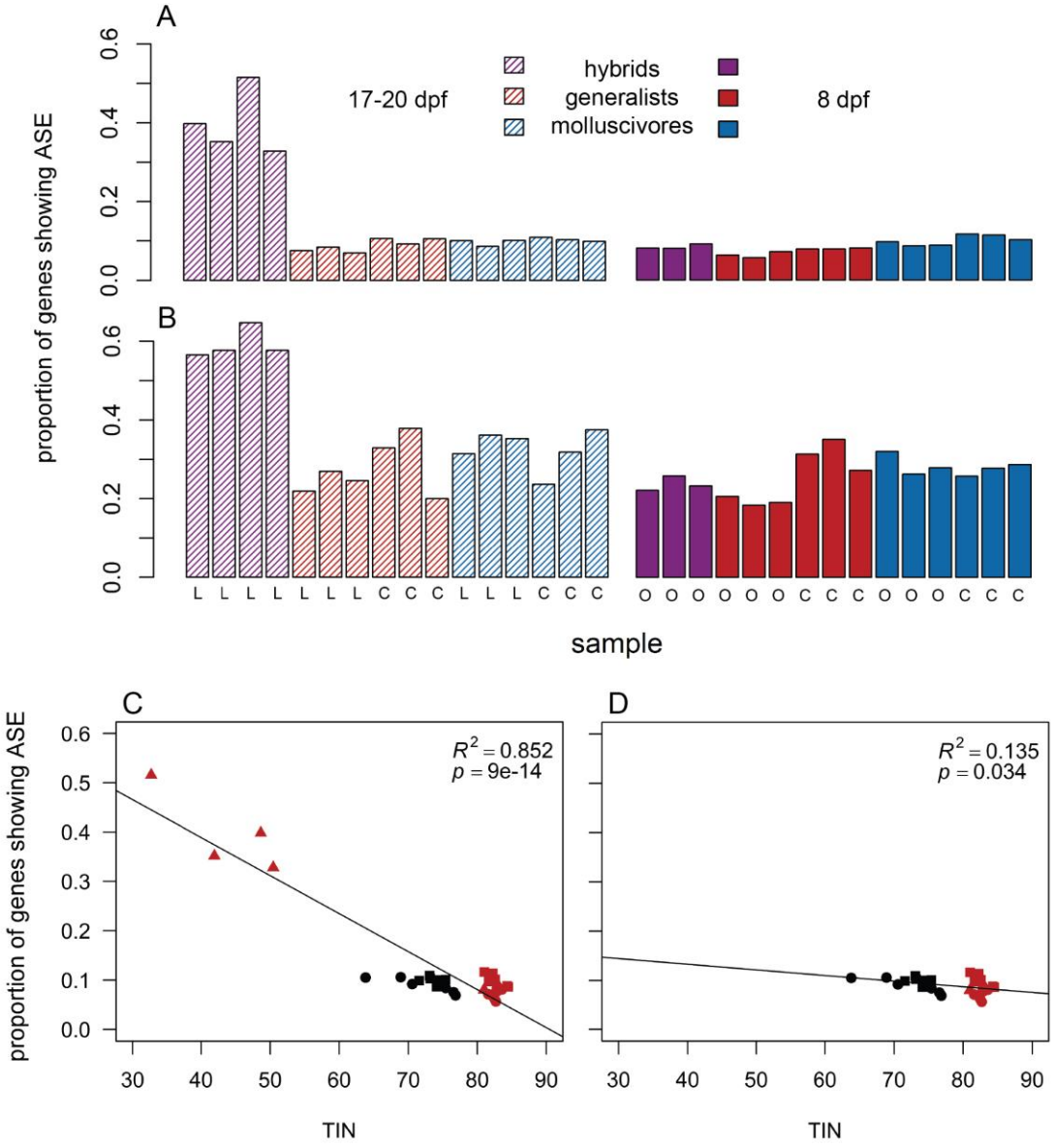

Figure 3.6. Hybrid craniofacial tissues show high levels of allele specific expression.

F1 hybrid craniofacial tissues sampled 17-20 days post fertilization (dpf; striped purple bars) showed a higher proportion of genes showing significant allele specific expression compared to all other samples using a coverage threshold of A) ≥ 10× reads supporting each heterozygous allele (ANOVA, $P = 2.81 \times 10^{-5}$) and B) ≥ 100× reads supporting each allele (ANOVA, $P = 3.85 \times 10^{-4}$). 8 dpf = solid, 17-20 dpf = striped; hybrids = purple, generalists = red, molluscivores = blue; L = Little Lake, C = Crescent Pond, O = Osprey Lake. C) TIN was significantly negatively correlated with ASE (linear regression; $P = 9.04 \times 10^{-14}$). D) This correlation persisted when 17-20 dpf hybrid craniofacial samples were excluded from the linear model (linear regression; $P = 0.034$). However, the observed proportion of genes showing ASE was much higher in 17-20 dpf hybrid craniofacial samples than predicted by the linear model in D.

CHAPTER 4: ECOLOGICAL DIVERGENCE IN SYMPATRY CAUSES GENE MISREGULATION IN HYBRIDS

Introduction

Adaptive radiations showcase dramatic instances of biological diversification resulting from ecological speciation, which occurs when reproductive isolation evolves as a by-product of adaptive divergence between populations (Schluter 2000; Nosil 2012). Ecological speciation predicts that populations adapting to different niches will accumulate genetic differences due to divergent ecological selection, indirectly resulting in reduced gene flow. Gene regulation is a major target of selection during adaptive divergence, with many known cases of divergent gene regulation underlying ecological traits (Parry et al. 2005; Abzhanov et al. 2011; Manceau et al. 2011; Jones et al. 2012; Thompson et al. 2018). However, it is still unknown whether selection on gene regulation can also contribute to reproductive isolation during ecological speciation (Pavey et al. 2010; Mack & Nachman 2017).

Hybridization between divergent populations can break up coadapted genetic variation, resulting in (Bateson) Dobzhansky-Muller incompatibilities (DMIs) if divergent alleles from parental populations are incompatible in hybrids and cause reduced fitness (Coyne & Orr, 2004; Orr, 1996). DMIs between divergent regulatory alleles can result in hybrid gene misregulation: transgressive expression levels that are significantly higher or lower in F1 hybrids than either parental population. Because gene expression is largely constrained by stabilizing selection, gene misregulation is expected to disrupt highly coordinated developmental processes and reduce hybrid fitness (Bedford and Hartl 2009; Signor and Nuzhdin 2018). Indeed, crosses between

distantly related species show that hybrid gene misregulation may be associated with strong intrinsic postzygotic isolation in the form of hybrid sterility and inviability (Landry, Hartl, & Ranz, 2007; Mack, Campbell, & Nachman, 2016; Ortíz-Barrientos, Counterman, & Noor, 2007), although other studies found no association (Wei et al. 2014; Guerrero et al. 2016; Kerwin and Sweigart 2019). Emerging empirical evidence suggests that weak intrinsic DMIs segregate within natural populations (Corbett-detig et al. 2013) and are abundant between recently diverged species, reaching hundreds of incompatibility loci within swordtail fish hybrid zones (Schumer et al. 2014; Schumer and Brandvain 2016). Additionally, hybrid gene misregulation has been reported at early stages of divergence within a species of intertidal copepod (Barreto et al. 2015) and between young species of lake whitefish (Renaut et al. 2009).

Since most studies on hybrid gene misregulation examine distantly related species pairs that exhibit strong intrinsic isolation, the role of regulatory divergence during speciation with gene flow remains largely unexplored. Furthermore, it is debated whether hybrid gene misregulation is driven largely by stabilizing selection or directional selection at early stages of species divergence. Under stabilizing selection, hybrid gene misregulation can evolve due to compensatory *cis*- and *trans*-acting variants with opposing effects on expression levels (Landry et al. 2005; Tulchinsky et al. 2014; Mack and Nachman 2017; Signor and Nuzhdin 2018). Compensatory evolution results in similar gene expression levels between species even though the underlying regulatory machinery has diverged (True and Haag 2001; Wray et al. 2003). Alternatively, directional selection could favor regulatory alleles causing divergent gene expression between species that are incompatible in hybrids (Pavey et al. 2010; Kulmuni and Westram 2017). In this scenario, the same genes showing expression divergence between species should also show misregulation in hybrids.

We examined genetic variation and gene expression divergence within an adaptive radiation to test whether genetic variants causing adaptive gene expression divergence between species may negatively interact to cause gene misregulation in F1 hybrids. If hybrid gene misregulation was influenced by adaptive divergence during ecological speciation, we predicted that 1) gene expression divergence and hybrid gene misregulation should evolve more quickly between ecologically diverged populations compared to populations adapted to similar ecological niches, 2) many of the same genes differentially expressed between species should also show misregulation in F1 hybrids, and 3) these genes should influence adaptive phenotypes and show signs of directional selection. We tested these predictions in a young (10 kya), sympatric radiation of *Cyprinodon* pupfishes endemic to San Salvador Island, Bahamas. This radiation consists of a dietary generalist and two derived specialists adapted to novel trophic niches: a molluscivore (*C. brontotheroides*) and a scale-eater (*C. desquamator*) (Martin & Wainwright, 2013a). All three species coexist in multiple hypersaline lake populations within the same littoral habitat. Hybrids among these species exhibit reduced fitness in the wild and impaired feeding performance in the lab (Martin, 2019; Martin & Wainwright, 2013b). We took a genome-wide approach to identify genetic variation underlying F1 hybrid gene misregulation and found 125 genes that were misregulated, showed high genetic differentiation between species, and were strikingly enriched for developmental functions related to trophic specialization. Our findings show that regulatory variation underlying adaptive changes in gene expression can interact to cause hybrid gene misregulation, which may contribute to reduced hybrid fitness and restrict gene flow between sympatric populations.

Methods

Study system and sample collection

We collected 51 wild-caught individuals from nine isolated hypersaline lakes on San Salvador Island, Bahamas (Great Lake, Stout's Lake, Oyster Lake, Little Lake, Crescent Pond, Moon Rock, Mermaid's Pond, Osprey Lake, Pigeon Creek) between 2011 and 2018 using seine-nets and hand nets. 18 scale-eaters (*Cyprinodon desquamator*) were sampled from six lake populations; 15 molluscivores (*C. brontotheroides*) were sampled from four populations; and 18 generalists (*C. variegatus*) were sampled from nine populations. The genomic dataset also included two *C. laciniatus* from Lake Cunningham, New Providence Island, Bahamas, one *C. bondi* from Etang Saumautre lake in the Dominican Republic, one *C. variegatus* from Fort Fisher, North Carolina, one *C. diabolis* from Devils Hole, Nevada, and captive-bred individuals of *C. simus* and *C. maya* from Laguna Chicancanab, Quintana Roo, Mexico. Sampling is further described in (McGirr & Martin, 2017; Richards & Martin, 2017). Fish were euthanized in an overdose of buffered MS-222 (Finquel, Inc.) following approved protocols from the University of California, Davis Institutional Animal Care and Use Committee (#17455), the University of California, Berkeley Animal Care and Use Committee (AUP-2015-01-7053), and the University of North Carolina Institutional Animal Care and Use Committee (18-061.0). Samples were stored in 95-100% ethanol.

Our total mRNA transcriptomic dataset consisted of 124 *Cyprinodon* exomes from embryos collected between 2017 and 2018. We collected fishes for breeding from two hypersaline lakes on San Salvador Island, Bahamas (Osprey Lake, and Crescent Pond), Lake Cunningham, New Providence Island, Bahamas, and Fort Fisher, North Carolina, United States. Wild-caught parents were reared in breeding tanks at 25–27°C, 10–15 ppt salinity, pH 8.3, and

fed a mix of commercial pellet foods and frozen foods. All purebred F1 offspring were collected from breeding tanks containing multiple F0 breeding pairs. All F1 offspring from crosses between species and populations were collected from individual F0 breeding pairs that were subsequently sequenced in our genomic dataset.

Methods for collecting and raising embryos were similar to previously outlined methods (McGirr & Martin, 2018; McGirr & Martin, 2019). All F1 embryos were collected from breeding mops within one hour of spawning and transferred to petri dishes incubated at 27°C. Embryo water was treated with Fungus Cure (API Inc.) and changed every 48 hours. Embryos were inspected for viability and sampled either 47-49 hours post fertilization (hereafter 2 days post fertilization (2 dpf)) or 190-194 hours (eight days) post fertilization (hereafter 8 dpf). These early developmental stages are described as stage 23 (2 dpf) and 34 (8 dpf) in a recent embryonic staging series of *C. variegatus* (Lencer and McCune 2018). The 2 dpf stage is comparable to the Early Pharyngula Period of zebrafish, when multipotent neural crest cells have begun migrating to pharyngeal arches that will form the oral jaws and most other craniofacial structures (Schilling and Kimmel 1994; Furutani-Seiki and Wittbrodt 2004; Lencer et al. 2017). Embryos usually hatch six to ten days post fertilization, with similar variation in hatch times among species (Lencer et al., 2017; McGirr & Martin, 2018). While some cranial elements are ossified prior to hatching, the skull is largely cartilaginous at 8 dpf (Lencer and McCune 2018). Embryos from each stage were euthanized in an overdose of buffered MS-222 and immediately preserved in RNA later (Ambion, Inc.) for 24 hours at 4°C and then - 20°C for up to 9 months following manufacturer's instructions.

Hybrid cross design

All parents used to generate F1 hybrids were collected from four locations: 1) Crescent Pond, San Salvador Island, 2) Osprey Lake, San Salvador Island, 3) Lake Cunningham, New Providence Island, or 4) Fort Fisher, North Carolina. In order to understand how varying levels of genetic divergence and ecological divergence between parents affected gene expression patterns in F1 offspring, we performed 11 separate crosses falling into three categories. 1) For purebred crosses, we collected F1 embryos from breeding tanks containing multiple breeding pairs from a single location. 2) For San Salvador Island species crosses, we crossed a single individual of one species with a single individual of another species from the same lake for all combinations of the three San Salvador Island species. In order to control for maternal effects on gene expression inheritance, we collected samples from reciprocal crosses for three San Salvador Island species crosses. 3) For outgroup generalist crosses, we bred a Crescent Pond generalist male with a Lake Cunningham female and a North Carolina female (Table D1.1).

Genomic sequencing and alignment

All DNA samples were extracted from muscle tissue or caudal fin clips using DNeasy Blood and Tissue kits (Qiagen, Inc.) and quantified on a Qubit 3.0 fluorometer (Thermofisher Scientific, Inc.). Sequencing methods for 43 of the 58 individuals in our genomic dataset were previously described (McGirr & Martin, 2017; Richards & Martin, 2017). We added 15 new individuals to this dataset that were crossed to generate F1 hybrids. These libraries were prepared at the Vincent J. Coates Genomic Sequencing Center (QB3, Berkeley, CA) using TruSeq kits on the automated Apollo 324 system (WaferGen BioSystems, Inc.). Samples were fragmented using Covaris sonication, barcoded with Illumina indices, quality checked using a Fragment Analyzer

(Advanced Analytical Technologies, Inc.), and sequenced on one lane of Illumina 150PE Hiseq4000 in June 2018.

We filtered raw reads using Trim Galore (v. 4.4, Babraham Bioinformatics) to remove Illumina adaptors and low-quality reads (mean Phred score < 20) and mapped 1,953,034,511 reads to the *Cyprinodon* reference genome (NCBI, *Cyprinodon variegatus* annotation release 100; total sequence length = 1,035,184,475; number of scaffolds = 9,259; scaffold N50 = 835,301; contig N50 = 20,803; (Lencer et al. 2017)) with the Burrows-Wheeler Alignment Tool (bwa mem; (Li and Durbin 2009) (v. 0.7.12)). The Picard software package (v. 2.0.1) and Samtools (v. 1.9) were used to remove duplicate reads (MarkDuplicates) and create indexes. We assessed mapping and read quality using MultiQC (Ewels et al. 2016).

Transcriptomic sequencing and alignment

We extracted RNA from a total of 348 individuals (whole-embryos and whole-larvae) using RNeasy Mini Kits (Qiagen catalog #74104). For samples collected at 2 dpf, we pooled 5 embryos together and pulverized them in a 1.5 ml Eppendorf tube using a plastic pestle washed with RNase Away (Molecular BioProducts). We used the same extraction method for samples collected at 8 dpf but did not pool larvae and prepared a library for each individual separately. Total mRNA sequencing libraries for the resulting 125 samples were prepared at the Vincent J. Coates Genomic Sequencing Center (QB3, Berkeley, CA) using the Illumina stranded Truseq RNA kit (Illumina RS-122-2001). Sequencing was performed on Illumina Hiseq4000 150PE. 72 and 53 total mRNA libraries were each pooled across three lanes and sequenced in May 2018 and July 2018, respectively.

We filtered raw reads using Trim Galore (v. 4.4, Babraham Bioinformatics) to remove Illumina adaptors and low-quality reads (mean Phred score < 20) and mapped 1,638,067,612

filtered reads to the *Cyprinodon* reference genome (NCBI, *Cyprinodon variegatus* annotation release 100; 1.035 Gb; scaffold N50 = 835,301; (Lencer et al. 2017)) using STAR with default parameters (v. 2.5 (Dobin et al. 2013a)). We assessed mapping and read quality using MultiQC (Ewels et al. 2016). We quantified the number of duplicate reads produced during sequence amplification and GC content of transcripts for each sample using RSeQC (Wang et al. 2012). We also used RSeQC to estimate transcript integrity numbers (TINs) which is a measure of potential *in vitro* RNA degradation within a sample (Wang et al. 2012, 2016). We performed one-way ANOVA to determine whether the GC content of reads, read depth across features, total normalized counts, or TINs differed between samples grouped by species and population. We did not find a difference between species or generalist populations for any quality control measure (Fig. D2.1; ANOVA, $P > 0.1$), except for a marginal difference in TIN (Fig. D2.2; ANOVA, $P = 0.041$) driven by slightly higher transcript quality in North Carolina samples (Tukey multiple comparisons of means; $P = 0.043$). We found no significant differences among San Salvador Island generalists, molluscivores, scale-eaters, and outgroup generalists in the proportion of reads that map to annotated features of the *Cyprinodon* reference genome (Fig. D2.3; ANOVA, $P = 0.17$). We did find that more reads mapped to features in 2 dpf samples than 8 dpf samples (Fig. D2.4; Student's *t*-test, $P < 2.2 \times 10^{-16}$).

Variant discovery and population genetic analyses

We used the Genome Analysis Toolkit (v. 3.5 (DePristo et al. 2011)) to call and refine SNP variants across 58 *Cyprinodon* genomes and across 124 *Cyprinodon* exomes using the Haplotype Caller function. For both datasets, we used conservative hard filtering criteria to call SNPs (DePristo et al., 2011; Marsden et al., 2014; McGirr & Martin, 2017). We filtered both SNP datasets to include individuals with a genotyping rate above 90% and SNPs with minor

allele frequencies higher than 5%. Our final filtered genomic SNP dataset included 13,838,603 variants with a mean sequencing coverage of 8.2× per individual.

We further refined our transcriptomic SNP dataset using the allele-specific software WASP (v. 0.3.3) to correct for potential mapping biases that would influence tests of allele-specific expression (ASE; (Degner et al. 2009; Van De Geijn et al. 2015)). While we showed that mapping bias does not significantly affect the proportion of reads mapped to features between species (Fig. D2.1), even a small number of biased sites would likely account for the majority of significant ASE at an exome-wide scale. WASP identified reads that overlapped sites in our original transcriptomic SNP dataset and re-mapped those reads after swapping the genotype for the alternate allele. Reads that failed to map to exactly the same location were discarded. We re-mapped unbiased reads using methods outlined above to create our final BAM files that were used for all downstream analyses. We re-called SNPs using unbiased BAMs for a final transcriptomic SNP dataset that included 413,055 variants with a mean coverage of 1,060× across gene features per individual.

We analyzed genomic SNPs to measure within-population diversity (π), between-population diversity (D_{xy}), relative genetic diversity (F_{st}), and Tajima's D. We measured π, D_{xy}, and F_{st} in 20 kb windows using the python script popGenWindows.py created by Simon Martin (github.com/simonhmartin/genomics_general; (Martin et al., 2013)). We calculated Tajima's D in 20 kb windows and per site F_{st} for each species and lake population genomic using VCFtools (v. 1.15). We chose to analyze 20 kb windows given previous estimates of pairwise linkage disequilibrium (measured as r^2) showing that linkage dropped to background levels between SNPs separated by >20 kb ($r^2 < 0.1$; (McGirr & Martin, 2017)). Tajima's D statistic compares observed nucleotide diversity to diversity under a null model assuming genetic drift, where

negative values indicate a reduction in diversity across segregating sites that may be due to positive selection (Tajima 1989). We also looked for evidence of hard selective sweeps using the SweepFinder method first developed by Nielsen et al. (2005) and implemented in the software package SweeD (Nielsen et al. 2005b; Pavlidis et al. 2013). SweeD separated scaffolds into 1000 windows of equal size and calculated a composite likelihood ratio (CLR) from a comparison of two contrasting models for each window. The first assumes a window has undergone a recent selective sweep, whereas the second assumes a null model where the site frequency spectrum of the window does not differ from that of the entire scaffold. Windows with a high CLR suggest a history of selective sweeps because the site frequency spectrum is shifted toward low-frequency and high-frequency derived variants (Nielsen et al. 2005b; Pavlidis et al. 2013).

We used ancestral population sizes (previously determined by the Multiple Sequentially Markovian Coalescent approach (McGirr & Martin, 2017; Schiffels & Durbin, 2014) to estimate the expected neutral SFS with SweeD, accounting for historical demographic effects on the contemporary shape of the SFS. SweeD identifies regions of a scaffold showing signs of a hard sweep relative to the rest of that scaffold. Thus, we normalized CLR values to be between zero and one to compare the strength of selection across scaffolds. We defined regions showing strong signs of a hard selective sweep as windows that showed CLRs above the 90^{th} percentile for a scaffold (normalized CLR > 0.9) and a negative value of Tajima's D less than the genome-wide 10^{th} percentile (range = -1.62 – -0.77 (see Table D1.2 for all population thresholds)). We also visually inspected regions near candidate incompatibility genes to identify CLRs and Tajima's D estimates indicating moderate signs of selection.

Read count abundance and differential expression analyses

We used the featureCounts function of the Rsubread package (Liao et al. 2014) to generate read counts across 36,511 previously annotated features for the *Cyprinodon* reference genome (Lencer et al. 2017). We aggregated read counts at the transcript isoform level (36,511 isoforms correspond to 24,952 protein coding genes).

We used DESeq2 (v. 3.5 (Love et al. 2014)) to normalize raw read counts and perform principal component analyses. Gene features showing less than 10 normalized counts in every sample were discarded from analyses. DESeq2 fits negative binomial generalized linear models for each gene across samples to test the null hypothesis that the fold change in gene expression between two groups is zero. The program uses an empirical Bayes shrinkage method to determine gene dispersion parameters, which model within-group variability in gene expression and logarithmic fold changes in gene expression. Significant differential expression between groups was determined with Wald tests by comparing normalized posterior log fold change estimates and correcting for multiple testing using the Benjamini–Hochberg procedure with a false discovery rate of 0.05 (Benjamini and Hochberg 1995).

We constructed a DESeqDataSet object in R using a multi-factor design that accounted for variance in F1 read counts influenced by parental population origin and sequencing date (design = ~sequencing_date + parental_breeding_pair_populations). Next, we used a variance stabilizing transformation on normalized counts and performed a principal component analysis to visualize the major axes of variation in 2 dpf and 8 dpf samples (Fig. D2.6). We removed one 8 dpf outlier so that the final count matrix used for differential expression analyses included 124 samples (2 dpf = 68, 8 dpf = 56). We contrasted gene expression in pairwise comparisons between populations grouped by developmental stage (Table D1.1). To determine within

population levels of expression divergence (Fig. 4.1B-E open circles), we down-sampled each population to perform every pairwise comparison between samples using the highest sample size possible between groups and calculated the mean number of genes differentially expressed across comparisons.

Hybrid misregulation and inheritance of gene expression patterns

We generated F1 hybrid offspring from crosses between populations and generated purebred F1 offspring from crosses within populations. We compared expression in hybrids to expression in purebred offspring to determine whether genes showed additive, dominant, or transgressive patterns of inheritance in hybrids. To categorize hybrid inheritance for F1 offspring generated from a cross between a female from population A and a male from population B ($F1_{(A\times B)}$), we conducted four pairwise differential expression tests with DESeq2:

1) $F1_{(A)}$ vs. $F1_{(B)}$

2) $F1_{(A)}$ vs. $F1_{(A\times B)}$

3) $F1_{(B)}$ vs. $F1_{(A\times B)}$

4) $F1_{(A)} + F1_{(B)}$ vs. $F1_{(A\times B)}$

Hybrid inheritance was considered additive if hybrid gene expression was intermediate between parental populations and significantly different between parental populations. Inheritance was dominant if hybrid expression was significantly different from one parental population but not the other. Genes showing misregulation in hybrids showed transgressive inheritance, meaning hybrid gene expression was significantly higher (overdominant) or lower (underdominant) than both parental species (Fig. D2.7-9). We describe transgressive expression as hybrid gene misregulation, which results from divergence in regulatory machinery between

parental species, rather than hybrid gene misexpression, which results from differences in developmental rates that lead to differences in the relative abundance of specific cell types in hybrids compared to parental species. We see no evidence for differences in hatch time between these very closely related species, nor between hybrids and parental species (Lencer, Riccio, & McCune, 2016; McGirr & Martin, 2017; McGirr & Martin, 2019). Furthermore, crosses between all San Salvador island species result in fertile F1 and later generation hybrids. This is contrasts with observations in other systems examining gene regulatory evolution between distantly related species pairs showing strong intrinsic postzygotic isolation (Ranz et al. 2004; Coolon et al. 2014; Mack et al. 2016). In these systems, differences in gene expression between hybrids and parents may be due to aberrant development of reproductive tissues (hybrid dysfunction).

Parallel changes in gene expression in specialists

We looked at the intersection of genes differentially expressed between generalists versus molluscivores and generalists versus scale-eaters to determine whether specialists showed parallel changes in expression relative to generalists (McGirr & Martin, 2018). We also examined the direction of expression divergence for each gene to evaluate the significance of parallel expression evolution (Fig 3E). Specifically, we wanted to know whether the fold change in expression for genes tended to show the same sign in both specialists relative to generalists (either up-regulated in both specialists relative to generalists or down-regulated in both specialists). Under a neutral model of gene expression evolution, half of the genes differentially expressed between generalists versus molluscivores and generalists versus scale-eaters would show fold changes in the same direction and half would show fold changes in opposite directions (Fig. 4.3E). Remarkably, 1,206 (96.6%) of the genes showing expression divergence between generalists versus molluscivores and generalists versus scale-eaters showed the same direction of

expression divergence in specialists. These results provide robust evidence for parallel changes in expression underlying divergent trophic adaptation and confirm our previous findings based on a smaller sample size (McGirr & Martin, 2018).

We wanted to determine whether significant parallelism at the level of gene expression in specialists was mirrored by parallel regulatory mechanisms. We predicted that genes showing parallel changes in specialists would show conserved expression levels in specialist hybrids if they were controlled by the same (or compatible) regulatory mechanisms, but would be misregulated in specialist hybrids if expression was controlled by different and incompatible regulatory mechanisms. We identified genes showing conserved levels of expression in specialist hybrids (no significant difference in expression between purebred specialist F1s and specialist hybrid F1s) and genes showing misregulation in specialist hybrids. We also identified genes showing extreme Caribbean-wide misregulation in specialists. These genes were differentially expressed in specialist hybrids relative to all other samples in our dataset from across the Caribbean (North Carolina to New Providence Island, Bahamas).

Allele specific expression and mechanisms of regulatory divergence

We partitioned hybrid gene expression divergence into patterns that could be attributed to *cis*-regulatory variation in cases where linked genetic variation affected proximal gene expression levels, and *trans*-regulatory variation in cases where genetic variation in unlinked factors bound to *cis*-regulatory elements affected gene expression levels. It is possible to identify mechanisms of gene expression divergence between parental species by bringing *cis* elements from both parents together in the same *trans* environment in F1 hybrids and quantifying allele specific expression (ASE) of parental alleles at heterozygous sites (Cowles et al. 2002; Wittkopp et al. 2004). A gene showing ASE in F1 hybrids that is differentially expressed between parental

species is expected to result from *cis*-regulatory divergence. *Trans*-regulatory divergence can be determined by comparing the ratio of gene expression in parents with the ratio of allelic expression in F1 hybrids. *Cis* and *trans* regulatory variants often interact to affect expression divergence of the same gene (Landry et al., 2005; McManus et al., 2010; Wittkopp et al., 2004).

Our genomic variant dataset included every parent used to generate F1 hybrids between populations (n = 15). We used the VariantsToTable function of the Genome Analysis Toolkit (DePristo et al. 2011) to output genotypes across 13.8 million variant sites for each parent and overlapped these sites with the 413,055 variant sites identified across F1 transcriptomes (corrected for mapping bias with WASP). To categorize mechanisms of regulatory divergence between two populations, we used custom R and python scripts (github.com/joemcgirr/fishfASE) to identify SNPs that were alternatively homozygous in breeding pairs and heterozygous in their F1 offspring. We counted reads across heterozygous sites using ASEReadCounter and matched read counts to maternal and paternal alleles. We calculated the significance of ASE per gene transcript. We identified significant ASE using a beta-binomial test comparing the maternal and paternal counts at each transcript with the R package MBASED (Mayba et al. 2014). For each F1 hybrid sample, we performed a 1-sample analysis with MBASED using default parameters run for 1,000,000 simulations to identify transcripts showing significant ASE ($P < 0.05$). Finally, we quantified allele counts across all heterozygous sites for each purebred F1 sample and ran the same analyses in MBASED to identify transcripts showing ASE in parental populations. A transcript was considered to show ASE if it showed significant ASE in all F1 hybrid samples generated from the same breeding pair and did not show significant ASE in purebred F1 offspring generated from the same parental populations.

In order to determine regulatory mechanisms controlling expression divergence between parental species, a transcript had to be included in differential expression analyses and ASE analyses. We were able to classify regulatory categories for more transcripts if breeding pairs were more genetically divergent because we could analyze more heterozygous sites in their hybrids (mean number of informative transcripts across crosses = 1,914; range = 812 – 3,543). For each hybrid sample and each transcript amenable to both types of analyses, we calculated H – the ratio of maternal allele counts compared to the number of paternal allele counts in F1 hybrids, and P – the ratio of normalized read counts in purebred F1 offspring from the maternal population compared to read counts in purebred F1 offspring from the paternal population. We performed a Fisher's exact test using H and P to determine whether there was a significant *trans*-contribution to expression divergence, testing the null hypothesis that the ratio of read counts in the parental populations was equal to the ratio of parental allele counts in hybrids (Wittkopp et al. 2004; McManus et al. 2010; Goncalves et al. 2012; Mack et al. 2016).

We classified expression divergence due to *cis*-regulation if a transcript showed significant ASE, significant differential expression between parental populations of purebred F1 offspring, and no significant *trans*- contribution. We identified expression divergence due to *trans*-regulation if transcripts did not show ASE, were differentially expressed between parental populations of purebred F1 offspring, and showed significant *trans*- contribution. We defined compensatory regulatory divergence (*cis*- and *trans*-regulatory factors had opposing effects on expression) as cases where a transcript showed ASE and was not differentially expressed between parental populations of purebred F1 offspring (Fig. D2.10-S12) because compensatory evolution is expected to cause divergence in gene regulatory elements that maintain similar

levels of gene expression between species (Landry et al., 2007; McGirr & Martin, 2019; Signor & Nuzhdin, 2018).

Phylogenetic analyses

In order to determine the relationship between expression divergence, hybrid gene misregulation, and phylogenetic distance, we constructed a maximum likelihood tree using RAxML. We excluded all missing sites and sites with more than one alternate allele from our genomic SNP dataset, leaving 1,737,591 variants across 58 individuals for analyses. We performed ten separate searches with different random starting trees under the GTRGAMMA model. Node support was estimated from 1,000 bootstrap samples. We used branch lengths from the best fitting tree as a measure of phylogenetic distance between populations.

We tested whether isolation by distance (kilometers separating populations) was a significant predictor of gene expression divergence between populations. We also tested whether isolation by distance explained patterns of misregulation in hybrids generated by inter-population crosses. Gene expression levels between species cannot be considered to be independent and identically distributed random variables (Felsenstein 1985) . We used phylogenetic generalized least-squares (PGLS) models in R, using the packages ape (Paradis and Schliep 2019) and nlme to assess whether gene expression patterns were predicted by distance between populations (measured in kilometers) after accounting for phylogenetic relatedness. We excluded Osprey Lake populations from these analyses because outgroup generalist hybrid crosses only involved Crescent Pond generalists. We used lake diameter as the maximum estimate of the distance between populations for sympatric comparisons.

Morphometrics

We used digital calipers to measure upper oral jaw length and body length from external landmarks on ethanol-preserved tissue specimens. Upper jaw length was measured from the quadroarticular joint to the tip of the most anterior tooth on the dentigerous arm of the premaxilla. Body length was measured from the midline of the posterior margin of the caudal peduncle to the tip of the lower jaw. We used this measure of body length rather than standard length to account for size variation because the nasal protrusion on some molluscivore samples extended beyond the upper jaw. One scale-eater specimen was removed from the analysis because the caudal region was missing, preventing an accurate measure of body length. All jaw length measurements were log-transformed and regressed against log-transformed body length to remove the effects of size variation among specimens. Size-corrected residuals were used for genome-wide association mapping

Association mapping

We employed a Bayesian Sparse Linear Mixed Model (BSLMM) implemented in the GEMMA software package ((Zhou et al. 2013) v. 0.94.1) to identify genomic regions associated with variation in upper oral jaw length. We previously used this program to identify candidate genes influencing jaw size (McGirr & Martin, 2017). Here, we used the same methods adding 15 individuals to our genomic dataset. Briefly, the BSLMM uses Markov Chain Monte Carlo sampling to estimate the proportion of phenotypic variation explained by every SNP included in the analysis (PVE), the proportion of phenotypic variation explained by SNPs of large effect (PGE), which are defined as SNPs with a non-zero effect on the phenotype, and the number of large-effect SNPs needed to explain PGE (nSNPs; Fig. D2.13). GEMMA also estimates an effect size coefficient (β) and a posterior inclusion probability (PIP) for each SNP. We used PIP (the

proportion of iterations in which a SNP is estimated to have a non-zero effect on phenotypic variation ($\beta \neq 0$)) to assess the significance of regions associated with jaw size variation. Because these statistics are difficult to interpret for causal SNPs tightly linked to neutral SNPs, we summed β and PIP parameters across 20 kb windows to avoid dispersion of the posterior probability density across SNPs in strong linkage disequilibrium (LD). GEMMA controls for background population structure by estimating and incorporating a kinship relatedness matrix as a covariate in the regression model. We performed 10 independent runs of the BSLMM for 57 individuals (following (Comeault et al. 2014)) using a step size of 100 million with a burn-in of 50 million steps. Independent runs were consistent in reporting the strongest associations for the same 20 kb windows. Windows that showed PIP values above the 99th percentile (0.00175) were considered to be strongly associated with oral jaw size variation within Caribbean pupfishes. Our PIP estimates for strongly associated windows suggest that jaw length may be controlled by several loci of moderate effect (see bimodal PGE distribution, Fig. D2.13 B). Indeed, a linkage mapping analysis of phenotypic diversity in an F_2 intercross between specialists estimated up to four QTL with moderate effects on oral jaw size explaining up to 15% of the phenotypic variation (Martin, Erickson, & Miller, 2017). Encouragingly, the window that showed the strongest association with jaw size (PIP = 0.1043; Fig. D2.13) contained a single gene associated with craniofacial deformities in humans (*samd12*; (Oliver et al. 2019)). Additionally, *clk2, gpr119, doc2b, rapgef4,* were also within the top four windows showing the highest PIP values.

Gene ontology enrichment analyses

We performed a gene ontology (GO) enrichment analysis for the 125 genes in differentiated genomic regions showing differential expression between species and misregulation in hybrids using ShinyGo v.0.51 (Ge and Jung 2018). The RefSeq genome records

for the *Cyprinodon* reference genome were annotated by the NCBI Eukaryotic Genome Annotation Pipeline, an automated pipeline that annotates genes, transcripts and proteins. Gene symbols for orthologs identified by this pipleline largely match human gene symbols. Thus, we searched for enrichment across biological process ontologies curated for human gene functions. We also determined whether genes sets showing other interesting patterns of expression were annotated for effects on cranial skeletal system development (GO:1904888).

Results

Trophic specialization, not geographic distance, drives major changes in gene expression and hybrid gene misregulation

Gene expression divergence is expected to increase with increasing phylogenetic distance between closely related species, and is expected to increase more rapidly when directional selection on gene expression in strong (Whitehead and Crawford 2006). Since allopatric generalist populations are adapted to similar ecological niches and sympatric specialist species are adapted to divergent niches (Martin, 2016b), stronger selection on gene expression in the specialist species may contribute to faster gene expression divergence between sympatric species than between allopatric generalists. However, gene expression levels among species cannot be considered to be independent and identically distributed random variables. Thus, we predicted that gene expression divergence should be higher between sympatric specialists than between allopatric generalists after controlling for genetic divergence among all populations. To test this, we determined whether isolation by distance explained patterns of gene expression divergence while controlling for phylogenetic relatedness using a maximum likelihood tree estimated with RAxML from 1.7 million SNPs (Fig. 4.1; Fig. D2.5).

Overall, genetic divergence increased with geographic distance between allopatric generalist populations and was lowest between sympatric populations (Table D1.3; genome-wide mean *Fst* measured across 13.8 million SNPs: San Salvador generalists vs. North Carolina = 0.217; vs. New Providence = 0.155; vs. scale-eaters = 0.106; vs. molluscivores = 0.056). Geographic distance among populations was a significant predictor of the proportion of differential gene expression between populations at two days post fertilization (2 dpf) (Fig. 4.1B; phylogenetic generalized least squares (PGLS); $P = 0.02$). This is consistent with a model of gene expression evolution governed largely by stabilizing selection and drift (Whitehead and Crawford 2006). However, at eight days post fertilization (8 dpf), when craniofacial structures of the skull begin to ossify (Lencer and McCune 2018), geographic distance was no longer associated with differential expression (Fig. 4.1C; PGLS; $P = 0.18$), which was higher between sympatric trophic specialist species on San Salvador Island than between generalist populations spanning 1000 km across the Caribbean. Thus, differential gene expression at 8 dpf was much higher than expected due to isolation by distance, suggesting that strong directional selection on gene expression was important during ecological divergence in sympatry.

Similar to expectations for gene expression divergence between species, the extent of F1 hybrid gene misregulation likely depends on genetic divergence between parental species (Coolon et al. 2014). Thus, we predicted to find higher levels of gene misregulation in specialist F1 hybrids than allopatric generalist F1 hybrids after accounting for phylogenetic relationships. Consistent with this prediction, geographic distance between parental populations was not associated with gene misregulation in F1 hybrids at either developmental stage (Fig. 4.1D and E; PGLS; 2 dpf $P = 0.17$; 8dpf $P = 0.38$). This was due to the high number of genes misregulated in Crescent Pond molluscivore × scale-eater hybrids (9.3% of genes) and Crescent Pond generalist

× scale-eater hybrids (7.6% of genes). This amount of gene misregulation is comparable to species pairs with much greater divergence times (Coolon et al. 2014; Mack et al. 2016). Together, these results suggest that positive selection on gene expression has shaped patterns of expression divergence between sympatric San Salvador Island species as well as patterns of gene misregulation in their F1 hybrids.

Genes differentially expressed between species are misregulated in F1 hybrids

Hybrid gene misregulation can result from stabilizing selection or directional selection (including divergent selection) on gene expression (Landry et al., 2007; Mack & Nachman, 2017; Signor & Nuzhdin, 2018). When stabilizing selection favors an optimal level of gene expression, hybrid gene misregulation is expected to result from epistasis between *cis* and *trans* compensatory variants that have accumulated between diverging lineages. In order to determine regulatory mechanisms underlying hybrid gene misregulation, we measured allele specific expression across genes containing heterozygous sites in F1 hybrids that were homozygous in their parents. Out of 3,669 misregulated genes amenable to this analysis, 819 (22.3%) showed allele specific expression and were not differentially expressed between parental populations. This expression pattern is consistent with compensatory regulation underlying misregulation, indicating stabilizing selection acting on gene expression (Fig. D2.10-12, Table D1.4).

Alternatively, if directional selection on regulatory variants contributed to hybrid gene misregulation, we would expect the same genes showing differential expression between species to show misregulation in F1 hybrids. Thus, we intersected genes that were differentially expressed between San Salvador Island species with genes showing misregulation in F1 hybrids to identify two types of expression patterns consistent with directional selection on regulatory genetic variants causing adaptive expression divergence between species.

First, we found 716 genes that showed differential expression between San Salvador Island species that were also misregulated in their F1 hybrids (Fig. 4.2, Table D1.5). We found that 69.8% of these genes were only misregulated at 8 dpf in comparisons involving scale-eaters (Fig. 4.2A-H). Additionally, nearly all of the 716 genes (712; 99.4%) were misregulated in only one lake population. This may suggest that incompatible alleles contributing to misregulation are segregating within species and between lake populations (Corbett-detig et al. 2013). However, we also found four genes that showed differential expression between species and misregulation their hybrids in both lake comparisons (*trim47*, *krt13*, *s100a1*, *elovl7*; Table D1.6).

Second, we identified genes showing parallel expression divergence in both specialist species relative to generalists that were misregulated in specialist F1 hybrids (Fig. 4.3). This pattern likely results from parallel expression in molluscivores and scale-eaters controlled by different genetic mechanisms (McGirr & Martin, 2018). Significantly more genes showed differential expression in both specialist comparisons than expected by chance (Fig. 4.3A-D; Fisher's exact test, $P < 2.7 \times 10^{-5}$). Of these, 96.6% (1,206) showed the same direction of expression in specialists relative to generalists. This was much more than expected under a neutral model of gene expression evolution, where a gene would be equally likely to show expression divergence in opposite directions in specialists (Fig. 4.3E and F; binomial test, $P < 1.0 \times 10^{-16}$). 45 of the 1,206 genes showing parallel expression divergence in specialists also showed misregulation in specialist F1 hybrids (Fig. 4.3F). Eight of these genes were severely misregulated to the extent that they were differentially expressed in hybrids relative to all other populations in our dataset. For example, *sypl1* showed significantly higher expression in 8 dpf Crescent Pond molluscivore × scale-eater F1 hybrids than all other crosses spanning 1000 km from San Salvador Island, Bahamas to North Carolina, USA ($P = 2.35 \times 10^{-4}$; Fig. 4.3G).

Overexpression of this gene is associated with epithelial-mesenchymal transition, an important process during cranial neural crest cell migration (Kang and Svoboda 2005; Chen et al. 2017). Similarly, *scn4a* showed significantly lower expression in 8 dpf Crescent Pond specialist F1 hybrids than all other crosses ($P = 5.49 \times 10^{-4}$; Fig. 4.3H). Mutations in this gene are known to cause paramyotonia congenita, a disorder causing weakness and stiffness of craniofacial skeletal muscles (Huang et al. 2019).

Misregulated genes under selection influence adaptive ecological traits in trophic specialists

If hybrid gene misregulation resulted from adaptive gene regulatory divergence between species, we predicted that these genes should influence the development of divergent traits and show genetic signatures of selection. Out of 750 total unique genes identified above as differentially expressed between populations and misregulated in F1 hybrids, 125 (17%) were within 20 kb of SNPs that were fixed between populations ($F_{st} = 1$) and within 20 kb windows showing high absolute genetic divergence between populations ($D_{xy} \geq$ genome-wide 90th percentile; range: 0.0031 – 0.0075; Table D1.3). These 125 genes were significantly enriched for functional categories highly relevant to divergent specialist phenotypes, including head development, brain development, muscle development, and cellular response to nitrogen (FDR = 0.05; Fig. 4.4A, Table D1.7). We refer to these 125 genes as ecological DMI candidate genes because 1) they showed high genetic differentiation between species, 2) were enriched for developmental functions related to divergent adaptive traits, 3) and showed expression patterns consistent with incompatible interactions between divergent regulatory alleles contributing to hybrid gene misregulation.

Twenty six (20.8%) of these ecological DMI candidate genes showed strong evidence of a hard selective sweep in specialists (negative Tajima's D < genome-wide 10th percentile; range:

-1.62 – -0.77; SweeD composite likelihood ratio > 90th percentile by scaffold; Table D1.2 and D1.8), and 16 of these showed at least a two-fold expression difference in F1 hybrids compared to purebred F1. Several ecological DMI candidate genes have known functions that are compelling targets for divergent ecological selection. For example, the autophagy-related gene *map1lc3c* has been shown to influence growth when cells are nitrogen deprived (Otto et al. 2004; Stadel et al. 2015). Given that specialists occupy higher trophic levels than generalists, as shown by stable isotope ratios (δ15N; Fig. 4.5B), expression changes in this gene may be important adaptations to nitrogen-rich diets. Similarly, expression changes in the ten genes annotated for effects on brain development may influence divergent behavioral adaptations associated with trophic specialists, including significantly increased aggression (St John et al. 2019) and female mate preferences (West and Kodric-Brown 2015).

Using a genome-wide association mapping method that accounts for genetic structure among populations (Zhou et al. 2013), we found that nine of the 125 genes in differentiated regions were significantly associated with oral jaw size – the most rapidly diversifying skeletal trait in this radiation (GEMMA PIP > 99th percentile; Table D1.9; Fig. D2.13). For example, we found that *mpp1* was near 170 SNPs fixed between Crescent Pond generalists and scale-eaters, showed evidence of a hard selective sweep in both populations, and was differentially expressed due to *cis* regulatory mechanisms (Fig. 4.4F-I). F1 hybrids showed a 3-fold decrease in expression of *mpp1* (P = 0.001; Fig. 4.4F). Knockouts of this gene were recently shown to cause severe craniofacial defects in humans and mice (Fritz et al. 2014). The other eight genes significantly associated with jaw size have not been previously shown to influence cranial phenotypes, but some have known functions in cell types relevant to craniofacial development (Table D1.9). For example, the gene *sema6c*, which shows strong signs of selection in both

scale-eaters and molluscivores (Fig. D2.14), is known to be expressed at neuromuscular junctions and is important for neuron growth and development within skeletal muscle (Svensson et al. 2008). Expression changes in this gene may influence the development of jaw closing muscles (adductor mandibulae), which tend to be larger in specialists relative to generalists (Fig. 4.4B). Overall, we found candidate regulatory variants under selection that likely contribute to hybrid gene misregulation and demonstrate that genes near these variants are strikingly enriched for developmental functions related to divergent adaptive traits.

Discussion

By combining whole genome sequencing with transcriptomic analyses of developing tissues in recently diverged trophic specialists and their F1 hybrids, we provide a genome-wide view of how ecological selection can influence gene misregulation in hybrids, which may contribute to reduced hybrid fitness. Unlike other studies that examined hybrid gene misregulation between distantly related species pairs exhibiting strong instinsic reproductive isolation (Kerwin & Sweigart, 2019; Landry et al., 2007; Mack & Nachman, 2017), we show that misregulation can evolve between recently diverged species that coexist in sympatry and still produce fertile hybrids. Our results are consistent with negative epistatic interactions between alleles from different parental genomes affecting 750 genes (3% of the transcriptome) that show differential expression between species and misregulation in F1 hybrids. 125 of these genes were in highly differentiated regions of the genome containing SNPs fixed between species which were enriched for developmental processes relevant to trophic specialization, suggesting that misregulation of these candidate genes in F1 and later generations of hybrids may disrupt the function of adaptive traits and contribute to reproductive isolation between these nascent species.

The negative fitness consequences associated with hybrid gene misregulation have been described in several systems (Landry et al., 2007; Mack et al., 2016; Maheshwari & Barbash, 2012; Malone & Michalak, 2008; Ortíz-Barrientos et al., 2007), but most of this research has focused on genes associated with sterility and inviability between highly divergent species (but see (Renaut et al. 2009)). It is clear that these strong intrinsic postzygotic isolating barriers evolve more slowly than premating barriers (Coyne & Orr, 2004; Coyne & Orr, 1989; Turissini, McGirr, Patel, David, & Matute, 2018); however, hybrid gene misregulation may also have non-lethal effects on fitness and performance that could evolve before or alongside premating isolating mechanisms. Additionally, if genes that are differentially expressed between species in developing tissues are important for adaptive trait divergence, then misregulation of those genes could contribute to abnormal phenotypes that are ecologically maladaptive (Renaut et al. 2009; Arnegard et al. 2014; Kulmuni and Westram 2017). We previously found extensive gene misregulation specific to craniofacial tissues, which were dissected from generalist × molluscivore F1 hybrids at an early developmental stage (McGirr & Martin, 2019). Furthermore, F2 and later generation hybrids showing more transgressive phenotypes exhibited the lowest survival and growth rate in field enclosures across multiple lakes and multiple independent field experiments on San Salvador Island (Martin, 2016a; Martin & Wainwright, 2013b). In the lab, generalist × scale-eater F1 hybrids exhibited non-additive and impaired feeding performance on scales (St. John et al. 2020). While it is difficult to demonstrate a causative link between gene misregulation and hybrid fitness without functional validation experiments or recombinant mapping populations, these independent lines of evidence suggest that hybrids among San Salvador Island species suffer reduced performance and survival in both laboratory and field

environments, which may partly result from misregulation of genes that are necessary for the normal development of adaptive traits.

If divergent ecological selection on adaptive traits also causes gene misregulation and subsequently reduced performance and survival of hybrids in the wild, then these ecological DMIs may promote rapid speciation, analogous to the mechanism of magic traits (Servedio et al. 2011). For example, whereas magic traits contribute to reproductive isolation through assortative mating as a by-product of divergent ecological selection, these ecological DMIs contribute to isolation through gene misregulation and reduced hybrid fitness (Kulmuni and Westram 2017). Thus, our results support a mechanism for divergent ecological selection to generate reproductive isolation as a by-product since many adaptive traits are expected to evolve by divergent gene regulation that may come into conflict in a hybrid genetic background (Pavey et al. 2010; Kulmuni and Westram 2017).

Mathematical models and simulations suggest that genetic incompatibilities evolve most rapidly under directional selection (Johnson and Porter 2000; Tulchinsky et al. 2014b), and evolve more slowly under stabilizing selection when compensatory *cis* and *trans* variants have opposing effects on expression levels (Tulchinsky et al. 2014b). We see evidence for both types of selection driving misregulation. Out of the genes showing hybrid misregulation that contained heterozygous variation, 819 showed expression patterns consistent with compensatory regulation, a signature of stabilizing selection (Table D1.4). Alternatively, 750 misregulated genes were differentially expressed between species, a signature of directional selection. Of these genes, 125 were in highly differentiated genomic regions containing SNPs fixed between populations, and 26 genes showed strong evidence of hard selective sweeps. (Table D1.8).

Importantly, even more genes may have experienced soft sweeps that were not detected by our methods.

Although scale-eaters from Crescent Pond and Osprey Lake form a monophyletic group (Fig. D2.5), we found little overlap in misregulated genes between lakes (Fig. 4.2). This may result from selection on Caribbean-wide standing genetic variation that has similar effects on expression, as we showed previously (McGirr & Martin, 2018), and could reflect polymorphic incompatibilities segregating within species (Corbett-detig et al. 2013). We also see distinct intraspecific differences between lake populations of trophic specialists in pigmentation, maxillary protrusion, and other traits (Martin & Feinstein, 2014), consistent with divergent regulatory variation underlying these adaptive phenotypes.

Identifying genetic variation that contributes to adaptive variation and studying its effect on reproductive isolation is important to understand the sequence of molecular changes leading to ecological speciation. We show that ecologically relevant genes near differentiated genetic regions between sympatric species are under selection and misregulated in F1 hybrids. Overall, our results are consistent with previous observations that hybrid incompatibility alleles are often segregating within populations (Reed and Markow 2004; Cutter 2012; Corbett-detig et al. 2013; Larson et al. 2018) and that hundreds of genetic incompatibilities can contribute to reproductive isolation between species at the earliest stages of divergence (Schumer et al. 2014). We extend this emerging consensus by showing that gene misregulation may result as a by-product of divergent ecological selection on a wide range of adaptive traits.

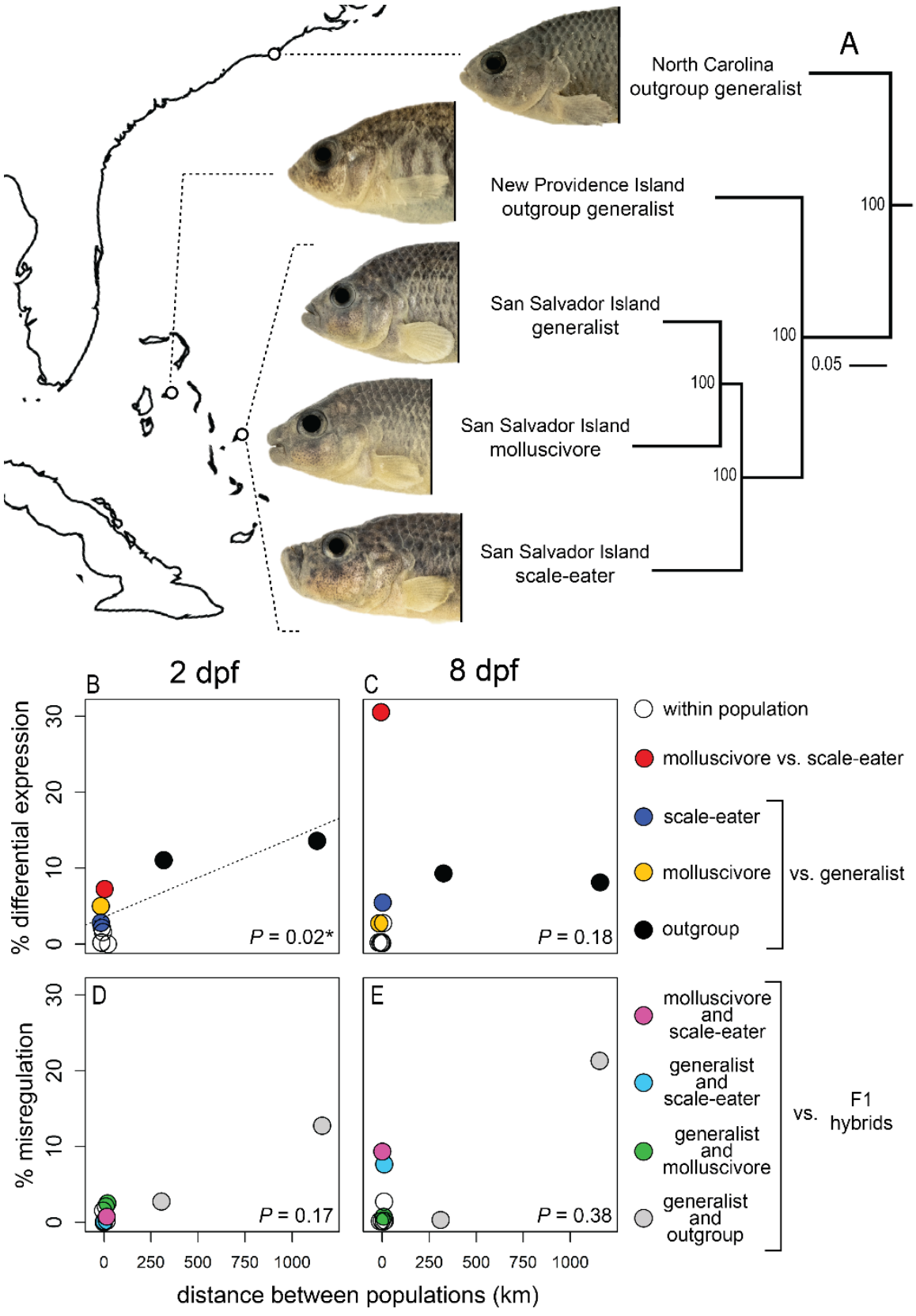
A
North Carolina outgroup generalist
New Providence Island outgroup generalist
San Salvador Island generalist
San Salvador Island molluscivore
San Salvador Island scale-eater
100
100
100
0.05
B
2 dpf
C
8 dpf
% differential expression
P = 0.02*
P = 0.18
within population
molluscivore vs. scale-eater
scale-eater
molluscivore
outgroup
vs. generalist
D
E
% misregulation
P = 0.17
P = 0.38
molluscivore and scale-eater
generalist and scale-eater
generalist and molluscivore
generalist and outgroup
vs. F1 hybrids
0
250
500
750
1000
distance between populations (km)

Figure 4.1. Caribbean-wide patterns of gene expression and misregulation across sympatric and allopatric populations of Cyprinodon pupfishes.

A) Maximum likelihood tree estimated from 1.7 million SNPs showing phylogenetic relationships among generalist populations and specialist species (100% bootstrap support indicated at nodes). B) Geographic distance separating populations was associated with differential gene expression levels in embryos at 2 days post fertilization (2 dpf; phylogenetic least squares $P = 0.02$, dotted regression line). C) In whole larvae at 8 dpf differential expression was not associated with geographic distance (PGLS; $P = 0.18$) and was higher between sympatric specialists (red) than between allopatric generalists separated by 300 and 1000 km (black). D and E) Hybrid gene misregulation for sympatric crosses at 2 dpf and 8 dpf. Geographic distance was not associated with hybrid misregulation at either developmental stage (PGLS; 2 dpf $P = 0.17$; 8dpf $P = 0.38$). Percentages in B-E were measured using Crescent Pond crosses.

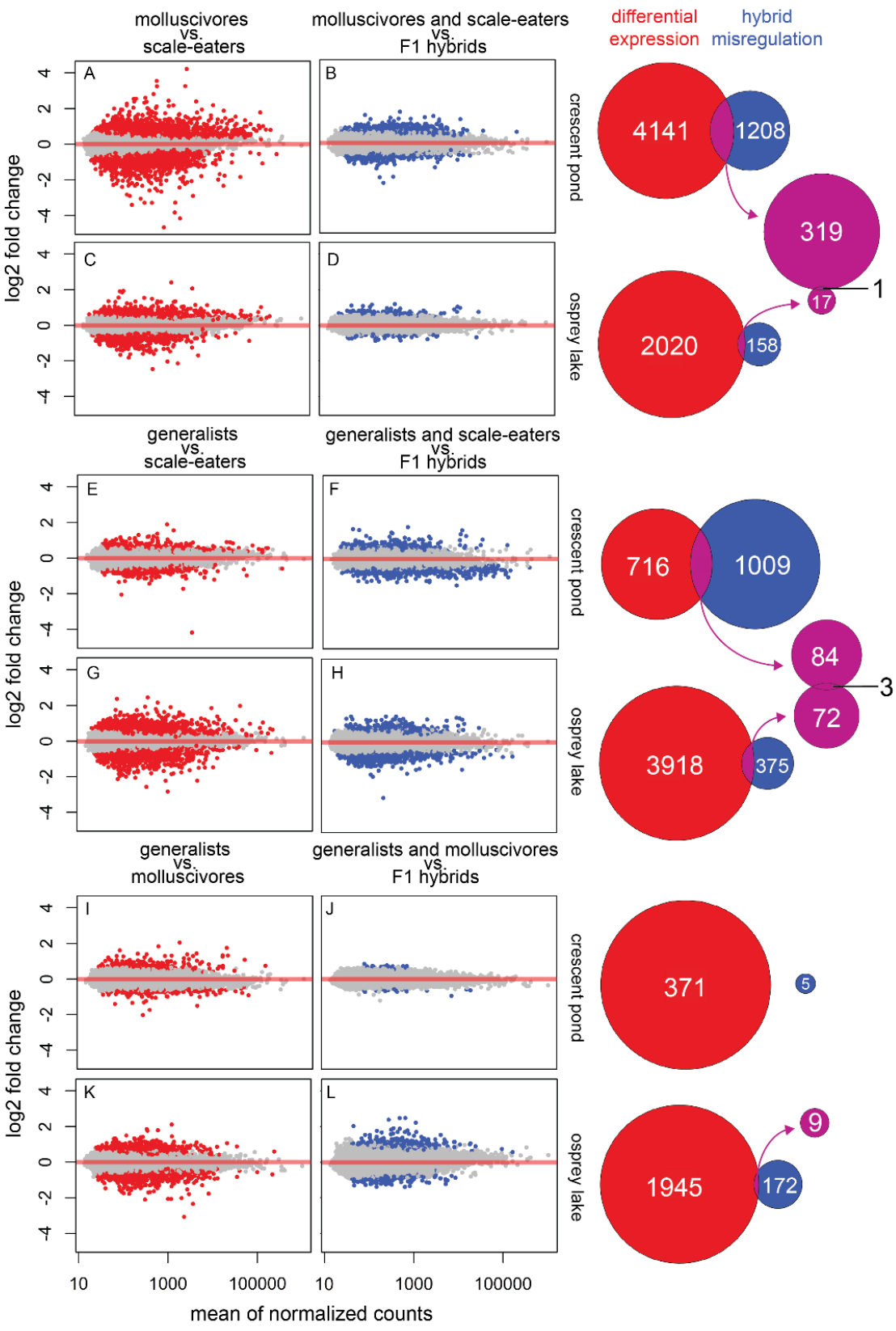
molluscivores vs. scale-eaters
molluscivores and scale-eaters vs. F1 hybrids
differential expression
hybrid misregulation
A
B
C
D
log2 fold change
crescent pond
osprey lake
4141
1208
319
17
1
2020
158
generalists vs. scale-eaters
generalists and scale-eaters vs. F1 hybrids
E
F
G
H
716
1009
84
72
3
3918
375
generalists vs. molluscivores
generalists and molluscivores vs. F1 hybrids
I
J
K
L
371
5
1945
172
9
10
1000
100000
mean of normalized counts

Figure 4.2. Genes differentially expressed between species are misregulated in their F1 hybrids at 8 days post fertilization.

Genes differentially expressed between San Salvador species from Crescent Pond and Osprey Lake are shown in red for molluscivore × scale-eater crosses (A-D), generalist × scale-eater crosses (E-H), and generalist × molluscivore crosses (I-L). Genes misregulated in F1 hybrids are shown in blue. In comparisons involving reciprocal crosses (D, J, and L), we only show genes misregulated in a single cross direction. A total of 716 genes (purple) were differentially expressed between species and also misregulated in their F1 hybrids. Purple Venn diagrams show overlap between lake population comparisons; 4 genes showed differential expression and misregulation in both lake comparisons.

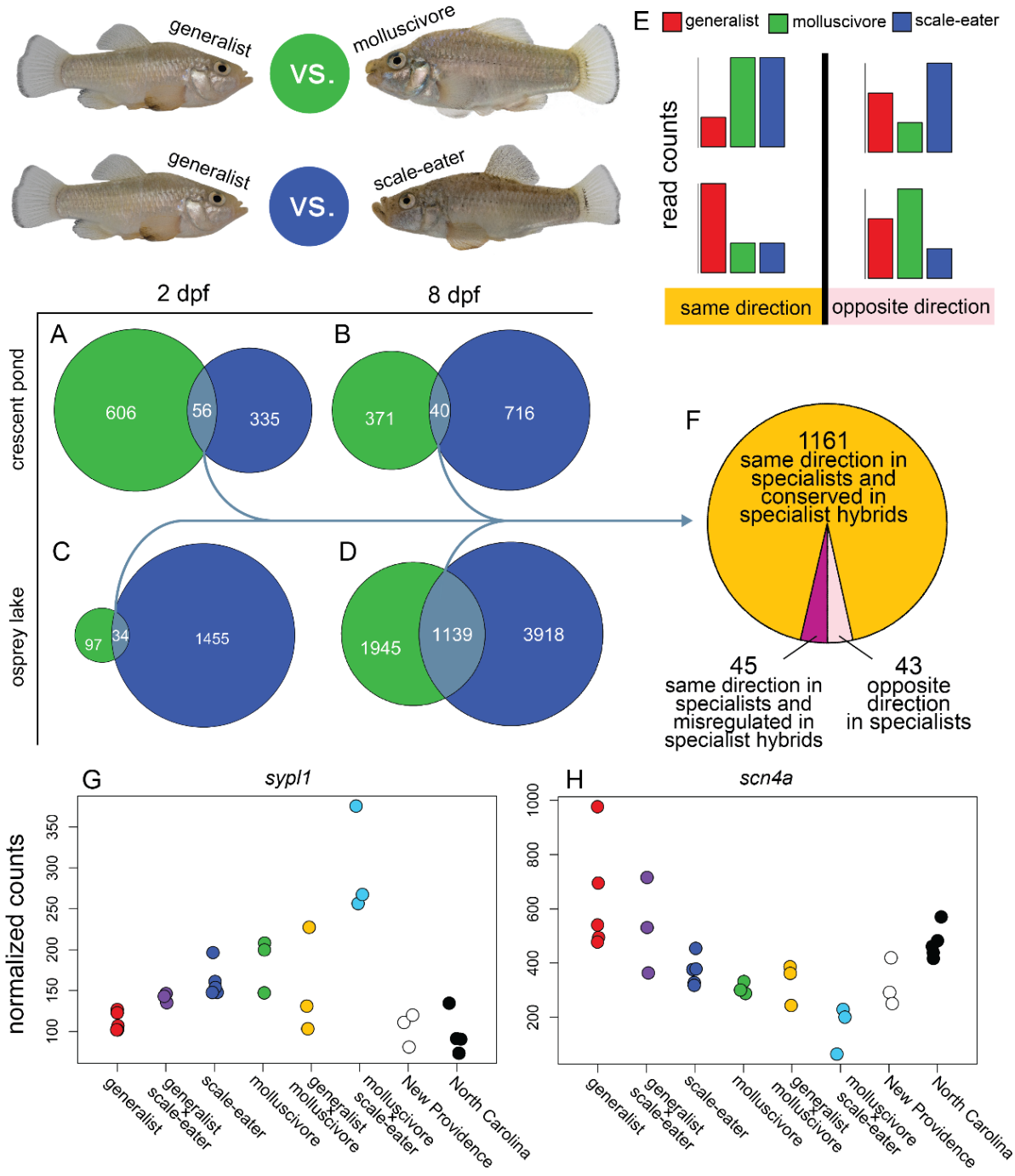
generalist
vs.
molluscivore
generalist
vs.
scale-eater
E
generalist
molluscivore
scale-eater
read counts
same direction
opposite direction
2 dpf
8 dpf
crescent pond
osprey lake
A
606
56
335
B
371
40
716
C
97
34
1455
D
1945
1139
3918
F
1161
same direction in specialists and conserved in specialist hybrids
45
same direction in specialists and misregulated in specialist hybrids
43
opposite direction in specialists
G
sypl1
normalized counts
100
150
200
250
300
350
generalist
generalist x scale-eater
scale-eater
molluscivore
generalist x molluscivore
molluscivore x scale-eater
New Providence
North Carolina
H
scn4a
200
400
600
800
1000

Figure 4.3. Genes showing parallel expression divergence in specialists are misregulated in specialist hybrids.

Genes differentially expressed between generalists and molluscivores (green) were compared to the set of genes differentially expressed between generalists and scale-eaters (dark blue). A-D) Significantly more genes showed differential expression in both specialist comparisons (light blue) than expected by chance in both lakes at both developmental stages (Fisher's exact test, $P < 2.7 \times 10^{-5}$). E) A neutral model of gene expression evolution would predict that only 50% of genes should show the same direction of expression in specialists relative to generalists (yellow). F) Instead, 96.6% of genes showed the same direction of expression in specialists, suggesting significant parallel expression divergence in specialists (Binomial exact test; $P < 1.0 \times 10^{-16}$). Consistent with incompatible regulatory mechanisms underlying parallel expression in specialists, 45 of these genes were misregulated in specialist F1 hybrids, including G) *sypl1* and H) *scn4a* which also showed expression levels outside the range of all other Caribbean populations examined.

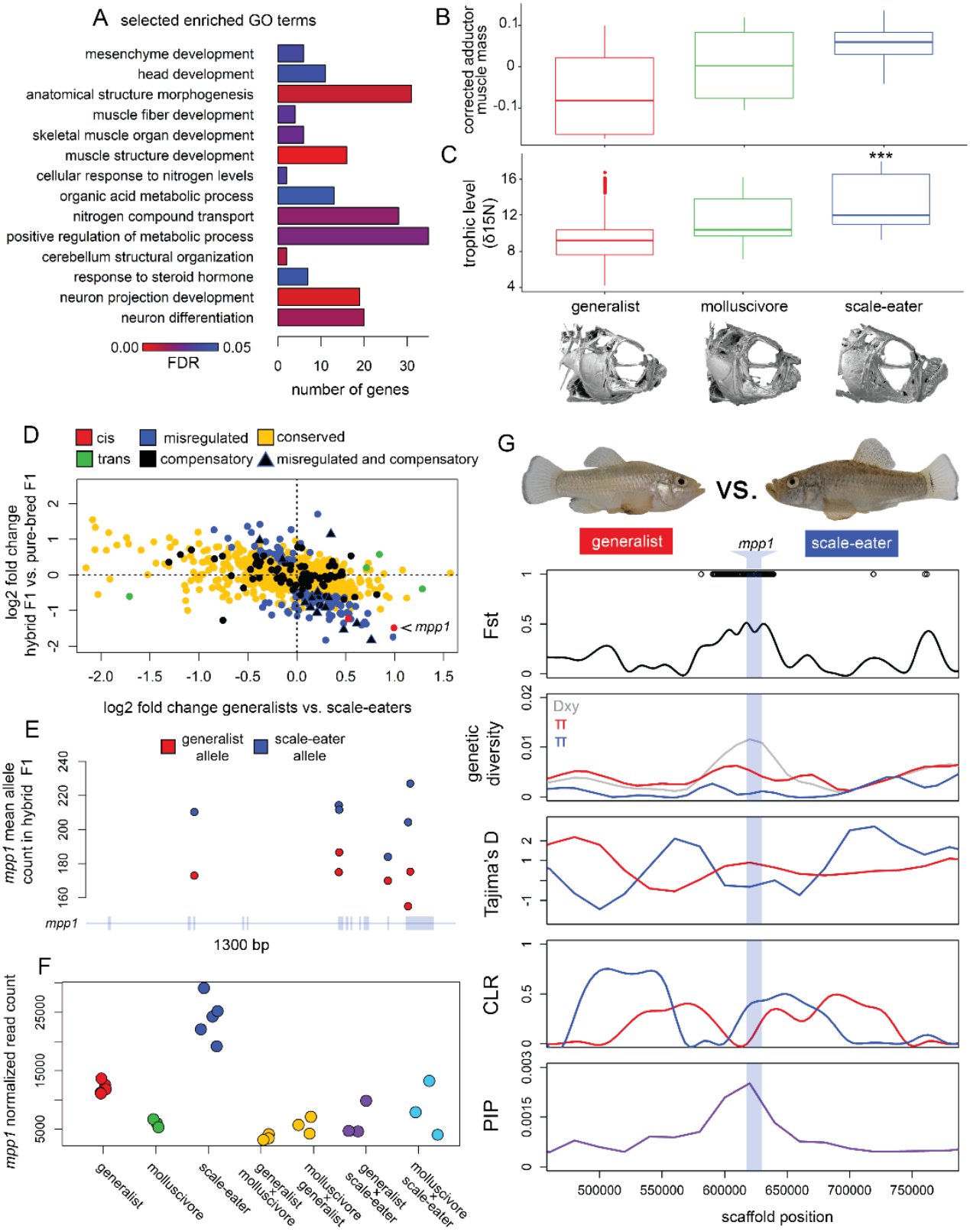
A selected enriched GO terms
mesenchyme development
head development
anatomical structure morphogenesis
muscle fiber development
skeletal muscle organ development
muscle structure development
cellular response to nitrogen levels
organic acid metabolic process
nitrogen compound transport
positive regulation of metabolic process
cerebellum structural organization
response to steroid hormone
neuron projection development
neuron differentiation
0.00 FDR 0.05
0 10 20 30
number of genes
B
corrected adductor muscle mass
0.1
0
-0.1
C
trophic level (δ15N)
16
12
8
4
***
generalist
molluscivore
scale-eater
D
cis
misregulated
conserved
trans
compensatory
misregulated and compensatory
log2 fold change hybrid F1 vs. pure-bred F1
mpp1
log2 fold change generalists vs. scale-eaters
E
generalist allele
scale-eater allele
mpp1 mean allele count in hybrid F1
mpp1
1300 bp
F
mpp1 normalized read count
generalist
molluscivore
scale-eater
generalist x molluscivore
generalist x scale-eater
molluscivore x scale-eater
G
vs.
generalist
mpp1
scale-eater
Fst
genetic diversity
Dxy
π
π
Tajima's D
CLR
PIP
500000 550000 600000 650000 700000 750000
scaffold position

Figure 4.4. Ecological divergence causes hybrid gene misregulation.

A) 14 selected gene ontology (GO) terms relevant to trophic specialization were significantly enriched for the set of 125 genes in highly differentiated genomic regions that showed differential expression between species and misregulation in F1 hybrids. Consistent with muscle development and nitrogen metabolism enrichment, B) adductor mandibulae muscle mass tends to be larger in specialists and C) stable nitrogen isotope ratios (δ15N) are significantly higher in scale-eaters, indicating that they occupy a higher trophic level (Tukey post-hoc test: $P <$ 0.001***). D) The gene *mpp1* is controlled by *cis*-regulatory divergence as shown by E) allele specific expression in F1 hybrids and F) differential expression between Crescent Pond generalists vs. scale-eaters and misregulation in their F1 hybrids. G) The gene *mpp1* (light blue band) is near 170 SNPs fixed between Crescent Pond generalists vs. scale-eaters (black points), shows high absolute divergence between species (D_{xy}), low within-species diversity (π), signatures of a hard selective sweep (Tajima's D and SweeD composite likelihood ratio (CLR)), and is significantly associated with oral jaw length (PIP; GEMMA genome-wide association mapping).

CHAPTER 5: CONSPICUOUS CANDIDATE ALLELES POINT TO *CIS*-REGULATORY DIVERGENCE UNDERLYING RAPIDLY EVOLVING CRANIOFACIAL PHENOTYPES

Introduction

Craniofacial anomalies account for approximately one-third of all birth defects (Gorlin et al. 1990). These include jaw deformities, oral clefts, defects in the ossification of facial or cranial bones, and facial asymmetries. Much of what is known about the developmental genetic basis of craniofacial morphology and function comes from mutagenesis screens and loss of function experiments in model organisms (Hall 2009). These types of studies have been critical to identifying genes essential for craniofacial development and alleles underlying monogenic disease conditions that exhibit Mendelian inheritance. However, screens are biased to detect alleles within protein-coding regions that severely disrupt gene function and are likely to cause lethality at early developmental stages (Nguyen and Tian 2008; Hall 2009). Furthermore, it is now understood that much of the natural and clinical variation in complex traits like craniofacial morphology results from interactions among hundreds to thousands of loci across the genome (Boyle et al. 2017; Sella et al. 2019). Genome-wide association studies (GWAS) have shown that the vast majority of genetic variants affecting complex traits and diseases are within non-coding regions, highlighting the importance of gene regulation influencing trait variation (Hindorff et al. 2009; Maurano et al. 2012; Schaub et al. 2012). Thus, complementary approaches to mutagenesis screens in model organisms are necessary to identify genes that influence craniofacial phenotypes at later stages in development though changes in gene regulation rather than gene function.

One such approach is to harness naturally occurring genetic variation between 'evolutionary mutants' – closely related species exhibiting divergent phenotypes that mimic human disease phenotypes (Albertson et al. 2008). Several fish systems have been particularly useful as models for craniofacial developmental disorders because closely related species are often distinguished by differences in morphological traits important for trophic niche specialization, such as the shape and dynamics of jaws and pharyngeal elements (Albertson et al. 2008; Schartl 2014; Powder and Albertson 2016). The process of identifying candidate genes and validating their effect on phenotypic divergence in evolutionary mutants typically involves population genomic analyses, gene expression analyses, GWAS, and functional validation experiments (Bono et al. 2015; Kratochwil and Meyer 2015). Using a combination of these approaches, research in fish systems has shown that the evolution of adaptive craniofacial traits often involve orthologs of genes implicated in human disorders (Albertson et al. 2005; Helms et al. 2005; Roberts et al. 2011; Ahi et al. 2014; Cleves et al. 2014; Lencer et al. 2017; Erickson et al. 2018; Gross and Powers 2018; Martin et al. 2019). Therefore, candidate genes identified in evolutionary mutant models that have orthologs with uncharacterized functions in humans warrant further study into their relationship with development and disease.

Advances in next generation sequencing technologies alongside substantial reductions in the cost of sequencing have made it possible to sequence the genomes of hundreds of individuals and identify millions of single nucleotide polymorphisms (SNPs) and structural variants (SVs) segregating between closely related species. Measuring relative and absolute genetic differentiation (estimated as *Fst* and *Dxy*) between species can reveal diverged regions of the genome that may influence trait development, but these statistics alone are insufficient to identify genetic mechanisms underlying evolutionary mutant phenotypes (Nachman and Payseur 2012;

Cruickshank and Hahn 2014). RNA sequencing across multiple developmental stages and tissue types can provide further evidence that differentiated regions influence phenotypic divergence if genes near genetic variants are differentially expressed between species (Whiteley et al. 2010; Poelstra et al. 2014; McGirr and Martin 2018; Verta and Jones 2019). However, this assumes that linked genetic variation within *cis*-acting regulatory elements affects proximal gene expression levels, and does not rule out the possibility of unlinked *trans*-acting regulatory variation binding regulatory regions to influence expression levels (Wittkopp and Kalay 2011; Signor and Nuzhdin 2018). Determining whether a candidate gene is differentially expressed due to *cis*- or *trans*-regulatory divergence is important to identify putatively causal alleles that can be further validated by genome editing or transgenesis experiments.

It is possible to use RNAseq to identify mechanisms of gene expression divergence between parental species by bringing *cis* elements from both parents together in the same *trans* environment in F1 hybrids and quantifying allele specific expression (ASE) of parental alleles at heterozygous sites (Cowles et al. 2002; Wittkopp et al. 2004; Signor and Nuzhdin 2018). ASE occurs when a heterozygous allele within a coding region that is alternatively homozygous in two parental species shows biased expression in F1 hybrids. *Cis*-regulatory divergence is expected when a gene is differentially expressed between species and shows ASE in F1 hybrids; whereas *trans*-regulatory divergence is expected when a gene is differentially expressed and does not show ASE (Wittkopp et al. 2004; Davidson and Balakrishnan 2016; Signor and Nuzhdin 2018). Thus, genes showing signs of *cis*-regulatory divergence that are near differentiated regions of the genome make promising candidates for causal variation underlying evolutionary mutant phenotypes, especially when the same genes show high genetic differentiation between species and are implicated by GWAS. Together, these strategies can target candidate variation

with nucleotide level resolution and provide a framework to prioritize variants for functional validation experiments.

Here, we combine whole-genome resequencing, RNAseq, and F1 hybrid allele specific expression analyses to identify candidate *cis*-acting genetic variation influencing rapidly evolving craniofacial phenotypes within an adaptive radiation of *Cyprinodon* pupfishes on San Salvador Island, Bahamas (Fig. 5.1). This sympatric radiation consists of a dietary generalist species (*C. variegatus*) and two endemic specialist species adapted to novel trophic niches – a molluscivore (*C. brontotheroides*) and a scale-eater (*C. desquamator*; (Martin and Wainwright 2013a)). Nearly all forty-nine pupfish species in the genus *Cyprinodon* distributed across North America and the Caribbean are dietary generalists with similar craniofacial morphology that is used for consuming algae and small invertebrates (Fig. 5.1A (Martin and Wainwright 2011, 2013b)). The molluscivore evolved short, thick oral jaws stabilized by a nearly immobile maxilla allowing it to specialize on hard-shelled prey including ostracods and gastropods (Fig. 5.1B). This morphology results in a larger in-lever to out-lever ratio compared with generalists, increasing mechanical advantage for strong biting (Hernandez et al. 2018). The molluscivore is also characterized by a prominent maxillary anteriodorsal protrusion that may be used as a wedge for extracting snails from their shells (Martin et al. 2017). The scale-eater is a predator that evolved to bite scales and protein-rich mucus removed from other pupfish species during rapid feeding strikes (Fig. 5.1C (St. John et al. 2020)). Scale-eaters have greatly enlarged oral jaws, larger adductor mandibulae muscles, darker breeding coloration, and a more elongated body compared with the generalist and molluscivore species (Martin and Wainwright 2013a). Exceptional craniofacial divergence despite extremely recent divergence times and low genetic differentiation between molluscivores and scale-eaters make this system a compelling

evolutionary model for human craniofacial developmental disorders. These trophic specialist species rapidly diverged from an ancestral generalist phenotype within the last 10-15k years (Turner et al. 2008; Martin and Feinstein 2014). Molluscivores and scale-eaters readily hybridize in the laboratory to produce fertile F1 offspring with approximately intermediate craniofacial phenotypes between the parents and no obvious sex ratio distortion (Martin and Wainwright 2013b; Martin and Feinstein 2014). These species show evidence of pre-mating isolation in the laboratory (West and Kodric-Brown 2015) and are genetically differentiated in sympatry (genome-wide mean *Fst* = 0.14 across 12 million SNPs; (McGirr and Martin 2017b)).

We previously identified 31 genomic regions (20 kb) containing SNPs fixed between species (*Fst* = 1), showed signs of a hard selective sweep, and were significantly associated with oral jaw size using multiple genome-wide association mapping approaches (McGirr and Martin 2017b). A subset of these fixed SNPs fell within significant QTL explaining 15% of variation in oral jaw size and were near genes annotated for effects on skeletal system development (Martin et al. 2017). Here we use complementary approaches to identify candidate causal variants putatively influencing craniofacial divergence by 1) incorporating transcriptomic data from 122 individuals sampled at three developmental stages (McGirr and Martin 2018, 2019c), 2) applying genome divergence scans to a much larger sample of whole genomes from San Salvador Island and surrounding Caribbean outgroup populations (increasing n = 37 to 258) aligned to a new high-quality *de novo* genome assembly (Richards et al. in prep.), 3) identifying structural variation fixed between species for the first time in this system, and 4) inferring *cis* and *trans* regulatory mechanisms underlying gene expression divergence between species using 12 F1 hybrid transcriptomes. We identified two genes showing *cis*-regulatory divergence that were near just one fixed variant each: a deletion upstream of a gene known to influence skeletal

development (*dync2li1*) and a SNP downstream of a novel candidate gene (*pycr3*). Our results highlight the utility of using these closely related species replicated across isolated lake populations as an evolutionary model for craniofacial development and provide highly promising candidate variants for future functional validation experiments.

Methods

Identifying genomic variation fixed between specialists

In order to identify SNPs fixed between species, we analyzed whole genome resequencing samples for 258 individuals from across the Caribbean (median coverage = 8×; (Richards et al. in prep.)). Briefly, 114 pupfishes from 15 isolated hypersaline lakes and one estuary on San Salvador Island were collected using hand and seine nets between 2011 and 2018. This included 33 generalists 46 molluscivores, and 35 scale-eaters. Eight of these individuals were bred to generate F1 offspring sampled for RNA sequencing (Table E1.2). This dataset also included 140 outgroup generalist pupfishes from across the Caribbean and North America, including two individuals belonging to the pupfish radiation in Lake Chichancanab, Mexico, and two individuals from the most closely related outgroups to *Cyprinodon* (*Megupsilon aporus* and *Cualac tessellatus* (Echelle et al. 2005)). Libraries for 150PE Illumina sequencing were generated from DNA extracted from muscle tissue and the resulting reads were mapped to the *C. brontotheroides* reference genome (v 1.0; total sequence length = 1,162,855,435 bp; number of scaffolds = 15,698, scaffold N50, = 32,000,000 bp; L50 = 15 scaffolds; Richards et al. in prep.). Variants were called using the HaplotypeCaller function of the Genome Analysis Toolkit (GATK v 3.5 (DePristo et al. 2011)) and filtered to include SNPs with a minor allele frequency

above 0.05, genotype quality above 20, and sites with greater than 50% genotyping rate across all individuals, resulting in 9.3 million SNPs.

Measuring relative genetic differentiation (*Fst*) between species can point to regions of the genome containing variation affecting divergent phenotypes (Jones et al. 2012; Poelstra et al. 2014; Lamichhaney et al. 2015). However, *Fst* is dependent on the many potential forces acting to reduce within-population nucleotide diversity, including selective sweeps, purifying selection, background selection, and low recombination rates (Noor and Bennett 2009; Cruickshank and Hahn 2014). Measuring between-population divergence (*Dxy*) can help distinguish between these possibilities because nucleotide divergence between species increases at loci under different selective regimes (Nachman and Payseur 2012; Cruickshank and Hahn 2014; Irwin et al. 2016). We measured *Fst* between species with vcftools (v. 0.1.15; weir-fst-pop function) and identified fixed SNPs (*Fst* = 1). We also measured *Fst* and *Dxy* in 10 kb windows using the python script popGenWindows.py created by Simon Martin (github.com/simonhmartin/genomics_general; (Martin et al. 2013)).

We identified structural variation (insertions, deletions, inversions, translocations, and copy number variants) fixed between specialist species with DELLY (v 0.8.1; (Rausch et al. 2012)). Unlike GATK HaplotypeCaller which is limited to identifying structural variants smaller than half the length of read size (DePristo et al. 2011), DELLY can identify small variants in addition to variants larger than 300 bp using paired-end clustering and split read analysis. We used DELLY to identify structural variants across eight whole genomes from the breeding pairs used to generate F1 hybrid RNA samples (Four scale-eaters from two lake populations and four molluscivores from the same two lake populations; Table E1.2). First, we trimmed reads using Trim Galore (v. 4.4, Babraham Bioinformatics), aligned them to the *C. brontotheroides* reference

genome with the Burrows-Wheeler Alignment Tool (v 0.7.12; (Li and Durbin 2011), and removed duplicate reads from the resulting .bam files with Picard MarkDuplicates (broadinstitute.github.io/picard). Second, we called variants with DELLY by comparing an individual of one species with all individuals of the other species, resulting in eight variant call files. Third, we identified structural variants fixed between species that were shared across all eight files, in which all molluscivores showed the reference allele and all scale-eaters showed the same alternate allele.

Transcriptomic sequencing, alignment, and variant discovery

Our transcriptomic dataset included 50 libraries from 122 individuals sampled across three early developmental stages (Table E1.2; (McGirr and Martin 2018, 2019c)). Breeding pairs used to generate F1 hybrids and purebred F1 offspring were collected from three hypersaline lakes on San Salvador Island: Crescent Pond, Osprey Lake, and Little Lake. For purebred crosses, we collected F1 embryos from breeding tanks containing multiple breeding pairs from a single lake population. For F1 hybrid samples, we crossed a single individual of one species with a single individual of another species from the same lake population.

RNA was extracted from samples collected two days after fertilization (2 dpf) eight days after fertilization (8 dpf), and 17-20 days after fertilization (20 dpf) using RNeasy Mini Kits (Qiagen catalog #74104). For samples collected at 2 dpf, we pooled 5 embryos together and pulverized them in a 1.5 ml Eppendorf tube using a plastic pestle washed with RNase Away (Molecular BioProducts). We used the same extraction method for samples collected at 8 dpf but did not pool larvae and prepared a library for each individual separately. We dissected samples collected at 20 dpf to isolate tissues from the anterior craniofacial region containing the dentary, angular articular, maxilla, premaxilla, palatine, and associated craniofacial connective tissues

using fine-tipped tweezers washed with RNase AWAY. The earlier developmental stages are described as stage 23 (2 dpf) and 34 (8 dpf) in a recent embryonic staging series of *C. variegatus* (Lencer and McCune 2018). The 2 dpf stage is comparable to the Early Pharyngula Period of zebrafish, when multipotent neural crest cells have begun migrating to pharyngeal arches that will form the oral jaws and most other craniofacial structures (Schilling and Kimmel 1994; Furutani-Seiki and Wittbrodt 2004; Lencer et al. 2017). Embryos usually hatch six to ten days post fertilization, with similar variation in hatch times among species (Lencer et al. 2017; McGirr and Martin 2018). While some cranial elements are ossified prior to hatching, the skull is largely cartilaginous at 8 dpf and ossified by 20 dpf (Lencer and McCune 2018). All samples were reared in breeding tanks at 25–27°C, 10–15 ppt salinity, pH 8.3, and fed a mix of commercial pellet foods and frozen foods.

Methods for total mRNA sequencing were previously described (McGirr and Martin 2018, 2019c). Briefly, 2 dpf and 8 dpf libraries were prepared using TruSeq stranded mRNA kits and sequenced on 3 lanes of Illumina 150 PE Hiseq4000 at the Vincent J. Coates Genomic Sequencing Center (McGirr and Martin 2019c). All 20 dpf libraries were prepared using the KAPA stranded mRNA-seq kit (KAPA Biosystems 2016) at the High Throughput Genomic Sequencing Facility at UNC Chapel Hill and sequenced on one lane of Illumina 150PE Hiseq4000 (McGirr and Martin 2018). We filtered raw reads using Trim Galore (v. 4.4, Babraham Bioinformatics) to remove Illumina adaptors and low-quality reads (mean Phred score < 20) and mapped 122,090,823 filtered reads to the *C. brontotheroides* reference genome (Richards et al. in prep.) using the RNAseq aligner STAR with default parameters (v. 2.5 (Dobin et al. 2013b)). We assessed mapping and read quality using MultiQC (Ewels et al. 2016) and quantified the number of duplicate reads and the median percent GC content of mapped reads for

each sample using RSeQC (Wang et al. 2012). Although all reads were mapped to a molluscivore reference genome, we did not find a significant difference between species in the proportion of reads uniquely mapped with STAR (Fig. E2.1 A; Student's t-test, $P = 0.061$). Additionally, we did not find a difference between species in the proportion of multimapped reads, GC content of reads, or number of duplicate reads (Fig. E2.1 B-D; Student's t-test, $P > 0.05$).

We used GATK HaplotypeCaller function to call SNPs across 50 quality filtered transcriptomes. We refined SNPs using the allele-specific software WASP (v. 0.3.3) to correct for potential mapping biases that would influence tests of allele-specific expression (ASE; (Van De Geijn et al. 2015)). WASP identified reads that overlapped SNPs in the initial .bam files and re-mapped those reads after swapping the genotype for the alternate allele. Reads that failed to map to exactly the same location were discarded. We re-mapped unbiased reads to create our final .bam files used for differential expression analyses. Finally, we re-called SNPs using unbiased .bam files for allele specific expression analyses.

Differential expression analyses

We used the featureCounts function of the Rsubread package (Liao et al. 2014) requiring paired-end and reverse stranded options to generate read counts across 19,304 genes and 156,743 exons annotated for the *C. brontotheroides* reference genome (Richards et al. in prep.). We used DESeq2 (v. 3.5 (Love et al. 2014)) to normalize raw read counts for library size and perform principal component analyses, and identify differentially expressed genes. DESeq2 fits negative binomial generalized linear models for each gene across samples to test the null hypothesis that the fold change in gene expression between two groups is zero. Significant differential expression between groups was determined with Wald tests by comparing normalized posterior

log fold change estimates and correcting for multiple testing using the Benjamini–Hochberg procedure with a false discovery rate of 0.01 (Benjamini and Hochberg 1995).

We constructed a DESeqDataSet object in R using a multi-factor design that accounted for variance in F1 read counts influenced by parental population origin and sequencing date (design = ~sequencing_date + parental_breeding_pair_populations). Next, we used a variance stabilizing transformation on normalized counts and performed a principal component analysis to visualize the major axes of variation in 2 dpf, 8 dpf, and 20 dpf samples (Fig. E2.2). We contrasted gene expression in pairwise comparisons between species grouped by developmental stage (sample sizes for comparisons (molluscivores vs. scale-eaters): 2 dpf = 6 vs. 6, 8 dpf = 8 vs. 10, 20 dpf = 6 vs. 2). We used plyranges (v. 1.6.5; (Lee et al. 2019)) to determine if genetic variation fixed between species fell within 10 kb of significantly differentially expressed genes (> 10 kb from the start of the first exon and <10 kb from the end of the last exon).

Allele specific expression analyses

It is possible to identify mechanisms of gene expression divergence between parental species by bringing *cis* elements from both parents together in the same *trans* environment in F1 hybrids and quantifying allele specific expression (ASE) of parental alleles at heterozygous sites (Fig. 5.5; (Cowles et al. 2002; Wittkopp et al. 2004)). A gene that is differentially expressed between parental species that also shows allele specific expression biased toward one parental allele is expected to result from *cis*-regulatory divergence. A gene that is differentially expressed between parental species that does not show ASE in F1 hybrids is expected to result from *trans*-regulatory divergence. After identifying genes differentially expressed between species that were also near fixed variants, we wanted to test whether those genes showed signs of *cis*-regulatory

divergence. This would strongly implicate that fixed variation contributed to expression divergence between species.

Our SNP dataset included every parent used to generate F1 hybrids between populations (n = 8). We used the GATK VariantsToTable function (DePristo et al. 2011) to output genotypes across 9.3 million SNPs for each parent and overlapped these sites with the variant sites identified in F1 hybrid transcriptomes. We used python scripts (github.com/joemcgirr/fishfASE) to identify SNPs that were alternatively homozygous in breeding pairs and heterozygous in their F1 offspring. We counted reads across heterozygous sites using ASEReadCounter (-minDepth 20 --minMappingQuality 10 --minBaseQuality 20 -drf DuplicateRead) and matched read counts to maternal and paternal alleles. For genes that were differentially expressed between species that were near fixed variants, we identified significant ASE using a beta-binomial test comparing the maternal and paternal counts at each gene with the R package MBASED (Mayba et al. 2014). For each F1 hybrid sample, we performed a 1-sample analysis with MBASED using default parameters run for 1,000,000 simulations to determine whether genes showed significant ASE in hybrids ($P < 0.05$).

For genes within 10 kb of variants fixed between species, we inferred *cis*-regulatory divergence if a gene was significantly differentially expressed between species (DeSeq2 $P < 0.01$) and showed significant ASE in all F1 hybrids from a cross at the same developmental stage (MBASED; $P < 0.05$). We inferred *trans*-regulatory divergence if a gene was significantly differentially expressed between species (DeSeq2 $P < 0.01$) and did not show significant ASE in all F1 hybrids (MBASED; $P > 0.05$; Fig. 5.5). We required that genes had at least two informative SNPs with ≥10× coverage to assign *cis*- or *trans*- regulatory divergence.

We most likely underestimated the number of fixed variants acting as *cis*-regulatory alleles influencing expression divergence between species. First, variants could affect expression at a developmental stage not included in our sampling. Second, because our approach to identify regulatory mechanisms underlying expression divergence relies on F1 hybrid expression, the advantage of having low genetic variation between species is counterbalanced by the disadvantage of limited heterozygosity within coding regions that provide informative sites to estimate allele-specific expression. Out of twelve genes near fixed variants that were differentially expressed between species, only five contained more than one heterozygous informative site to assign *cis*- or *trans*- regulatory divergence. Third, some of the genes that we classified as *trans*-regulated showed low overall levels of expression, reducing our power to detect significant differences in expression levels of parental alleles. Fourth, we required that genes show allele-specific expression across the entire coding region to assign *cis*-regulatory divergence, which ignored the possibility of alleles affecting the expression of specific transcript isoforms. For these reasons, our estimation of *cis*-regulatory divergence was highly conservative but still provided promising candidate genes for future study.

Gene ontology enrichment and transcription factor binding site analyses

We performed gene ontology (GO) enrichment analyses for genes near candidate adaptive variants using ShinyGo v.0.51 (Ge and Jung 2018, unpublished, biorXiv doi.org/10.1101/315150). The *C. brontotheroides* reference genome was annotated using MAKER, a genome annotation pipeline that annotates genes, transcripts, and proteins (Cantarel et al 2008). Gene symbols for orthologs identified by this pipeline largely match human gene symbols. Thus, we searched for enrichment across biological process ontologies curated for human gene functions.

We searched the JASPAR database (Fornes et al. 2019) to identify whether fixed variation near genes showing *cis*-regulatory divergence altered potential transcription factor binding sites. We generated fasta sequences for the molluscivore containing the variant site and 20 bp on either end of the site and searched across all 1011 predicted vertebrate binding motifs in the database using a 95% relative profile score threshold. We then preformed the same analysis for scale-eater fasta sequences containing the alternate allele.

Genotyping fixed variants

In order to confirm the genotypes of putative *cis*-acting variants, we performed Sanger sequencing on four additional individuals that were not included in our whole-genome dataset. We extracted DNA from muscle tissue using DNeasy Blood and Tissue kits (Qiagen, Inc.) from two molluscivores and two scale-eaters (one sample from Crescent Pond and one from Osprey Lake for both species). We designed primers targeting the regions containing variation fixed between species near the two genes showing evidence for *cis*-regulatory divergence (*pycr3* and *dync2li1*) using the NCBI primer design tool (Ye et al. 2012). We designed primers targeting a 446 bp region containing the SNP fixed between species (scaffold: HiC_scaffold_16 ; position: 1,0043,644) that was 1,808 bp downstream of *pycr3* (forward: 5′-ACCATTCCAGAAGACAAAAAGCG-3′; reverse: 5′-GGCCCTATATATGGGATGCACAA-3′). Sequences were amplified with PCR using New England BioLabs *Taq* polymerase (no. 0141705) and dNTP solution (no. 0861609) and Sanger sequencing was performed at Eton Bioscience Inc. (Research Triangle Park, North Carolina). Aligning the resulting sequences using the Clustal Omega Multiple Sequence Alignment Tool (Madeira et al. 2019)) confirmed the A-to-C transversion in scale-eaters (Fig. 5.8). We designed two additional primer sets targeting the deletion region near *dync2li1* (scaffold: HiC_scaffold_43 ; position: 26,792,380-26,792,471).

While both primer sets amplified the sequence in molluscivore samples (not shown), we were unable to amplify this region in scale-eaters, potentially due to high polymorphism in this region.

Results

Few fixed variants between young species showing drastic craniofacial divergence

We analyzed whole genome resequencing samples for 258 *Cyprinodon* pupfishes (median coverage = 8×; (Richards et al. in prep.)). This included 114 individuals from multiple isolated lake populations on San Salvador Island (33 generalists, 46 molluscivores, and 35 scale-eaters) and 140 outgroup generalist pupfishes from across the Caribbean and North America. Libraries for 150PE Illumina sequencing were generated from DNA extracted from muscle tissue and the resulting reads were mapped to the *C. brontotheroides* reference genome (v 1.0; total sequence length = 1,162,855,435 bp; number of scaffolds = 15,698, scaffold N50, = 32,000,000 bp; L50 = 15 scaffolds; Richards et al. in prep.). Variants were called using the Genome Analysis Toolkit (GATK v 3.5 (DePristo et al. 2011)) and filtered to include SNPs with a minor allele frequency above 0.05, genotype quality above 20, and sites with greater than 50% genotyping rate across all individuals.

Out of 9.3 million SNPs identified in our dataset, we found a mere 157 SNPs fixed between molluscivore and scale-eater specialist species showing *Fst* = 1 (Fig. 5.2A; mean genome-wide *Fst* = 0.076). Of these 157 variants, 46 were within 10 kb of 27 genes and none were in coding regions. These 27 genes were enriched for 27 biological processes, including several ontologies describing neuronal development and activity of cell types within bone marrow (Fig. 5.2B; Table E1.1).

Structural variants (including insertions, deletions, inversions, translocations, and copy number variants) have been traditionally difficult to detect in non-model systems and ignored by many early whole-genome comparative studies (Stapley et al. 2010; Ho et al. 2019; Wellenreuther et al. 2019). We identified 80,012 structural variants across eight molluscivore and scale-eater individuals using a method that calls variants based on combined evidence from paired-end clustering and split read analysis (Rausch et al. 2012). Just 87 structural variants were fixed between species and, strikingly, all of these variants were deletions fixed in scale-eaters. These deletions ranged in size between 55 bp and 4,703 bp (Fig. 5.2C). Of these, 34 fixed deletions were near 34 genes (Table E1.1). Only a single fixed deletion (1,256 bp) was found within a protein coding region, spanning the entire fifth exon of *gpa33* (Fig. 5.3). The 34 genes near fixed deletions were enriched for 36 biological processes, including ontologies describing bone development, mesenchyme development, fibroblast growth, and digestive tract development (Fig 2D).

Including SNPs and deletions, we found a total of 80 fixed variants within 10 kb of 59 genes (Table E1.1). Encouragingly, 41 of these genes (70%) also showed high between population nucleotide divergence ($Dxy > 0.0083$ (genome-wide 90^{th} percentile)), strengthening evidence for adaptive divergence at these loci. There are likely many alleles contributing to craniofacial divergence that are segregating between populations of molluscivores and scale-eaters. However, variants with larger effect sizes are predicted to fix faster than variants with smaller effects, especially given short divergence times (Griswold 2006; Yeaman and Whitlock 2011). Thus, these 80 fixed variants provided a promising starting point to identify causal alleles influencing craniofacial phenotype.

Genes near fixed variants are differentially expressed throughout development

All but one of the 80 variants fixed between species were in non-coding regions, suggesting that they may affect species-specific phenotypes through regulation of nearby genes. In order to identify patterns of gene expression divergence between species, we combined two previous transcriptomic datasets spanning three developmental stages and three San Salvador Island lake populations (McGirr and Martin 2018, 2019c). F1 offspring were sampled at 2 days post-fertilization (dpf), 8 dpf, and 20 dpf. RNA was extracted from whole body tissue at 2 dpf and 8 dpf; whereas 20 dpf samples were dissected to only extract RNA from craniofacial tissues (Table E1.2). We used DEseq2 (Love et al. 2014) to contrast gene expression in pairwise comparisons between species grouped by developmental stage (sample sizes for comparisons (molluscivores vs. scale-eaters): 2 dpf = 6 vs. 6, 8 dpf = 8 vs. 10, 20 dpf = 6 vs. 2).

Out of 19,304 genes annotated for the *C. brontotheroides* reference genome, we found 770 (5.93%) significantly differentially expressed at 2 dpf, 1,277 (9.48%) at 8 dpf, and 312 (2.50%) at 20 dpf (Fig. 5.4A-D). The lower number of genes differentially expressed at 20 dpf likely reflects reduced power to detect expression differences due to the small scale-eater sample size. Nonetheless, we found four genes differentially expressed throughout development at all three stages (*filip1*, *c1galt1*, *klhl24*, and *oit3*) and 248 genes were differentially expressed during two of the three stages examined. Of the 59 genes near SNPs or deletions fixed between species, we found 12 differentially expressed during at least one developmental stage (Table 5.1; Fig. 5.4E). Two of these genes (*dync2li1* and *pycr3*) were differentially expressed at 2 dpf and 8 dpf.

Fixed variants near genes showing cis-regulatory divergence

In order to determine whether the 12 genes near fixed variants showed differential expression due to *cis*- or *trans*-regulatory divergence, we analyzed expression patterns across 12 F1 hybrid transcriptomes generated from crosses between molluscivores and scale-eaters. The parents used as breeding pairs for these crosses were included in the genomic SNP dataset, allowing us to identify sites that were alternatively homozygous in parents and heterozygous in their F1 hybrids. We measured allele-specific expression (ASE) for genes containing heterozygous sites to identify mechanisms of regulatory divergence (Cowles et al. 2002; Wittkopp et al. 2004). We identified significant ASE using a beta-binomial test comparing the maternal and paternal counts at each gene with the R package MBASED (Mayba et al. 2014). We inferred *cis*-regulatory divergence if a gene was significantly differentially expressed between species and showed significant ASE in all F1 hybrids from a cross at the same developmental stage (Fig. 5.5A). We inferred *trans*-regulatory divergence if a gene was significantly differentially expressed between species and did not show significant ASE in all F1 hybrids (Fig. 5.5B).

Of the 12 genes differentially expressed near fixed variants, five contained at least two informative heterozygous sites that could be used to measure ASE (Fig. 5.6 and 5.7). The same two genes that were differentially expressed at 2 dpf and 8 dpf (*dync2li1* and *pycr3*) also showed significant allele specific expression in F1 hybrids at both developmental stages (Fig. 5.6A and B; MBASED $P < 0.05$). This provided strong evidence that differential regulation of these genes may have been caused by nearby fixed variation within putative *cis*-regulatory elements. The three other genes with informative sites (*eef1d, washc5,* and *pxk*) did not show significant ASE,

suggesting that *trans*-acting variation may have influenced expression divergence between species for these genes (Fig. 5.7A-C).

The two genes showing *cis*-regulatory divergence were near just one fixed variant each: a 91 bp deletion located 7,384 bp upstream of *dync2li1* and an A-to-C transversion 1,808 bp downstream of *pycr3* (Fig. 5.6). The next closest fixed variants were separated by greater than 600 kb and 31 kb, respectively. We searched the JASPAR database (Fornes et al. 2019) to identify potential transcription factor binding sites that could be altered by these candidate *cis*-acting variants. The 91 bp deletion near *dync2li1* contained binding motifs corresponding to seven transcription factors (*nfia*, *nfix*, *nfic*, *znf384*, *hoxa5*, *gata1*, *myb*; Table E1.3). Two binding motifs spanned the *pycr3* SNP region (*gata2*, *mzf1*), one of which, *mfz1*, was altered by the alternate allele in scale-eaters. The scale-eater allele created a new potential binding motif matching the transcription factor *plagl2*. Sanger sequencing confirmed the A-to-C transversion near *pycr3* in four additional individuals not included in the whole-genome resequencing dataset (Fig. 5.8).

Discussion

Understanding the developmental genetic basis of complex traits by investigating natural variation among closely related species is a powerful complementary approach to traditional genetic screens in model systems. The San Salvador Island *Cyprinodon* pupfish system is a useful evolutionary model for understanding the genetic basis of craniofacial defects and natural diversity given extensive morphological divergence between these young species (Fig. 5.1). We found just 244 genetic variants fixed between molluscivores and scale-eaters across 9.3 million SNPs and 80,012 structural variants (Fig. 5.2A and C). Almost all variants were in non-coding

regions, with the exception of an exon-spanning deletion (Fig. 5.3). In support of these variants affecting divergent adaptive phenotypes, 80 variants were near 59 genes that were enriched for developmental functions related to divergent specialist traits (Fig. 5.2B and D). Furthermore, twelve of these genes were highly differentially expressed between species across three developmental stages (Fig. 5.4E). By measuring allele-specific expression in F1 hybrids from multiple crosses between species, we found two variants strongly implicated as *cis*-regulatory alleles affecting expression divergence between species: a fixed deletion near *dync2li1* and a fixed SNP near *pycr*3 (Fig. 5.6).

Fixed genetic variation underlying trophic specialization

In a previous analysis of SNPs from a smaller whole genome dataset, *dync2li1* was one of 30 candidate genes that showed signs of a hard selective sweep and was significantly associated with variation in jaw size between molluscivores and scale-eaters using multiple genome-wide association mapping approaches (McGirr and Martin 2017b). Here we show that a fixed deletion near *dync2li1* may influence expression divergence between species through *cis*-acting regulatory mechanisms. This gene (dynein cytoplasmic 2 light intermediate chain 1) is known to influence skeletal morphology in humans (Kessler et al. 2015; Taylor et al. 2015; Niceta et al. 2018). It is a component of the cytoplasmic dynein 2 complex which is important for intraflagellar transport – the movement of protein particles along the length of eukaryotic cilia (Cole 2003; Pfister et al. 2006). Due to the vital role that cilia play in the transduction of signals in the *hedgehog* pathway and other pathways important for skeletal development, disruptions in dynein complexes cause a variety of skeletal dysplasias collectively termed skeletal ciliopathies (Huber and Cormier-Daire 2012; Taylor et al. 2015). Mutations in *dync2li1* have been linked with ciliopathies that result from abnormal cilia shape and function including Ellis-van Creveld

syndrome, Jeune syndrome, and short rib polydactyly syndrome (Kessler et al. 2015; Taylor et al. 2015; Niceta et al. 2018). These disorders are characterized by variable craniofacial malformations including micrognathia (small jaw), hypodontia (tooth absence), and cleft palate (Brueton et al. 1990; Ruiz-Perez and Goodship 2009; Taylor et al. 2015). The discovery of *dync2li1* as a candidate gene influencing differences in oral jaw length between molluscivores and scale-eaters suggests that this system is particularly well-suited as an evolutionary mutant model for clinical phenotypes involving jaw size, such as micrognathia and macrognathia.

We also identified a fixed SNP near the gene *pycr3* (pyrroline-5-carboxylate reductase 3; also denoted *pycrl*) which showed *cis*-regulatory divergence. This gene is not currently known to influence craniofacial phenotypes in humans or other model systems. However, one study investigating gene expression divergence between beef and dairy breed bulls found that *pycr3* was one of the most highly differentially expressed genes in skeletal muscle tissues. The authors found nearly a three-fold difference in expression of *pycr3* between the two bull breeds that are primarily characterized by differences in muscle anatomy (Sadkowski et al. 2009). Similarly, expression changes in this gene may influence skeletal muscle development in specialists species, which differ in the size of their adductor mandibulae muscles (Martin and Wainwright 2011; Hernandez et al. 2018). The A-to-C transversion near *pycr3* could influence differences in expression by altering transcription factor binding. We found that the molluscivore allele matches the binding motif of *mzf1* (myeloid zinc finger 1; Fig. 5.8), a transcription factor known to influence cell proliferation (Gaboli et al. 2001), whereas the scale-eater allele alters this motif. This type of binding motif analyses has a high sensitivity (*mzf1* is known to bind this motif) but extremely low selectivity (*mzf1* does not bind nearly every occurrence of this motif, which appears 1,430,540 times in the molluscivore reference genome).

While oral jaw size is the primary axis of phenotypic divergence in the San Salvador pupfish system, adaptation to divergent niches required changes in a suite of morphological and behavioral phenotypes (St John et al. 2019; St. John et al. 2020). The majority of genes differentially expressed between species were found within whole embryo tissues (Fig. 5.4A-D), suggesting we should find candidate genes influencing the development of craniofacial phenotypes and other divergent traits. Of the 244 variants fixed between species, the only coding variant was a 1,256 bp deletion that spanned the fifth exon of *gpa33* (glycoprotein A33), which is expressed exclusively in intestinal epithelium (Fig. 5.3). Knockouts of this gene in mice cause increased hypersensitivity to food allergens and susceptibility to a range of related inflammatory intestinal pathologies (Williams et al. 2015). The gut contents of wild-caught scale-eaters are comprised of 40-51% scales (Martin and Wainwright 2013c). The exon deletion of *gpa33* may play a metabolic role in this unique adaptation that allows scale-eaters to occupy a higher trophic level than molluscivores. Future studies in this system will benefit from sequencing and analyses that target specific tissues and cell types to determine whether candidate variants affect a single phenotype or have pleiotropic effects.

The effectiveness of Cyprinodon pupfishes for identifying candidate cis-regulatory variants

One major advantage of investigating the genetic basis of craniofacial divergence between molluscivores and scale-eaters is the low amount of genetic divergence between species. Species-specific phenotypes are replicated across multiple isolated lake populations that exhibit substantial ongoing gene flow. This has resulted in small regions of the genome showing strong genetic differentiation, with some regions containing just a single variant fixed between species. Furthermore, a previous study found a significant QTL explaining 15% of variation in oral jaw size and three more potential moderate-effect QTL, suggesting that we may expect to

find variants with moderate effects on craniofacial divergence. Thus, we chose to focus our analyses on genes near fixed variation because variants with larger effect sizes are predicted to fix faster than variants with smaller effects, especially given short divergence times (Griswold 2006; Yeaman and Whitlock 2011). The low number of fixed variants dispersed across the genome makes this system relatively unique compared to other systems with similar divergence times (Whiteley et al. 2010; Jones et al. 2012; Martin et al. 2019). Although our approach ignores segregating variation, which likely influences the majority of craniofacial divergence between species, it provides a strategy to identify novel candidate genes like *pycr3* that have not been previously associated with craniofacial development and prioritize such candidates for functional validation experiments.

Conclusions

Overall, our results highlight the utility of the San Salvador pupfish system as an evolutionary mutant model for natural and clinical variation in human craniofacial phenotypes. Similar rapid speciation replicated across many environments can be found in other adaptive radiations (Martin et al. 2019; Martin and Richards 2019), which could also prove useful as evolutionary models for a variety of other human traits. We found that a combination of SNPs and deletions likely contribute to the evolution of highly divergent craniofacial morphology through *cis*-acting effects on the expression of skeletal genes. Future studies will attempt to validate the effect of candidate variation on gene expression and craniofacial development *in vivo.*

Table 5.1. Twelve genes differentially expressed between molluscivores and scale-eaters at 2 days post fertilization (dpf), 8 dpf, and/or 20 dpf.

Twelve genes differentially expressed between molluscivores and scale-eaters at 2 days post fertilization (dpf), 8 dpf, and/or 20 dpf. Differentially expressed genes showing *cis* regulation showed significant allele-specific expression in F1 hybrids (MBASED $P < 0.05$), while genes showing *trans* regulation did not (MBASED $P > 0.05$). Genes with undetermined regulatory mechanisms underlying expression divergence (NA) had fewer than two informative heterozygous sites within the coding region. MNC = mean normalized counts across all samples. LFC = log2 fold change in expression (positive values indicate higher expression in scale-eaters than molluscivores). *P* = adjusted *P*-value for differential expression.

		2 dpf			8 dpf			20 dpf		
gene	mechanism	MNC	LFC	*P*	MNC	LFC	*P*	MNC	LFC	*P*
dync2li1	*cis*	96.09	-0.70	**3.7E-05**	34.05	-1.05	**5.2E-05**	23.83	-1.10	1.2E-01
pycr3	*cis*	221.91	0.49	**2.5E-03**	56.19	1.09	**1.5E-08**	38.16	0.13	8.9E-01
eef1d	*trans*	1984.23	0.18	1.3E-01	1076.82	0.51	**8.8E-07**	1265.39	0.08	8.9E-01
washc5	*trans*	293.53	-0.14	5.0E-01	141.55	-0.40	**9.2E-04**	143.95	-0.03	9.6E-01
pxk	*trans*	205.36	0.19	2.9E-01	183.15	0.67	**1.9E-04**	120.35	0.65	7.3E-02
hint1	NA	1719.70	0.28	2.6E-01	824.17	0.46	**9.4E-03**	336.79	-1.03	**9.7E-03**
nsmce2	NA	260.89	-0.48	**1.4E-04**	79.51	-0.44	1.5E-02	82.97	-0.80	6.1E-02
gimap2	NA	17.46	2.14	**5.5E-04**	46.44	0.04	9.5E-01	57.94	1.89	1.6E-02
cdk5r1	NA	106.52	-0.59	**3.7E-03**	292.02	0.31	9.2E-02	7.22	-1.18	3.6E-01
dph5	NA	344.39	0.51	**2.8E-03**	108.03	0.20	2.9E-01	63.25	-0.28	6.4E-01
pdhb	NA	662.23	0.41	**6.9E-03**	2359.84	0.06	8.1E-01	680.86	-0.29	5.8E-01
irf1	NA	5.62	0.32	7.6E-01	142.62	-1.19	**2.9E-04**	360.24	1.17	1.0E-01

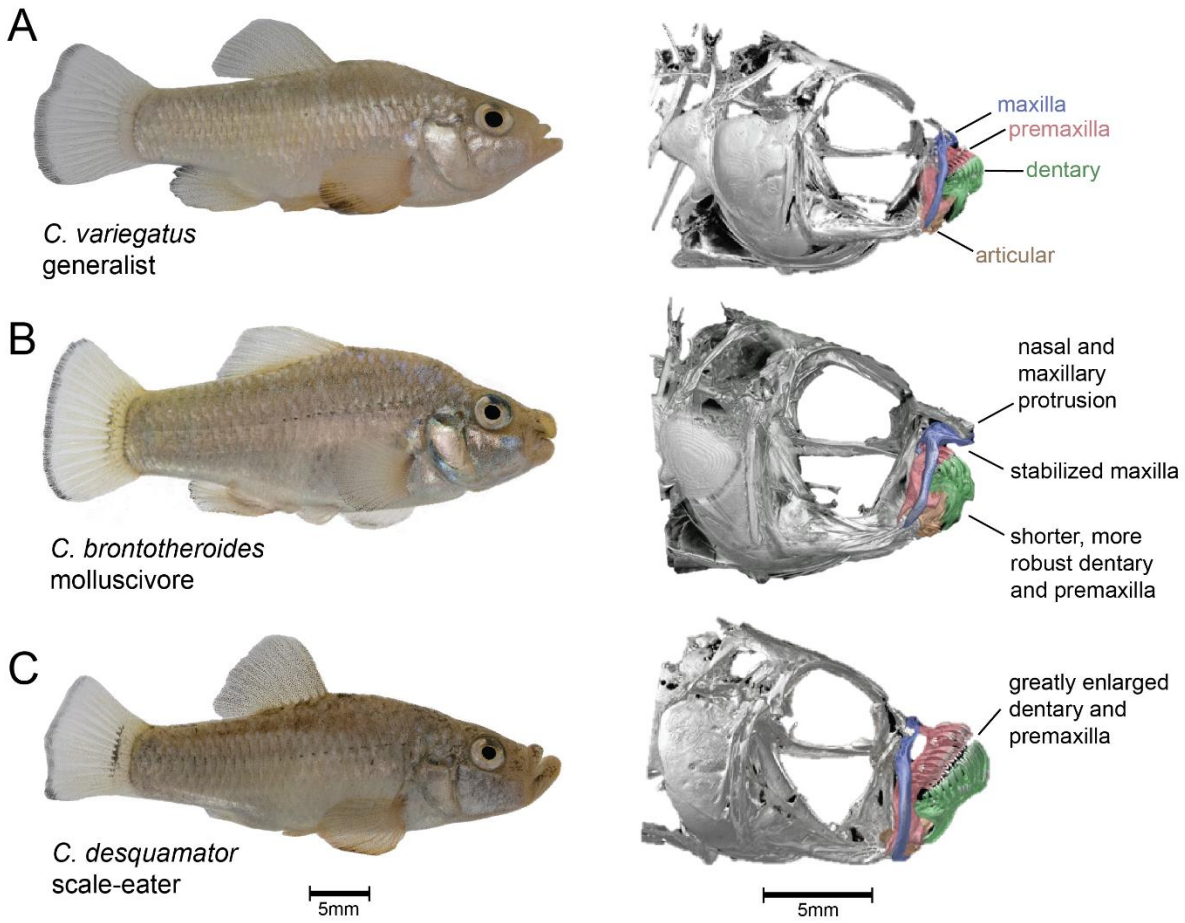

Figure 5.1. San Salvador Island pupfishes exhibit exceptional craniofacial divergence despite recent divergence times.

A) *Cyprinodon variegatus* (generalist), B) *C. brontotheroides* (molluscivore), C) *C. desquamator* (scale-eater). µCT scans modified from (Hernandez et al. 2018) show major craniofacial skeletal structures diverged among species including the maxilla (blue), premaxilla (red), dentary (green), and articular (brown).

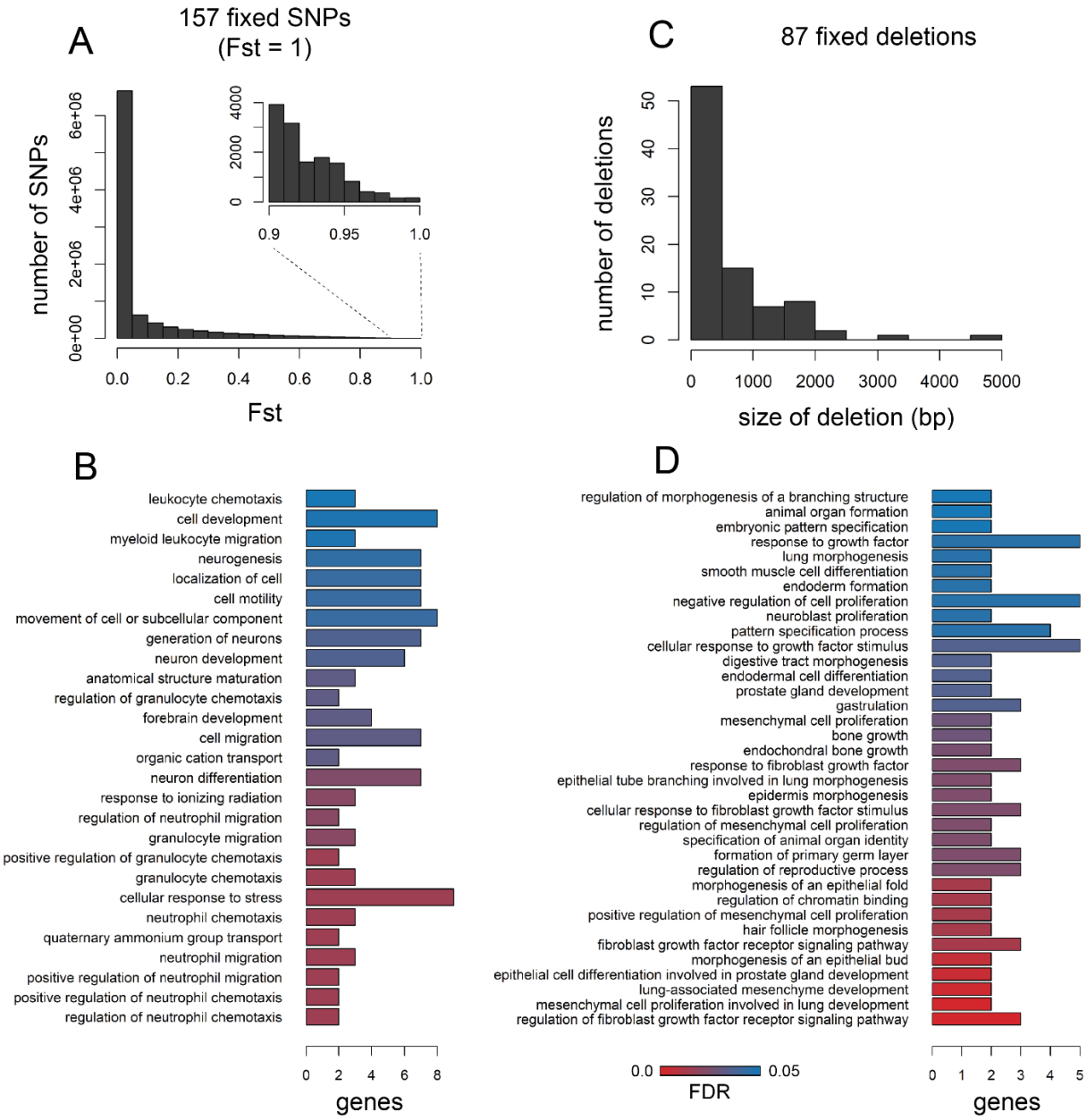

Figure 5.2. Very few SNPs and structural variants are fixed between trophic specialists.

A) Distribution of Weir and Cokerham *Fst* values across 9.3 million SNPs. 157 were fixed between species (*Fst* = 1). B) 46 of the 157 SNPs were located near 27 genes that were enriched for 27 biological processes (FDR < 0.05). C) Size distribution of the 87 deletions are fixed between species out of 80,012 structural variants. D) 34 of the 87 fixed deletions were within 10 kb of 34 genes that were enriched for 36 biological processes.

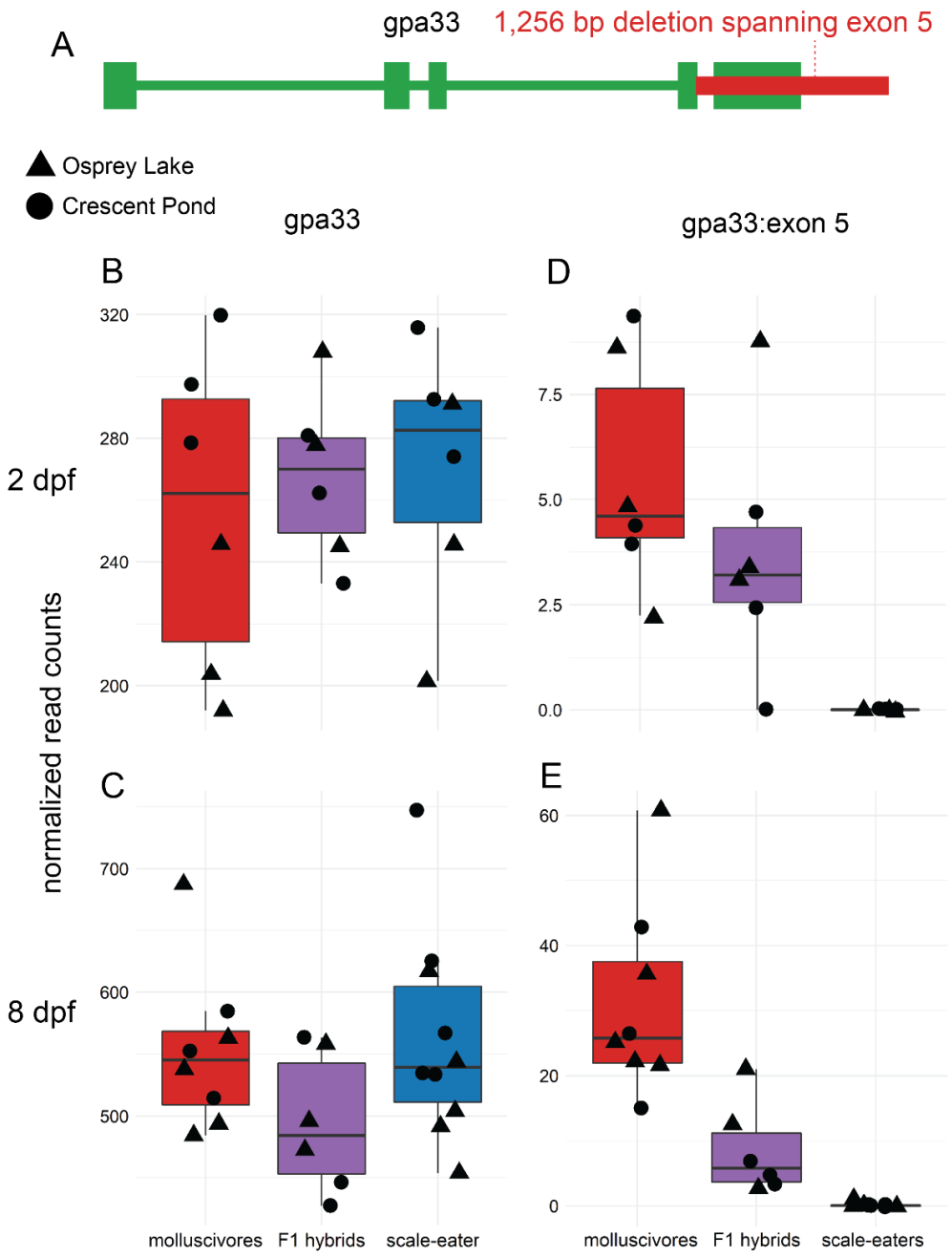

Figure 5.3. The only fixed variant within a protein coding region is an exon deletion of gpa33.

A) A 1,256 bp deletion (red) identified by DELLY spans the entire fifth exon of *gpa33* and is fixed in scale-eaters. B and C) The gene is not significantly differentially expressed between molluscivores (red) and scale-eaters (blue) at 2 days post fertilization (dpf) or 8 dpf when considering read counts across all exons ($P > 0.05$). D and E) However, when only considering the fifth exon, scale-eaters show no expression and F1 hybrids (purple) show reduced expression, supporting evidence for the deletion.

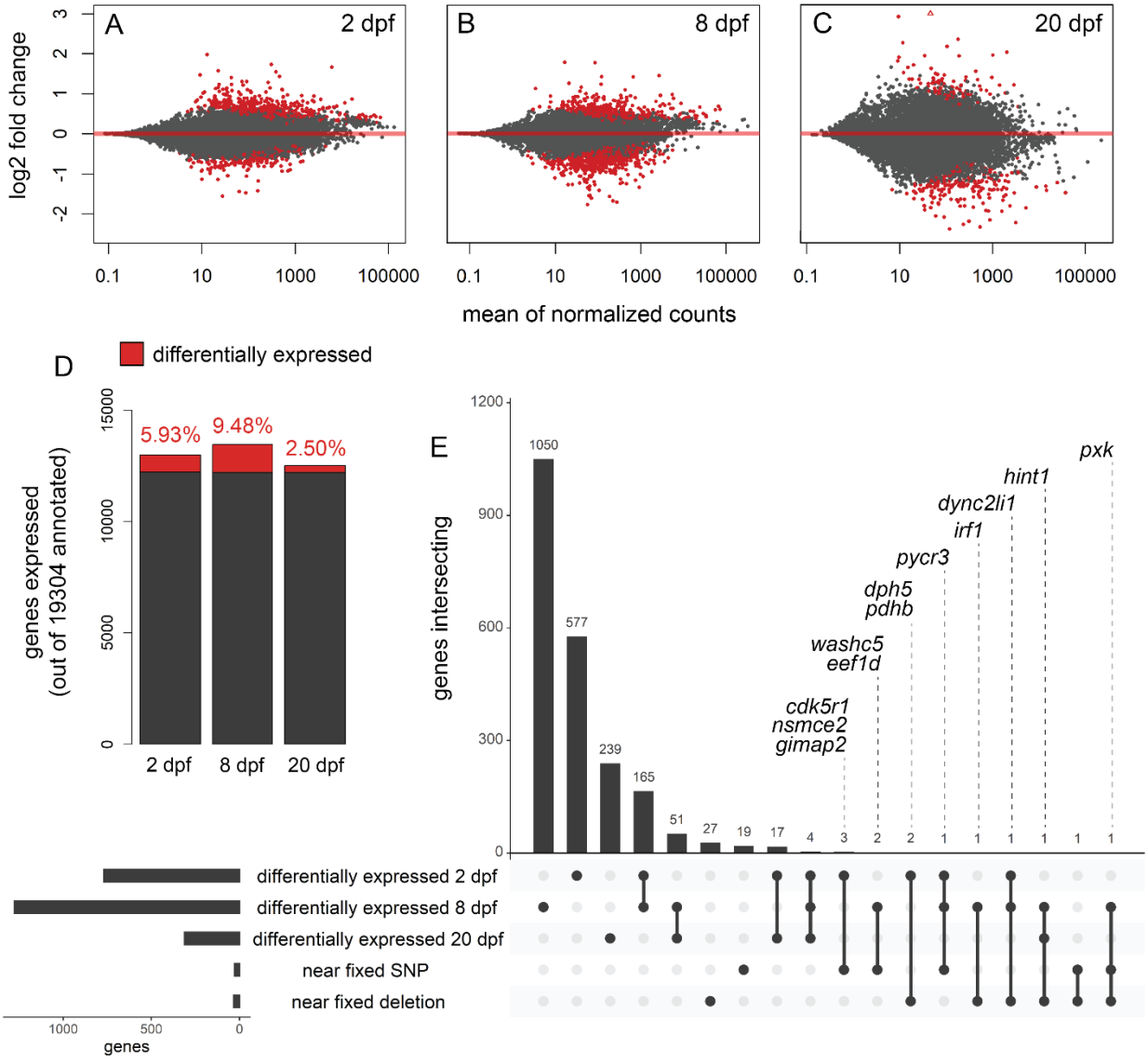

Figure 5.4. Genes near fixed variants are differentially expressed between species across three developmental stages.

Plots show genes differentially expressed (red; $P < 0.01$) between molluscivores and scale-eaters at A) 2 days post fertilization (dpf), B) 8 dpf, and C) 20 dpf. Positive log2 fold changes indicate higher expression in scale-eaters relative to molluscivores. D) Proportion of genes differentially expressed out of the total number of genes expressed across three stages. E) UpSet plot (Conway et al. 2017) showing intersection across five sets: genes differentially expressed at each of the three stages, genes within 10 kb of fixed SNPs, and genes within 10 kb of fixed deletions. The twelve labeled genes were differentially expressed during at least one stage and within 10 kb of fixed variants.

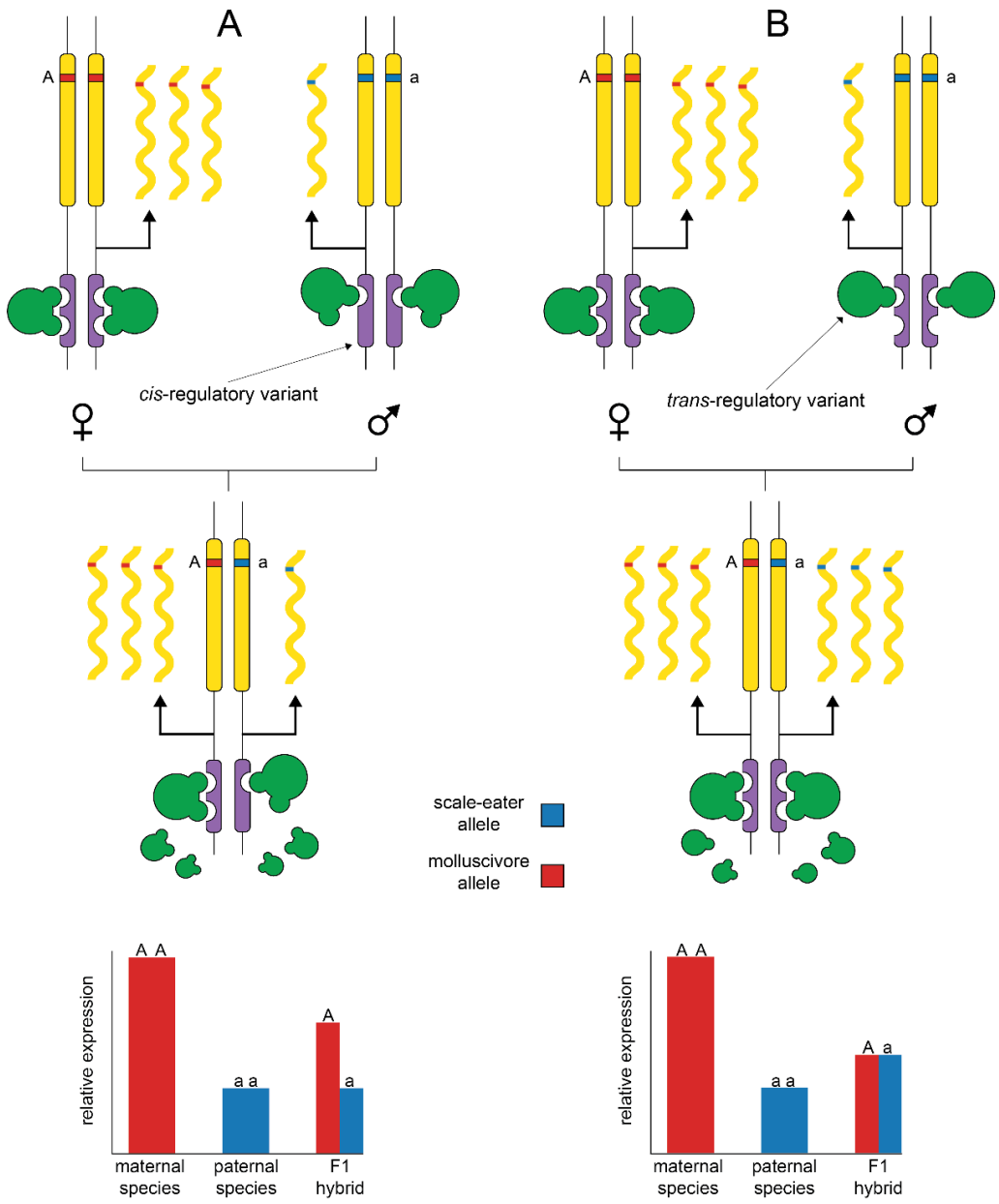

Figure 5.5. Deciphering between cis- and trans-regulatory divergence influencing gene expression.

Diagrams show protein coding gene regions (yellow) regulated by linked *cis*-acting elements (purple) and *trans*-acting binding proteins (green). In the examples, a female molluscivore is crossed with a male scale-eater to produce an F1 hybrid. The two species are alternatively homozygous for an allele within the coding region of a gene that shows higher expression in the molluscivore than the scale-eater. A) A *cis*-acting variant causing reduced expression results in low expression of the scale-eater allele in the F1 hybrid. B) Lower expression in the scale-eater is caused by a *trans*-acting variant, resulting in similar expression levels of both parental alleles in the F1 hybrid.

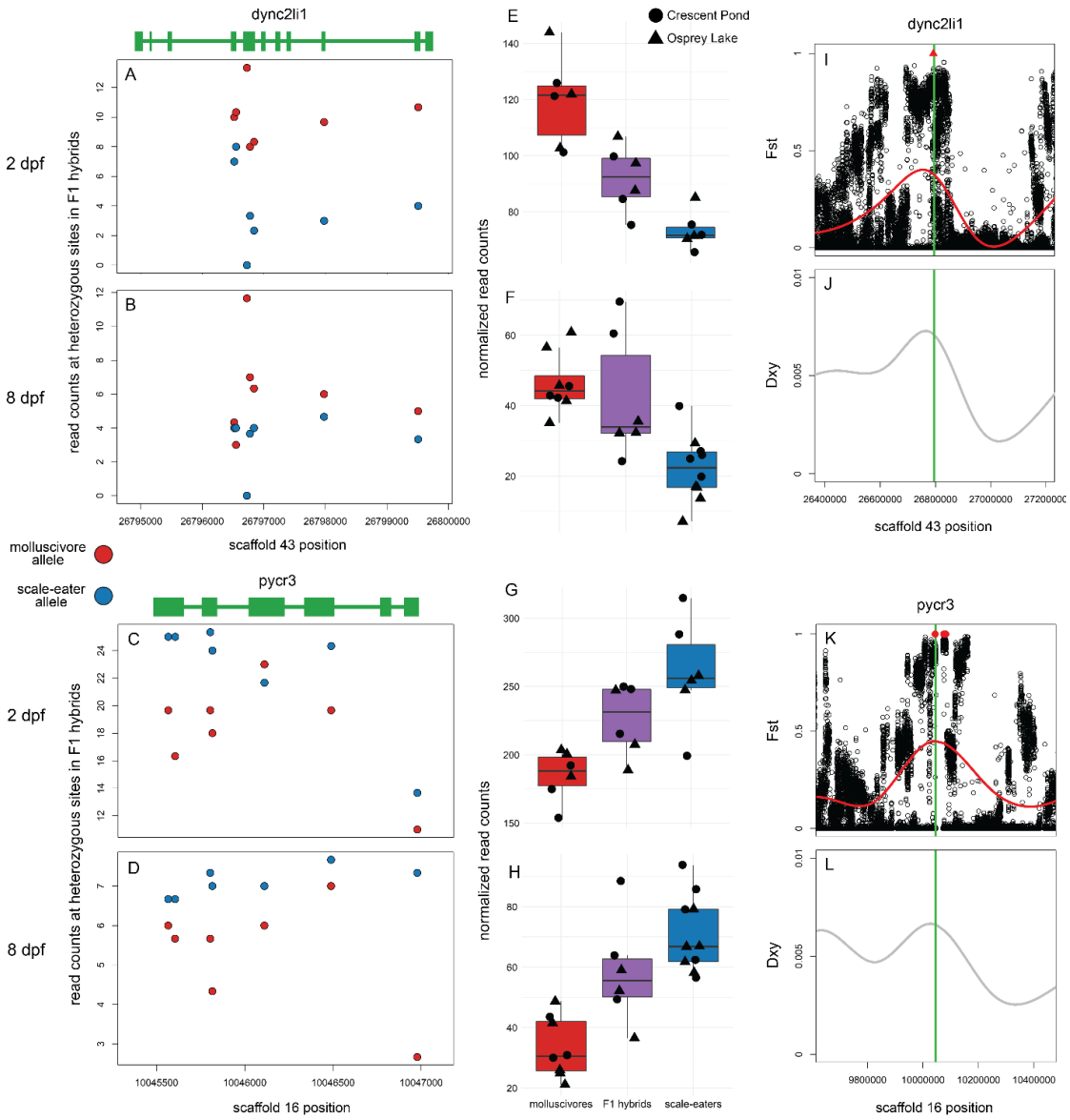

Figure 5.6. Two genes near fixed variants show cis-regulatory divergence between trophic specialists.

A-D) Mean counts for reads spanning *dync2li1* and *pycr3* that match parental alleles at heterozygous sites are shown for crosses between Crescent Pond molluscivores (red) and scale-eaters (blue) at 2 dpf (A and C) and 8 dpf (B and D). E-H) Normalized read counts for F1 offspring from Crescent Pond (circles) and Osprey Lake (triangles) crosses. Both genes are differentially expressed between molluscivores (red) and scale-eaters (blue) at both developmental stages ($P < 0.01$) and show additive inheritance in F1 hybrids (purple). For both genes, F1 hybrids show higher expression of alleles derived from the parental species that shows higher gene expression in purebred F1 offspring (MBASED $P < 0.05$), consistent with *cis*-regulatory divergence between species. I-L) Both genes (green lines) are within regions showing high relative genetic differentiation (*Fst*; I and K) and high absolute genetic divergence (*Dxy*; J and L). Red triangle shows fixed deletion. Red points show fixed SNPs (*Fst* = 1).

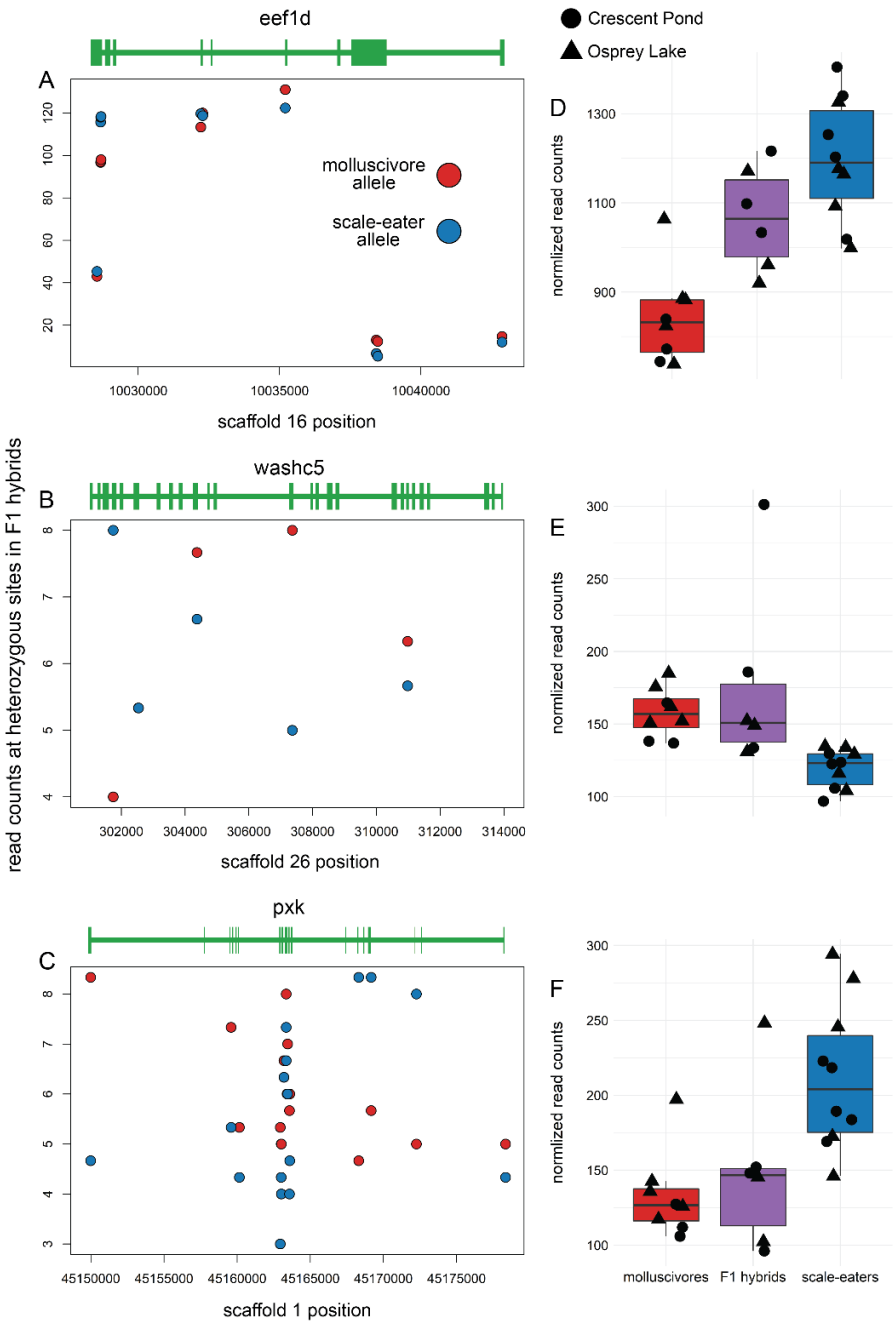

Figure 5.7. Three genes near fixed variants show trans-regulatory divergence between trophic specialists.

A-D) Mean counts for reads spanning A) *eef1d*, B) *washc5*, and C) *pxk* that match parental alleles at heterozygous sites are shown for crosses between Crescent Pond molluscivores (red) and scale-eaters (blue) at 8 dpf. D-F) Library size normalized read counts for F1 offspring from Crescent Pond (circles) and Osprey Lake (triangles) crosses. All three genes are differentially expressed between molluscivores (red) and scale-eaters (blue) at 8 dpf. None of these genes showed significant allele-specific expression in F1 hybrids (purple; MBASED, $P > 0.05$), indicating *trans*-regulatory mechanisms underlying expression divergence.

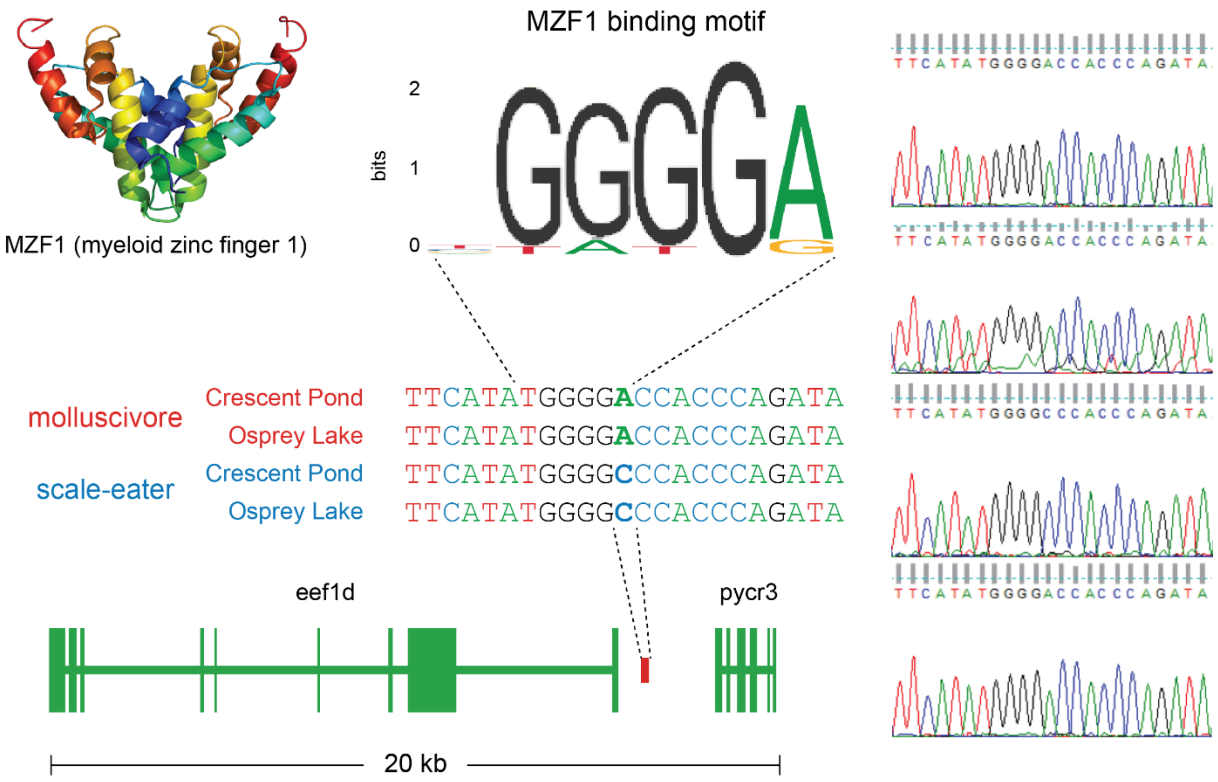

Figure 5.8. Sanger sequencing confirms fixed SNP that could alter transcription factor binding near pycr3.

Chromatograms on the right confirm the A-to-C transversion fixed in scale-eaters that falls between *eef1d* (Fig. 5.7A) and *pycr3* (Fig. 5.6C-L). The myeloid zinc finger transcription factor binds a motif that matches the molluscivore (JASPAR database matrix ID MA0056.1), however, the scale-eater allele alters this motif.

APPENDIX A: SUPPLEMENTARY MATERIAL FOR CHAPTER 1

A1. Supplemental Tables

Table A1.1. SNPs Fixed Between C. variegatus (generalist) and C. desquamator (scale-eater).

Asterisks (*) show SNPs in gene regions (bold) annotated for skeletal system effects.

SNP	Scaffold	Median PIP	PIP percentile	Median β	*P*	Gene
1	KL652900.1	0.00245	0.988	0.617581	5.36E-10	NA
2	KL653356.1	0.00205	0.966	0.458927	2.59E-09	NA
3	KL652510.1	0.002	0.961	0.460259	1.53E-09	NA
4	KL653712.1	0.0019	0.950	0.121984	1.07E-05	FAM49B
5	KL652554.1	0.00185	0.941	0.302317	2.43E-09	PIGR
6	KL653302.1	0.0016	0.888	0.136604	4.36E-10	MAGI3
7	KL653302.1	0.0016	0.888	0.136604	8.26E-10	MAGI3
8	KL653302.1	0.0016	0.888	0.136604	1.67E-08	MAGI3
9	KL652758.1	0.0013	0.778	0.128798	4.87E-06	FBXO32
10	KL652758.1	0.0013	0.778	0.128798	4.87E-06	FBXO32
11	KL652583.1	0.0012	0.729	0.018276	0.01289	EIF2B3
12	KL652583.1	0.00115	0.698	0.01545	0.02878	EIF2B3
13	KL652584.1	0.0011	0.674	-0.047672	1.14E-09	LINGO1
14	KL652584.1	0.0011	0.674	-0.047672	2.31E-09	LINGO1
15*	KL652632.1	0.001	0.614	0.126082	3.08E-09	**CABP2**
16*	KL652632.1	0.001	0.614	0.126082	1.17E-08	**CABP2**
17*	KL653422.1	9.00E-04	0.551	-0.030165	0.0811	**COL11A1**
18*	KL653422.1	9.00E-04	0.551	-0.030165	0.0811	**COL11A1**
19*	KL653422.1	9.00E-04	0.551	-0.030165	0.0811	**COL11A1**
20*	KL653422.1	9.00E-04	0.551	-0.030165	0.128	**COL11A1**
21	KL652603.1	0.00065	0.386	0.005032	0.000628	MEF2C
22	KL652585.1	6.00E-04	0.362	0.021429	9.90E-08	FAM172A

Table A1.2. Top 20 SNPs associated with jaw size after correcting for population structure in PLINK with the top two principle components.

None reach our Bonferroni corrected level of significance ($P < 4.0 \times 10^{-9}$).

SNP	Scaffold	Beta	*P*
1	KL653294.1	-6.432	6.83E-07
2	KL653071.1	-6.403	7.36E-07
3	KL653414.1	6.393	7.57E-07
4	KL653264.1	6.37	8.03E-07
5	KL652620.1	6.534	9.31E-07
6	KL653172.1	6.368	9.58E-07
7	KL652789.1	6.217	1.20E-06
8	KL653414.1	6.158	1.40E-06
9	KL653049.1	6.207	1.45E-06
10	KL652731.1	6.179	1.55E-06
11	KL652573.1	6.013	2.04E-06
12	KL652868.1	-6.137	2.05E-06
13	KL653566.1	6.188	2.15E-06
14	KL652753.1	5.973	2.27E-06
15	KL652694.1	-6.08	2.36E-06
16	KL652841.1	5.955	2.38E-06
17	KL652723.1	-5.931	2.54E-06
18	KL653042.1	6.035	2.65E-06
19	KL652694.1	-5.966	2.69E-06
20	JPKM01108474.1	5.891	2.82E-06

A2. Supplemental Figures

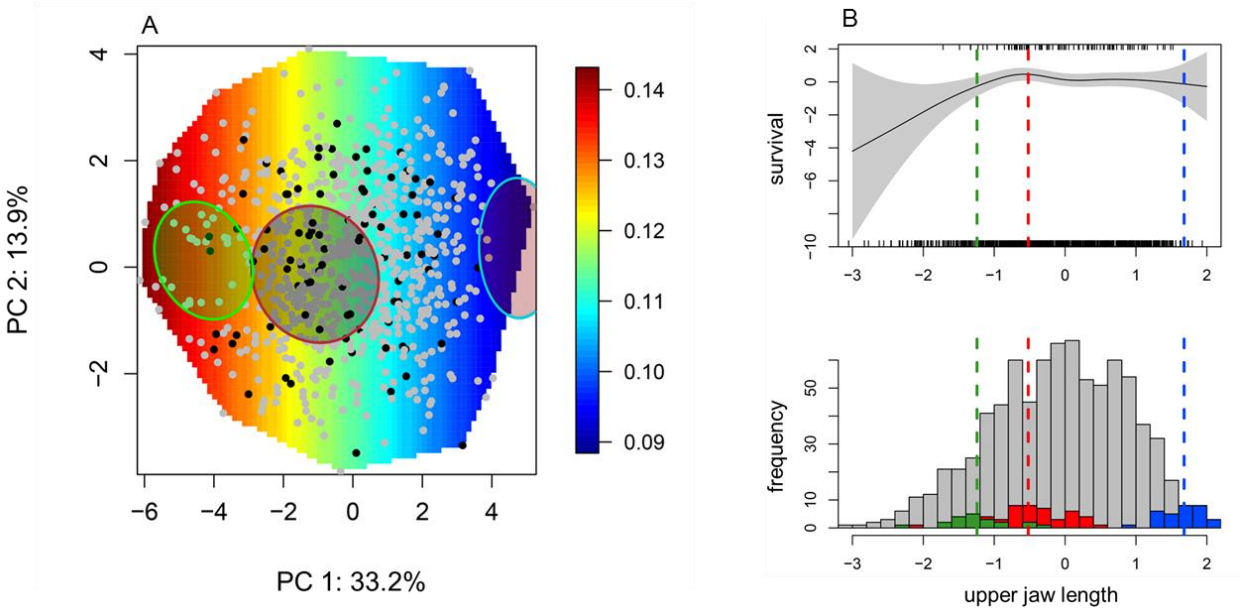

Figure A2.1. Large phenotypic distance between C. desquamator (scale-eater) and C. variegatus (generalist).

C. variegatus (red), *C. desquamator* (blue), and *C. brontotheroides* (green) from each lake population were intercrossed in every direction to produce F_2 hybrids which were left for three months in an enclosure on San Salvador. Survival probability is plotted against two axes of the discriminant morphospace, indicating a wide range of jaw phenotypes in the F_2 hybrids (modified from Martin and Wainwright, 2013). Heat colors correspond to survival probability (with blue being low and red being high). A) F_2 hybrid survivors (grey dots) and deaths (black dots) plotted against principal components together explaining 47% of the variation across measurements for 16 morphological traits. Phenotypes for all lab-raised purebred species are represented by 95% confidence ellipses. The phenotypic distance is greater between *C. desquamator* and *C. variegatus* than *C. brontotheroides* and *C. variegatus*. B) Smoothing splines with 95% shaded confidence regions show survival probability (upper panel) and histograms (lower panel) show the range of upper jaw measurements within the F_2 hybrid population relative to parental trait ranges. Rug plots indicate jaw lengths of F_2 hybrid survivors (upper axis) and deaths (lower axis). Dashed lines show mean jaw length for each species (modified from Martin 2016a).

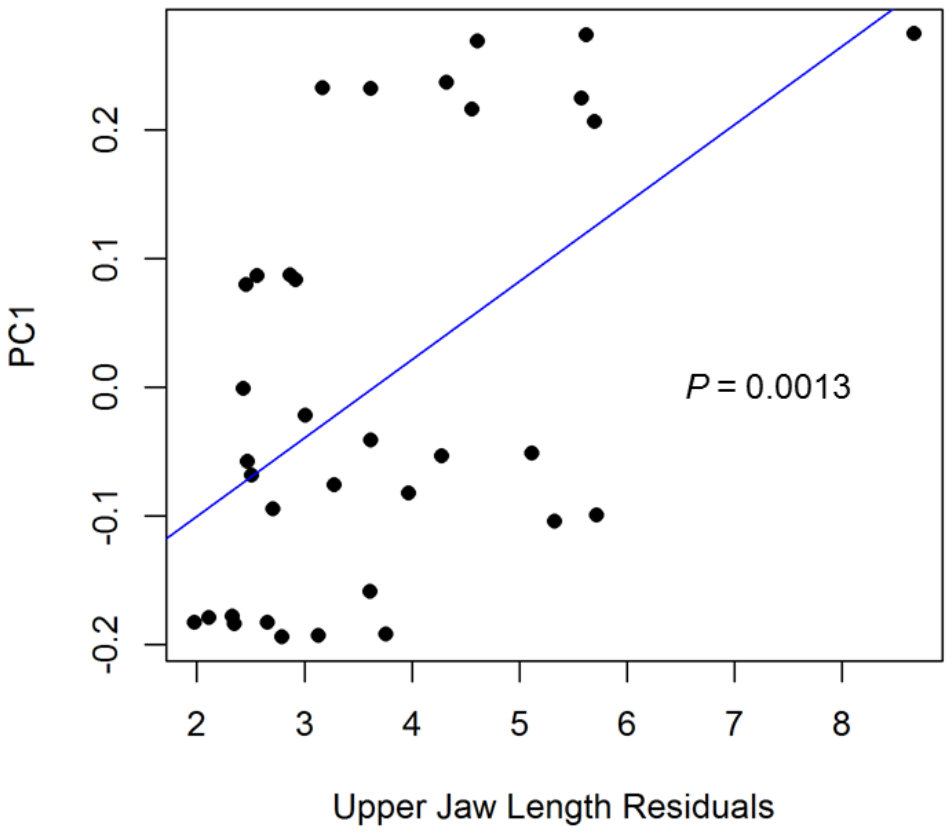

Figure A2.2. Principal Component 1 Correlated with Jaw Length.

Jaw measurements were log transformed and regressed against log-transformed body length. We fit a log-transformed trait by log-transformed body length linear regression and plot the residuals versus the top principal component that explains 5.45% of the variation in our genomic dataset. The correlation between jaw size and PC1 is significant (P = 0.0013).

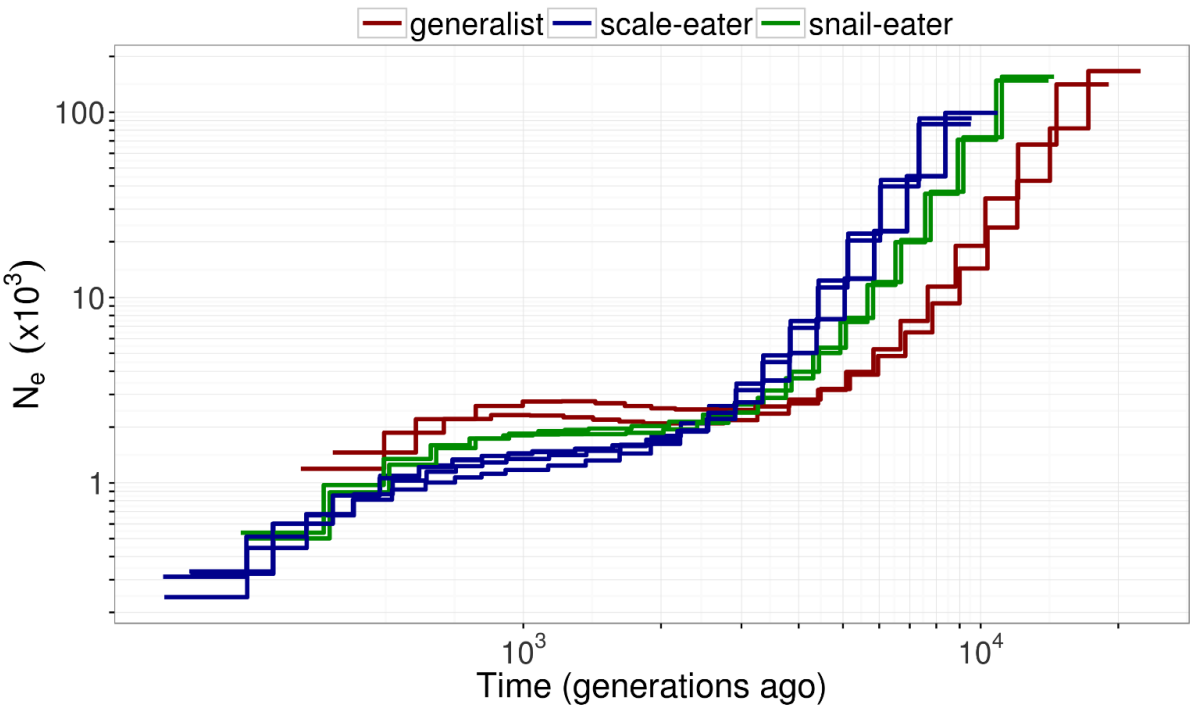

Figure A2.3. Ancestral Population Size of San Salvador Pupfish Species.

Historical effective population sizes estimated by the Multiple Sequentially Markovian Coalescent (MSMC) for *C. variegatus* (red), *C. desquamator* (blue) and *C. brontotheroides* (green) using a six-month generation time and mutation rate estimated for cichlids (6.6×10^{-8} mutations per site per year).

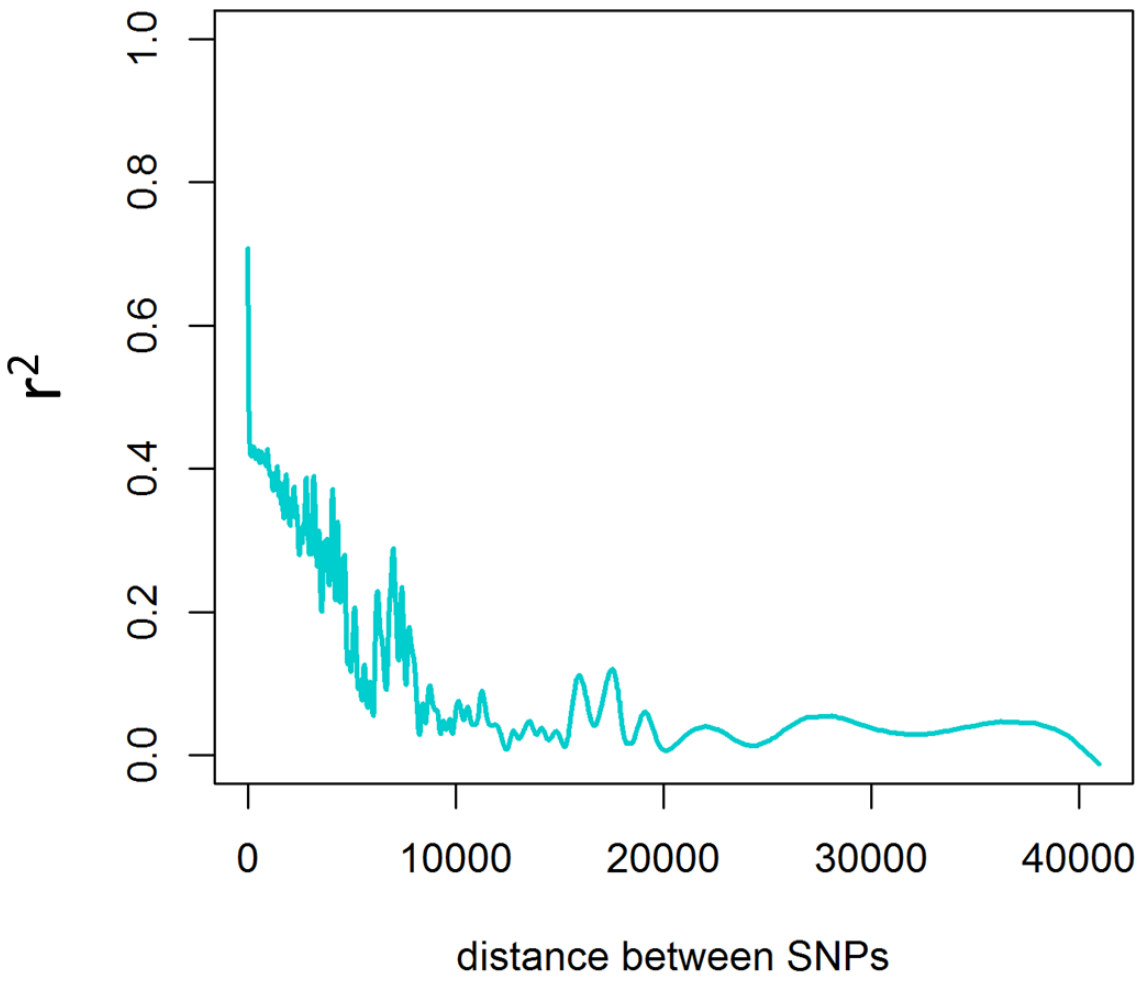

Figure A2.4. Decay of linkage disequilibrium across a 4.5 Mb scaffold.

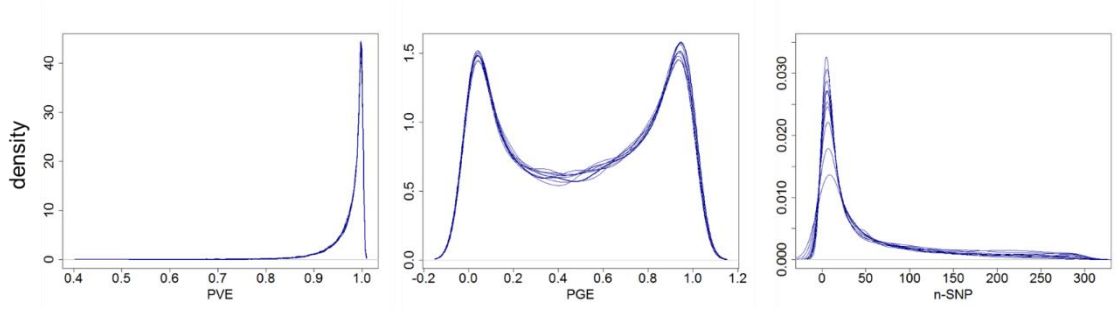

Figure A2.5. Posterior density distributions for hyperparameters obtained from GEMMA's Bayesian sparse linear mixed model.

A) The proportion of variance in phenotypes explained by every SNP (PVE), B) the proportion of phenotypic variation explained by SNPs of large effect (PGE), and C) the number of large effect SNPs required to explain PGE (nSNP). Individual lines represent ten independent MCMC runs.

APPENDIX B: SUPPLEMENTARY MATERIAL FOR CHAPTER 2

B1. Supplemental Tables

Table B1.1. Total mRNA sequencing sampling design.

	Crescent Pond		Little Lake	
species	8-10 dpf	17-20 dpf	8-10 dpf	17-20 dpf
generalist	3 F_2	3 F_2	3 F_1	3 F_1
snail-eater	3 F_2	3 F_2	3 F_2	3 F_2
scale-eater	0	0	3 F_1	2 F_1

Table B1.2. Four genes showing opposite expression patterns in specialists relative to generalists.

				snail-eater *vs.* generalist		scale-eater *vs.* generalist	
scaffold	stage	symbol	zebrafish ortholog	$\log_2$ FC	*P*	$\log_2$ FC	*P*
015150477	17-20 dpf	LOC107096735	mybpc2a	2.13	0.05	-3.27	0.00
015150518	8-10 dpf	LOC107082892	si:ch211-197h24.9	2.90	0.01	-2.61	0.02
015150587	8-10 dpf	agxt2	agxt2	1.48	0.01	-1.43	0.01
015150546	8-10 dpf	plin2	plin2	-0.90	0.03	2.43	0.00

Table B1.3. Enriched gene ontologies for genes showing parallel changes in expression between specialists.

Representative terms were determined using REVIGO (Tomislav 2011).

ID	description	representative term
GO:0006950	response to stress	response to stress
GO:0006302	double-strand break repair	response to stress
GO:0006974	cellular response to DNA damage stimulus	response to stress
GO:0000725	recombinational repair	response to stress
GO:0000724	double-strand break repair via homologous recombination	response to stress
GO:0008150	biological_process	biological_process
GO:0008152	metabolic process	metabolism
GO:0009987	cellular process	cellular process
GO:0032501	multicellular organismal process	multicellular organismal process
GO:0032502	developmental process	developmental process
GO:0044085	cellular component biogenesis	cellular component biogenesis
GO:0048285	organelle fission	cellular component biogenesis
GO:0000280	nuclear division	cellular component biogenesis
GO:0071826	ribonucleoprotein complex subunit organization	cellular component biogenesis
GO:0043933	macromolecular complex subunit organization	cellular component biogenesis
GO:0033043	regulation of organelle organization	cellular component biogenesis
GO:0042254	ribosome biogenesis	cellular component biogenesis
GO:0070925	organelle assembly	cellular component biogenesis
GO:0016570	histone modification	cellular component biogenesis
GO:0006325	chromatin organization	cellular component biogenesis
GO:0007010	cytoskeleton organization	cellular component biogenesis
GO:0006996	organelle organization	cellular component biogenesis
GO:0051128	regulation of cellular component organization	cellular component biogenesis
GO:0030261	chromosome condensation	cellular component biogenesis
GO:0044699	single-organism process	single-organism process
GO:0044707	single-multicellular organism process	single-multicellular organism process
GO:0044710	single-organism metabolic process	single-multicellular organism process
GO:0044763	single-organism cellular process	single-multicellular organism process
GO:0044281	small molecule metabolic process	single-multicellular organism process
GO:0060041	retina development in camera-type eye	single-multicellular organism process
GO:0048519	negative regulation of biological process	negative regulation of biological process
GO:0080090	regulation of primary metabolic process	negative regulation of biological process

GO:0031324	negative regulation of cellular metabolic process	negative regulation of biological process
GO:0019222	regulation of metabolic process	negative regulation of biological process
GO:0050794	regulation of cellular process	negative regulation of biological process
GO:0048856	anatomical structure development	anatomical structure development
GO:0051303	establishment of chromosome localization	establishment of chromosome localization
GO:0065007	biological regulation	biological regulation
GO:0071840	cellular component organization or biogenesis	cellular component organization or biogenesis
GO:1901575	organic substance catabolic process	organic substance catabolism
GO:0009056	catabolic process	catabolism
GO:0009058	biosynthetic process	biosynthesis
GO:0006479	protein methylation	protein methylation
GO:0090304	nucleic acid metabolic process	protein methylation
GO:0055086	nucleobase-containing small molecule metabolic process	protein methylation
GO:0019538	protein metabolic process	protein methylation
GO:0006310	DNA recombination	protein methylation
GO:0002181	cytoplasmic translation	protein methylation
GO:0006518	peptide metabolic process	protein methylation
GO:0006271	DNA strand elongation involved in DNA replication	protein methylation
GO:0010467	gene expression	protein methylation
GO:0006275	regulation of DNA replication	protein methylation
GO:0006261	DNA-dependent DNA replication	protein methylation
GO:0018205	peptidyl-lysine modification	protein methylation
GO:0022616	DNA strand elongation	protein methylation
GO:0019752	carboxylic acid metabolic process	protein methylation
GO:0016070	RNA metabolic process	protein methylation
GO:0016072	rRNA metabolic process	protein methylation
GO:0044260	cellular macromolecule metabolic process	protein methylation
GO:0043412	macromolecule modification	protein methylation
GO:0051052	regulation of DNA metabolic process	protein methylation
GO:0006082	organic acid metabolic process	protein methylation
GO:0008213	protein alkylation	protein methylation
GO:0044267	cellular protein metabolic process	protein methylation
GO:0006413	translational initiation	protein methylation
GO:1901564	organonitrogen compound metabolic process	protein methylation
GO:1901566	organonitrogen compound biosynthetic process	protein methylation
GO:0006139	nucleobase-containing compound metabolic process	protein methylation

GO:0043603	cellular amide metabolic process	protein methylation
GO:0072521	purine-containing compound metabolic process	protein methylation
GO:0045005	DNA-dependent DNA replication maintenance of fidelity	protein methylation
GO:0034641	cellular nitrogen compound metabolic process	protein methylation
GO:0006464	cellular protein modification process	protein methylation
GO:0006259	DNA metabolic process	protein methylation
GO:0006260	DNA replication	protein methylation
GO:0043038	amino acid activation	protein methylation
GO:0034660	ncRNA metabolic process	protein methylation
GO:0006807	nitrogen compound metabolic process	nitrogen compound metabolism
GO:0019693	ribose phosphate metabolic process	ribose phosphate metabolism
GO:0019637	organophosphate metabolic process	ribose phosphate metabolism
GO:0009141	nucleoside triphosphate metabolic	ribose phosphate metabolism
GO:1901135	carbohydrate derivative metabolic process	carbohydrate derivative metabolism
GO:0044711	single-organism biosynthetic process	carbohydrate derivative metabolism
GO:0009059	macromolecule biosynthetic process	carbohydrate derivative metabolism
GO:0044249	cellular biosynthetic process	carbohydrate derivative metabolism
GO:0043170	macromolecule metabolic process	carbohydrate derivative metabolism
GO:0044271	cellular nitrogen compound biosynthetic process	carbohydrate derivative metabolism
GO:1901576	organic substance biosynthetic process	carbohydrate derivative metabolism
GO:1901362	organic cyclic compound biosynthetic process	carbohydrate derivative metabolism
GO:1901360	organic cyclic compound metabolic process	carbohydrate derivative metabolism
GO:0034645	cellular macromolecule biosynthetic process	carbohydrate derivative metabolism
GO:0007059	chromosome segregation	chromosome segregation
GO:0000278	mitotic cell cycle	chromosome segregation
GO:0061640	cytoskeleton-dependent cytokinesis	chromosome segregation
GO:0007049	cell cycle	chromosome segregation
GO:0051301	cell division	chromosome segregation
GO:0007017	microtubule-based process	chromosome segregation
GO:0007018	microtubule-based movement	chromosome segregation
GO:0045787	positive regulation of cell cycle	chromosome segregation
GO:0044238	primary metabolic process	primary metabolism
GO:0006793	phosphorus metabolic process	primary metabolism
GO:0044237	cellular metabolic process	primary metabolism
GO:0046483	heterocycle metabolic process	primary metabolism
GO:0071704	organic substance metabolic process	primary metabolism
GO:0006725	cellular aromatic compound metabolic process	primary metabolism

Table B1.4. Enriched gene ontologies for genes showing divergent expression in specialists.

Representative terms were determined using REVIGO (Tomislav 2011).

ID	description	representative term
GO:0002376	immune system process	immune system process
GO:0006839	mitochondrial transport	mitochondrial transport
GO:0015031	protein transport	mitochondrial transport
GO:0006820	anion transport	mitochondrial transport
GO:0071705	nitrogen compound transport	mitochondrial transport
GO:0034504	protein localization to nucleus	mitochondrial transport
GO:0016192	vesicle-mediated transport	mitochondrial transport
GO:0015931	nucleobase-containing compound transport	mitochondrial transport
GO:0050658	RNA transport	mitochondrial transport
GO:0000041	transition metal ion transport	mitochondrial transport
GO:0033036	macromolecule localization	mitochondrial transport
GO:0042886	amide transport	mitochondrial transport
GO:0015833	peptide transport	mitochondrial transport
GO:0006403	RNA localization	mitochondrial transport
GO:1990542	mitochondrial transmembrane transport	mitochondrial transport
GO:0048193	Golgi vesicle transport	mitochondrial transport
GO:0051641	cellular localization	mitochondrial transport
GO:0015711	organic anion transport	mitochondrial transport
GO:0009620	response to fungus	response to fungus
GO:1901698	response to nitrogen compound	response to fungus
GO:0034976	response to endoplasmic reticulum stress	response to fungus
GO:0048583	regulation of response to stimulus	response to fungus
GO:0036503	ERAD pathway	response to fungus
GO:0009628	response to abiotic stimulus	response to fungus
GO:0009605	response to external stimulus	response to fungus
GO:0042221	response to chemical	response to fungus
GO:0010033	response to organic substance	response to fungus
GO:0006974	cellular response to DNA damage stimulus	response to fungus
GO:0010243	response to organonitrogen compound	response to fungus
GO:0032259	methylation	methylation
GO:0032501	multicellular organismal process	multicellular organismal process
GO:0032502	developmental process	developmental process
GO:0040007	growth	growth
GO:0050790	regulation of catalytic activity	regulation of catalytic activity

GO:0048519	negative regulation of biological process	regulation of catalytic activity
GO:0048518	positive regulation of biological process	regulation of catalytic activity
GO:0023051	regulation of signaling	regulation of catalytic activity
GO:0051246	regulation of protein metabolic process	regulation of catalytic activity
GO:0006357	regulation of transcription from RNA polymerase II promoter	regulation of catalytic activity
GO:0035556	intracellular signal transduction	regulation of catalytic activity
GO:0042592	homeostatic process	regulation of catalytic activity
GO:0010646	regulation of cell communication	regulation of catalytic activity
GO:0007267	cell-cell signaling	regulation of catalytic activity
GO:0010608	posttranscriptional regulation of gene expression	regulation of catalytic activity
GO:0065009	regulation of molecular function	regulation of catalytic activity
GO:0065008	regulation of biological quality	regulation of catalytic activity
GO:0007186	G-protein coupled receptor signaling pathway	regulation of catalytic activity
GO:0034101	erythrocyte homeostasis	regulation of catalytic activity
GO:0048872	homeostasis of number of cells	regulation of catalytic activity
GO:0009893	positive regulation of metabolic process	regulation of catalytic activity
GO:0009892	negative regulation of metabolic process	regulation of catalytic activity
GO:0034248	regulation of cellular amide metabolic process	regulation of catalytic activity
GO:0034249	negative regulation of cellular amide metabolic process	regulation of catalytic activity
GO:0061077	chaperone-mediated protein folding	chaperone-mediated protein folding
GO:1904888	cranial skeletal system development	cranial skeletal system development
GO:0048589	developmental growth	cranial skeletal system development
GO:0061061	muscle structure development	cranial skeletal system development
GO:0055002	striated muscle cell development	cranial skeletal system development
GO:0010927	cellular component assembly involved in morphogenesis	cranial skeletal system development
GO:0048863	stem cell differentiation	cranial skeletal system development
GO:0048856	anatomical structure development	cranial skeletal system development
GO:0002072	optic cup morphogenesis involved in camera-type eye development	cranial skeletal system development

GO:0014706	striated muscle tissue development	cranial skeletal system development
GO:0060322	head development	cranial skeletal system development
GO:0048705	skeletal system morphogenesis	cranial skeletal system development
GO:0046148	pigment biosynthetic process	pigment biosynthesis
GO:0044723	single-organism carbohydrate metabolic process	pigment biosynthesis
GO:0006629	lipid metabolic process	pigment biosynthesis
GO:0042440	pigment metabolic process	pigment biosynthesis
GO:1901615	organic hydroxy compound metabolic process	organic hydroxy compound metabolism
GO:0006914	autophagy	autophagy
GO:0016236	macroautophagy	macroautophagy
GO:0007033	vacuole organization	vacuole organization
GO:0071826	ribonucleoprotein complex subunit organization	vacuole organization
GO:0044802	single-organism membrane organization	vacuole organization
GO:0006325	chromatin organization	vacuole organization
GO:0097435	supramolecular fiber organization	vacuole organization
GO:0044085	cellular component biogenesis	vacuole organization
GO:0044087	regulation of cellular component biogenesis	vacuole organization
GO:0000469	cleavage involved in rRNA processing	vacuole organization
GO:0000466	maturation of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	vacuole organization
GO:0043933	macromolecular complex subunit organization	vacuole organization
GO:0042254	ribosome biogenesis	vacuole organization
GO:0070925	organelle assembly	vacuole organization
GO:0016570	histone modification	vacuole organization
GO:0007010	cytoskeleton organization	vacuole organization
GO:0007005	mitochondrion organization	vacuole organization
GO:0006996	organelle organization	vacuole organization
GO:0030036	actin cytoskeleton organization	vacuole organization
GO:0051128	regulation of cellular component organization	vacuole organization
GO:0043254	regulation of protein complex assembly	vacuole organization
GO:0006457	protein folding	protein folding
GO:0006790	sulfur compound metabolic process	sulfur compound metabolism

GO:1901137	carbohydrate derivative biosynthetic process	carbohydrate derivative biosynthesis
GO:0009100	glycoprotein metabolic process	carbohydrate derivative biosynthesis
GO:0009101	glycoprotein biosynthetic process	carbohydrate derivative biosynthesis
GO:0033013	tetrapyrrole metabolic process	tetrapyrrole metabolism
GO:0016072	rRNA metabolic process	tetrapyrrole metabolism
GO:0016071	mRNA metabolic process	tetrapyrrole metabolism
GO:0006260	DNA replication	tetrapyrrole metabolism
GO:0090305	nucleic acid phosphodiester bond hydrolysis	tetrapyrrole metabolism
GO:0006396	RNA processing	tetrapyrrole metabolism
GO:0008380	RNA splicing	tetrapyrrole metabolism
GO:0006397	mRNA processing	tetrapyrrole metabolism
GO:0072528	pyrimidine-containing compound biosynthetic process	tetrapyrrole metabolism
GO:0072527	pyrimidine-containing compound metabolic process	tetrapyrrole metabolism
GO:0072521	purine-containing compound metabolic process	tetrapyrrole metabolism
GO:0034660	ncRNA metabolic process	tetrapyrrole metabolism
GO:0043414	macromolecule methylation	macromolecule methylation
GO:0018193	peptidyl-amino acid modification	macromolecule methylation
GO:0031329	regulation of cellular catabolic process	macromolecule methylation
GO:0009057	macromolecule catabolic process	macromolecule methylation
GO:0070647	protein modification by small protein conjugation or removal	macromolecule methylation
GO:0044270	cellular nitrogen compound catabolic process	macromolecule methylation
GO:0030163	protein catabolic process	macromolecule methylation
GO:0016567	protein ubiquitination	macromolecule methylation
GO:0009894	regulation of catabolic process	macromolecule methylation
GO:0044282	small molecule catabolic process	macromolecule methylation
GO:1901361	organic cyclic compound catabolic process	macromolecule methylation
GO:0046777	protein autophosphorylation	macromolecule methylation
GO:0009451	RNA modification	macromolecule methylation
GO:0030029	actin filament-based process	actin filament-based process
GO:0032787	monocarboxylic acid metabolic process	actin filament-based process
GO:0046394	carboxylic acid biosynthetic process	actin filament-based process
GO:0051301	cell division	actin filament-based process
GO:1901605	alpha-amino acid metabolic process	actin filament-based process

GO:0009132	nucleoside diphosphate metabolic process	actin filament-based process
GO:0008610	lipid biosynthetic process	actin filament-based process
GO:0009141	nucleoside triphosphate metabolic process	actin filament-based process
GO:0006720	isoprenoid metabolic process	actin filament-based process
GO:0007049	cell cycle	actin filament-based process
GO:0006643	membrane lipid metabolic process	actin filament-based process
GO:0008219	cell death	actin filament-based process
GO:1903047	mitotic cell cycle process	actin filament-based process
GO:0006915	apoptotic process	actin filament-based process
GO:0006091	generation of precursor metabolites and energy	generation of precursor metabolites and energy
GO:0006732	coenzyme metabolic process	coenzyme metabolism
GO:0006779	porphyrin-containing compound biosynthetic process	coenzyme metabolism
GO:0044707	single-multicellular organism process	single-multicellular organism process
GO:0072358	cardiovascular system development	single-multicellular organism process
GO:0055123	digestive system development	single-multicellular organism process
GO:0061008	hepaticobiliary system development	single-multicellular organism process
GO:0001501	skeletal system development	single-multicellular organism process
GO:0001889	liver development	single-multicellular organism process
GO:0031016	pancreas development	single-multicellular organism process
GO:0048732	gland development	single-multicellular organism process
GO:0051186	cofactor metabolic process	cofactor metabolism

Table B1.5. Eleven genes previously described as candidates influencing craniofacial divergence are differentially expressed between generalists and specialists.

scaffold	log_2 FC	*P*	symbol	comparison	stage
15151665	4.49	<0.01	znf664	generalist vs. scale-eater	17-20 dpf
15150619	2.79	<0.01	abcg5	generalist vs. scale-eater	8-10 dpf
15150999	1.22	0.04	lrp1b	generalist vs. snail-eater	17-20 dpf
15151015	1.07	0.02	gmds	generalist vs. scale-eater	17-20 dpf
15150619	0.91	0.01	dync2li1	generalist vs. scale-eater	8-10 dpf
15151665	0.7	0.04	fam49b	generalist vs. scale-eater	8-10 dpf
15150480	-0.51	0.01	tmem30a	generalist vs. scale-eater	8-10 dpf
15150538	-0.62	0.01	fam172a	generalist vs. scale-eater	8-10 dpf
15151075	-0.64	<0.01	atp8a1	generalist vs. scale-eater	8-10 dpf
15150670	-0.73	0.01	ash1l	generalist vs. scale-eater	8-10 dpf
15150619	-1.16	0.03	hint1	generalist vs. scale-eater	17-20 dpf

Table B1.6. 68 out of 84 gene regions containing fixed variants show signs of a hard sweep.

Estimated by SweeD. Composite likelihood ratio (CLR) > 95th percentile across their respective scaffolds.

scaffold	fixed SNPs	log2 fold change	adjusted *P*	stage	CLR	*Cyprinodon* gene symbol
15150501	1	0.68	0.03	8-10dpf	0.42	LOC107082156
15150501	1	-1.13	0.03	17-20dpf	0.42	LOC107082156
15150501	4	-1.26	0.00	8-10dpf	0.42	LOC107082264
15151439	2	-0.62	0.02	8-10dpf	0.42	LOC107100553
15151189	10	0.84	0.00	8-10dpf	0.39	LOC107097945
15151015	1	-1.23	0.03	8-10dpf	0.36	LOC107095655
15150556	1	-0.63	0.03	8-10dpf	0.33	mef2c
15150556	1	-1.13	0.00	17-20dpf	0.33	mef2c
15150680	1	-1.25	0.00	8-10dpf	0.31	plgrkt
15150776	1	-1.02	0.03	17-20dpf	0.30	LOC107091063
15151452	1	0.74	0.00	8-10dpf	0.28	reck
15150730	1	1.32	0.00	8-10dpf	0.24	exosc4
15151162	26	-1.20	0.02	8-10dpf	0.23	LOC107097607
15151066	2	1.43	0.00	8-10dpf	0.23	dbf4
15151810	3	-0.97	0.01	8-10dpf	0.23	LOC107103000
15151726	1	0.87	0.00	8-10dpf	0.22	LOC107102549
15150691	2	0.92	0.04	8-10dpf	0.20	LOC107089095
15150455	3	0.58	0.05	8-10dpf	0.20	fam188a
15150455	1	-1.30	0.00	8-10dpf	0.20	LOC107102995
15151892	1	-1.34	0.01	17-20dpf	0.19	loxl3
15151892	1	1.02	0.01	8-10dpf	0.19	loxl3
15151400	2	-1.48	0.00	8-10dpf	0.18	LOC107100233
15150688	2	0.62	0.01	8-10dpf	0.18	LOC107089013
15150854	2	-1.13	0.00	8-10dpf	0.18	lmo7
15150495	1	-1.35	0.01	8-10dpf	0.18	adgrg2
15150702	2	-0.54	0.00	8-10dpf	0.17	cct8
15150702	3	0.69	0.03	8-10dpf	0.17	LOC107089362
15150702	15	1.73	0.00	8-10dpf	0.17	LOC107089382
15150702	15	-2.32	0.00	17-20dpf	0.17	LOC107089382
15150924	1	-0.92	0.00	8-10dpf	0.17	LOC107094239
15150533	1	1.11	0.02	8-10dpf	0.16	erap2
15150763	1	-2.68	0.00	8-10dpf	0.16	LOC107090753
15150467	3	-0.79	0.04	17-20dpf	0.15	nxn
15151167	19	1.34	0.00	17-20dpf	0.15	LOC107097675
15151665	2	0.70	0.04	8-10dpf	0.15	fam49b
15151665	6	4.49	0.00	17-20dpf	0.15	znf664
15150634	1	0.61	0.02	8-10dpf	0.15	xpo4

15151075	7	-0.64	0.00	8-10dpf	0.15	atp8a1
15151905	1	0.61	0.02	8-10dpf	0.14	LOC107103455
15151905	1	0.88	0.03	8-10dpf	0.14	pxmp4
15150457	5	-0.64	0.00	8-10dpf	0.12	ppp1r13b
15150457	5	0.96	0.00	17-20dpf	0.12	ppp1r13b
15152211	5	0.68	0.04	17-20dpf	0.12	atf6b
15150651	2	-2.40	0.00	17-20dpf	0.12	LOC107087896
15150673	4	-0.86	0.00	8-10dpf	0.12	papd5
15151409	2	-1.00	0.04	8-10dpf	0.12	LOC107100292
15150711	10	-5.42	0.00	8-10dpf	0.12	fbxo32
15150711	8	-6.16	0.00	8-10dpf	0.12	klhl38
15150599	1	0.66	0.04	8-10dpf	0.11	snx29
15150825	1	4.68	0.00	8-10dpf	0.11	pkd1l1
15150825	1	1.34	0.00	8-10dpf	0.11	skida1
15150487	2	-0.83	0.00	8-10dpf	0.10	st7l
15150621	20	-1.51	0.01	17-20dpf	0.09	kcnab1
15150536	40	0.66	0.01	8-10dpf	0.09	eif2b3
15150536	1	-0.96	0.04	8-10dpf	0.09	LOC107083768
15150536	18	-1.04	0.02	8-10dpf	0.09	plk3
15150548	2	-0.56	0.05	17-20dpf	0.09	LOC107084243
15150538	1	-0.62	0.01	8-10dpf	0.09	fam172a
15150538	2	0.73	0.05	17-20dpf	0.09	rtkn
15150538	2	-0.87	0.01	8-10dpf	0.09	rtkn
15150453	3	-1.73	0.00	8-10dpf	0.08	LOC107084596
15150453	3	-0.85	0.05	17-20dpf	0.08	LOC107084596
15150453	3	-0.59	0.01	8-10dpf	0.08	LOC107084689
15151058	1	1.34	0.02	8-10dpf	0.08	LOC107096196
15150508	1	-0.72	0.04	17-20dpf	0.07	atic
15150508	1	0.65	0.01	8-10dpf	0.07	atic
15150480	1	-0.51	0.01	8-10dpf	0.07	tmem30a
15150463	1	-0.66	0.05	17-20dpf	0.07	stx5
15151111	1	-2.18	0.00	8-10dpf	0.07	LOC107096914
15151111	1	-3.83	0.00	17-20dpf	0.07	LOC107096914
15151111	1	-1.55	0.00	8-10dpf	0.07	LOC107096921
15150652	6	0.71	0.05	8-10dpf	0.07	LOC107087924
15151119	21	-2.06	0.00	17-20dpf	0.06	LOC107097014
15151119	21	-0.92	0.03	8-10dpf	0.06	LOC107097014
15151119	22	-0.92	0.00	8-10dpf	0.06	LOC107097016
15150922	2	-0.78	0.02	17-20dpf	0.06	LOC107094191

B2. Supplemental Figures

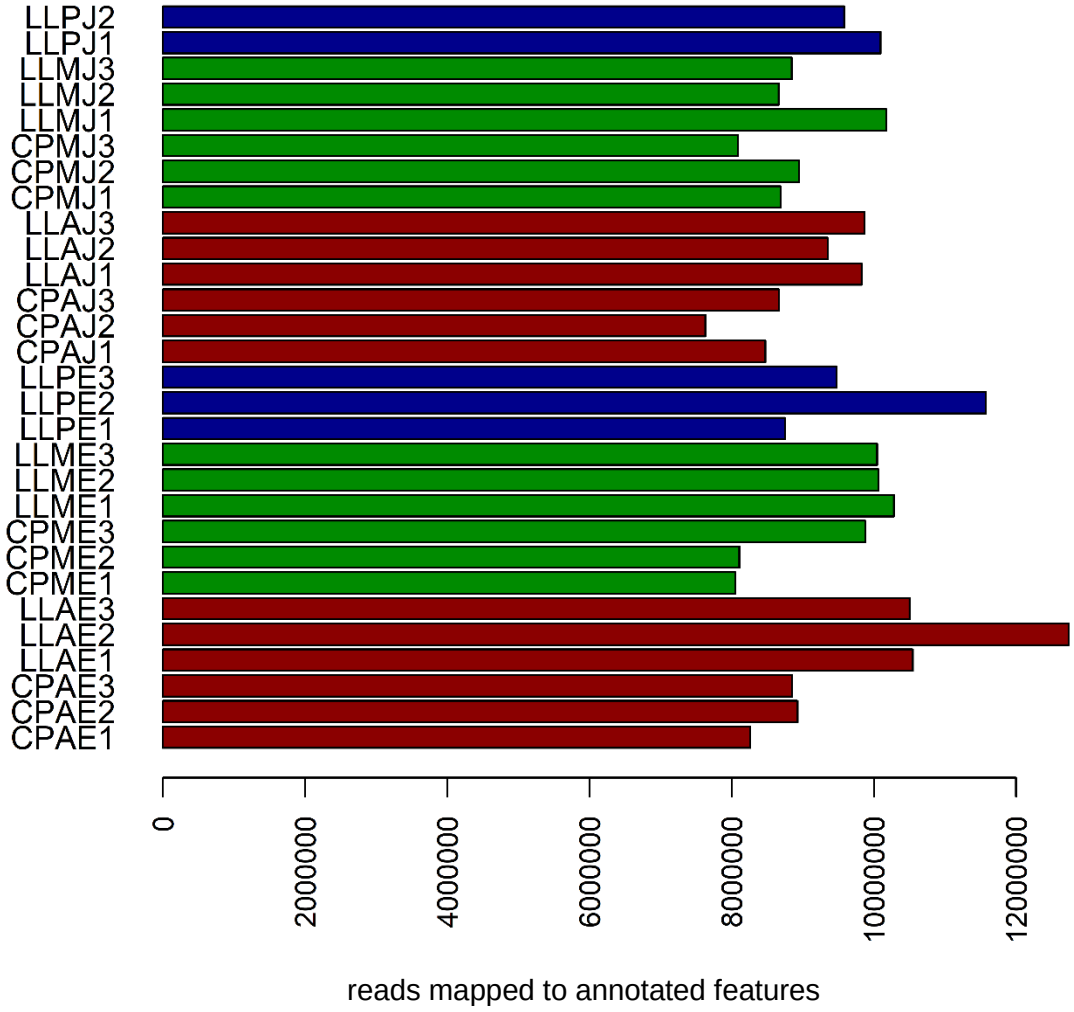

Figure B2.1. A similar number of reads map to annotated features across generalists, snail-eaters, and scale-eaters.

Generalists (red), snail-eaters (green), and scale-eaters (blue) (ANOVA; 8-10 dpf $p = 0.47$; 17-20 dpf $p = 0.33$). CP = Crescent Pond, LL = Little Lake, E = 8-10 dpf, J = 17-20 dpf).

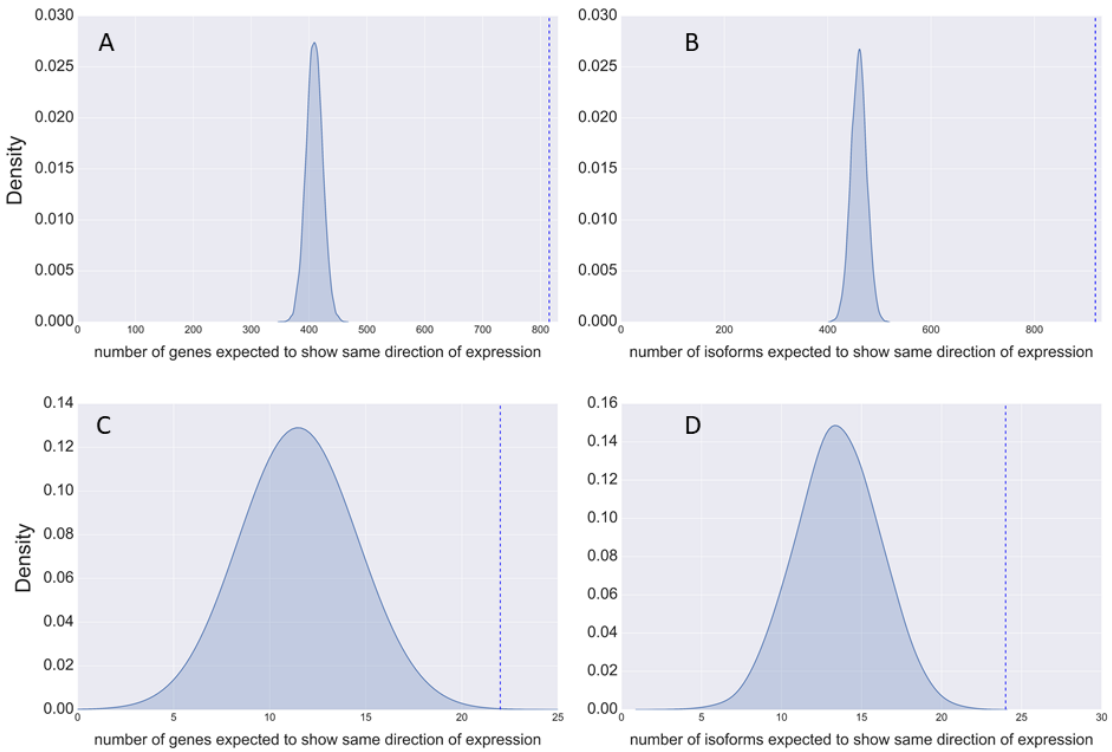

Figure B2.2. Null distributions of parallel changes in gene expression between specialists.

Kernel density plots show the null distribution for the number of genes (A and C) and isoforms (B and D) expected to show the same direction of expression in specialists relative to generalists. We performed 10,000 permutations sampling from a binomial distribution to estimate the expected number of genes and isoforms showing shared expression. The actual number of genes and isoforms showing shared directions of expression are indicated by blue dotted lines. A and B show distributions for gene and isoform expression at 8-10 dpf. C and D show distributions for gene and isoform expression at 17-20 dpf. Significantly more genes and isoforms show the same expression pattern in specialists relative to generalists at both developmental time points ($P < 1.0 \times 10^{-4}$).

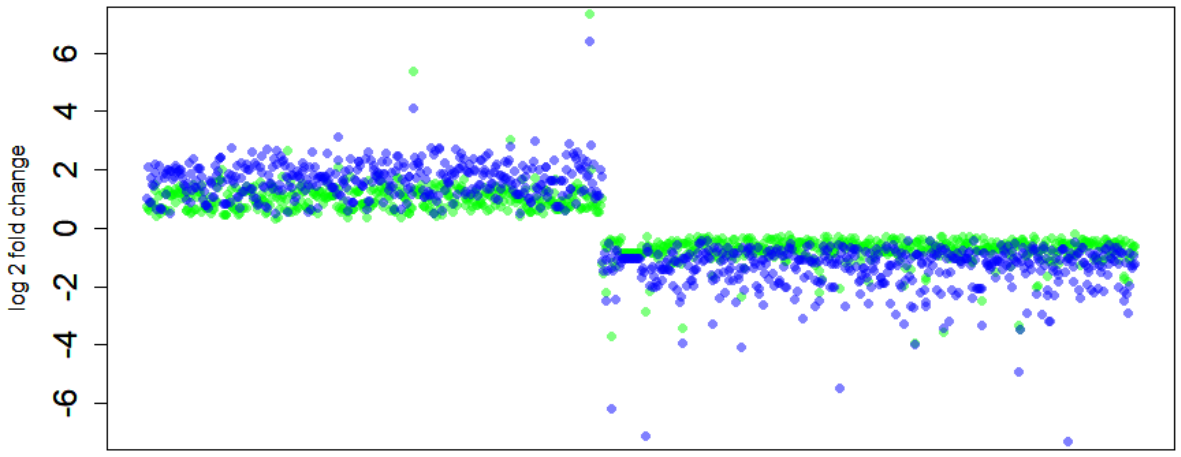

Figure B2.3. Parallel changes in isoform expression between specialists at 8-10 dpf.

497 differentially expressed isoforms showed lower expression in both specialist species compared to generalists, while 424 showed higher expression in specialists. Blue points indicate $\log_2$ fold change for genes differentially expressed between generalists and scale-eaters and green shows $\log_2$ fold change for genes differentially expressed between generalists and snail-eaters.

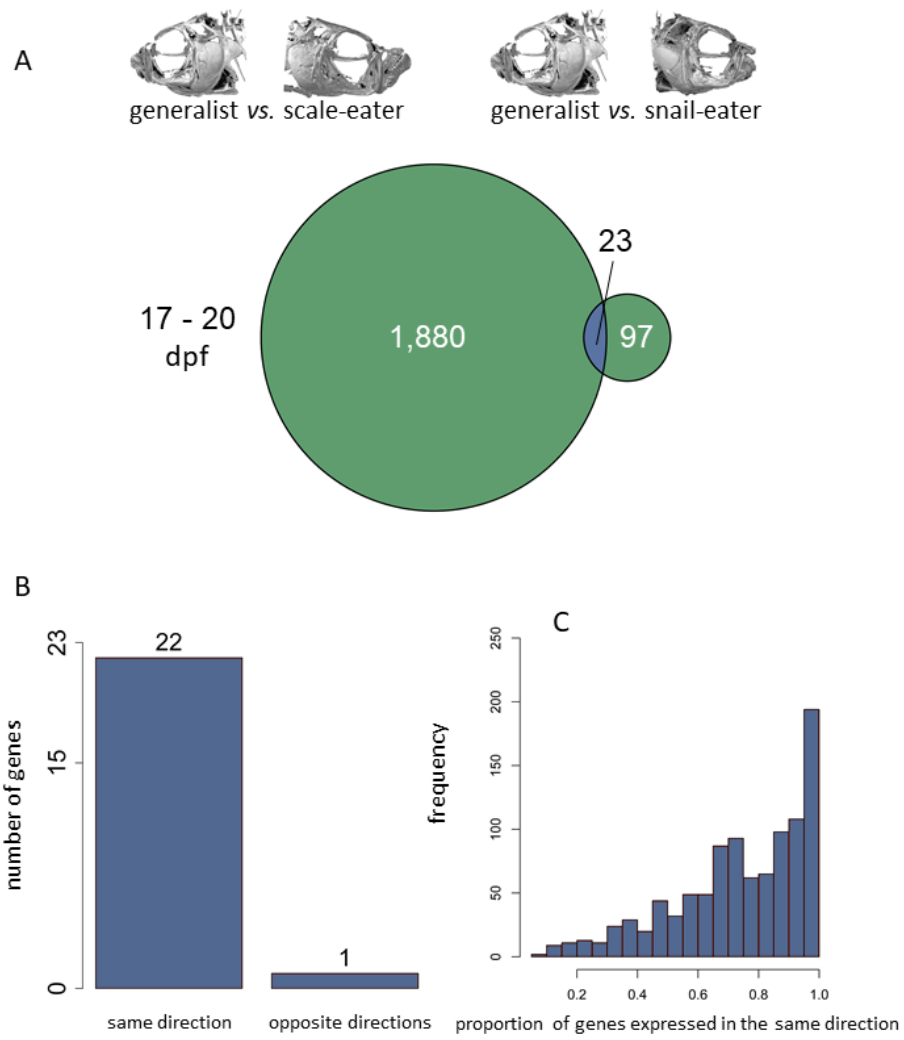

Figure B2.4. Significant parallel evolution of gene expression between specialists despite divergent trophic adaptation.

Circles illustrate genes differentially expressed in 17-20 dpf whole-body tissue for generalists *vs.* scale-eaters (left) and generalists *vs.* snail-eaters (right). Genes showing differential expression in both comparisons are shown in blue, and those showing divergent expression patterns unique to each specialist are green. Significantly more genes show differential expression in both comparisons than expected by chance (Fisher's exact test, $P < 1.0 \times 10^{-16}$). B) Significantly more genes show the same direction of expression in specialists relative to generalists than expected by chance (10,000 permutations; $P < 1.0 \times 10^{-4}$; Fig. B2.2). C) Distribution of the proportion of genes differentially expressed in the same direction between specialists relative to generalists after 1,000 down sampling permutations where groups of generalists and snail-eaters were randomly sampled to match scale-eater sample sizes (n = 2) show that parallel expression is robust to variation in sample size (median number of genes common to both comparisons = 16).

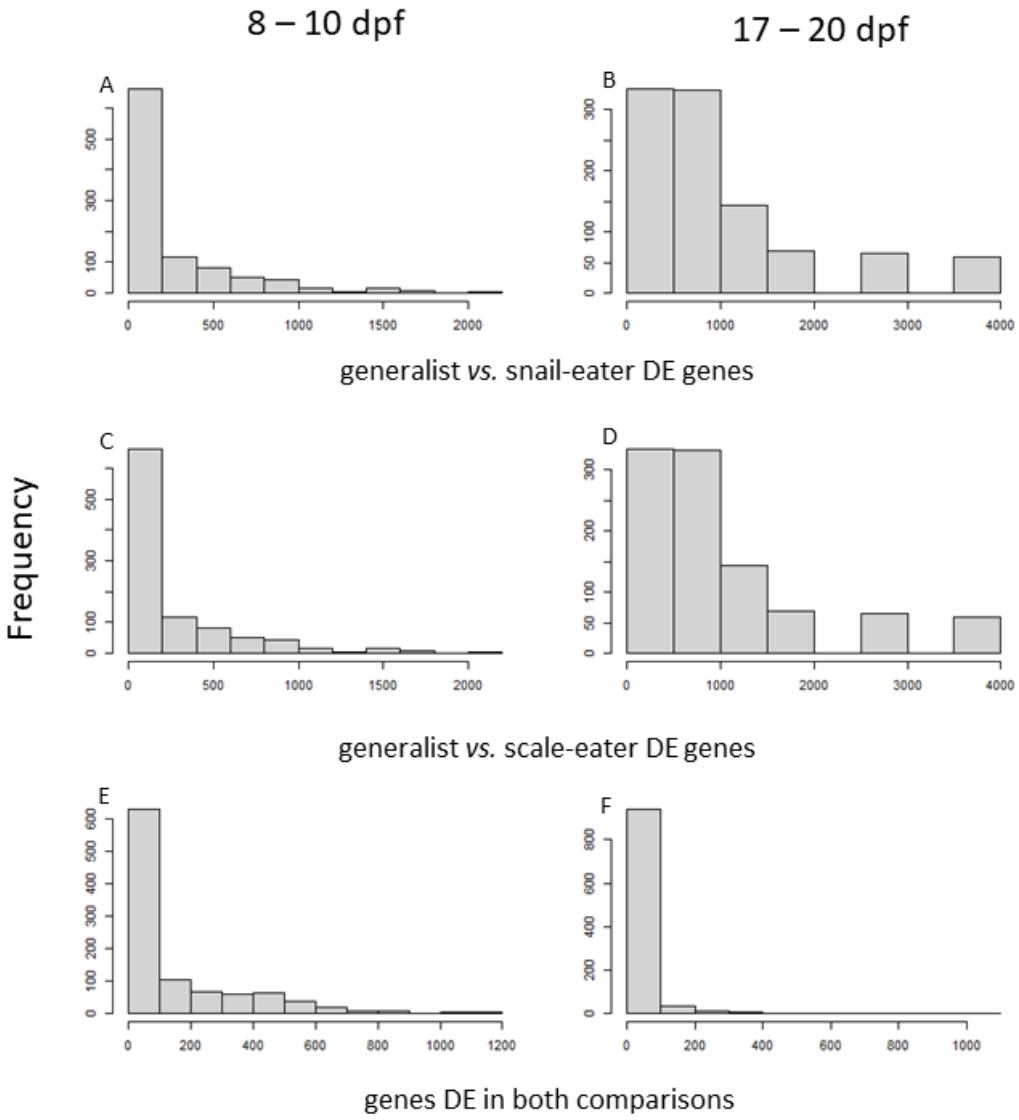

Figure B2.5. Down sampling permutations.

Distribution of genes differentially expressed (DE) between generalists and snail-eaters (A and B), generalists and scale-eaters (C, and D), and genes DE in both comparisons (E and F) for 8-10 dpf (left) and 17-20 dpf (right) samples after 1000 down sampling permutations where groups of generalists and snail-eaters were randomly sampled to match scale-eater sample sizes (8-10 dpf, n = 3; 17-20 dpf, n = 2).

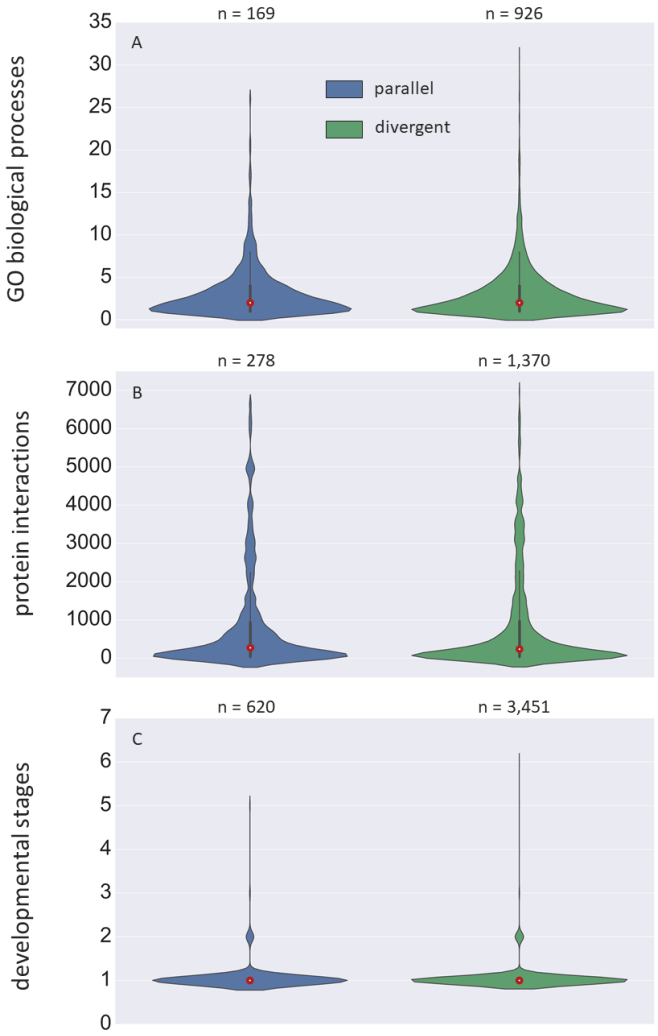

Figure B2.6. Genes showing parallel expression patterns in specialists are not more pleiotropic than genes showing divergent expression.

Violin plots show the distribution of pleiotropy estimates (GO biological processes (A), protein-protein interactions (B), and developmental stages expressed (C)) for genes showing parallel changes in expression (blue) and divergent changes in expression (green) between specialists relative to generalists at 8-10 dpf. Red dots show the median, thick black bars show interquartile ranges and thin bars show 95% confidence intervals. Genes showing parallel expression are not significantly more or less pleiotropic than divergently expressed genes (GLM; biological processes: $P = 0.67$; PPIs: $P = 0.09$; developmental stages: $P = 0.89$).

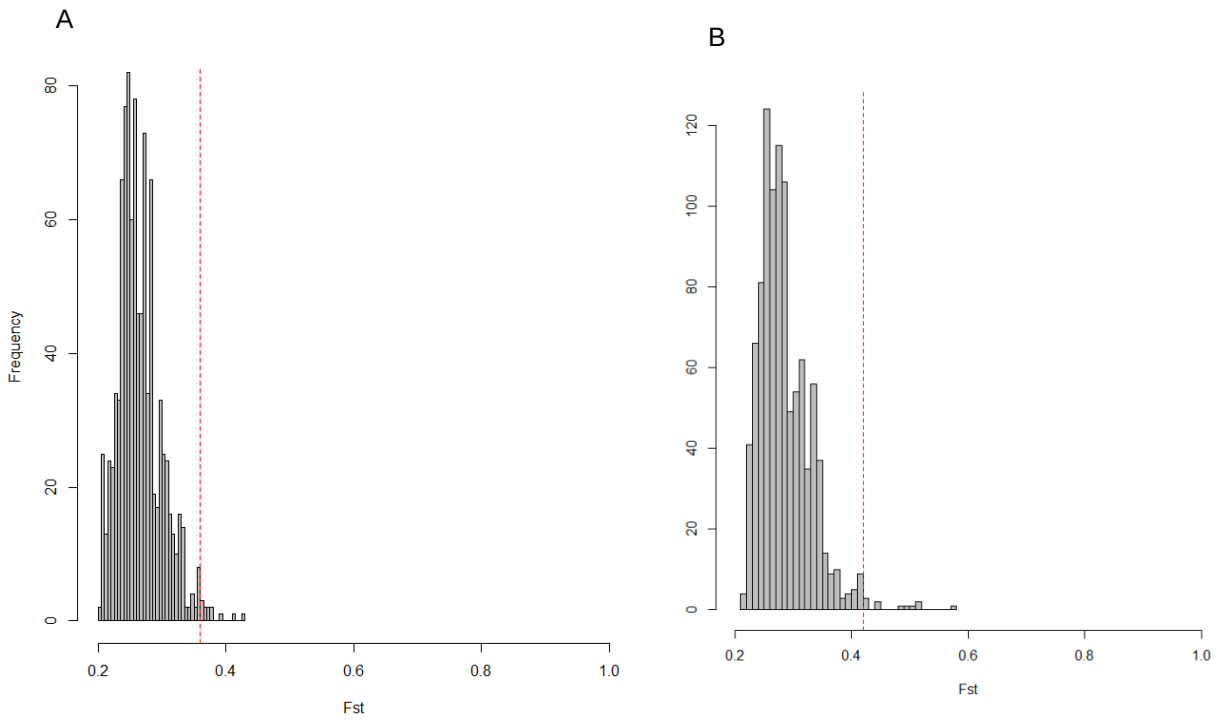

Figure B2.7. Fst permutations to determine significantly differentiated SNPs.

We performed 1,000 permutations calculating genome-wide F_{st} between randomly subsampled groups in order to identify non-randomly differentiated genomic regions between species A) 99th percentile estimates of F_{st} across all SNPs between randomly sampled generalists and snail-eaters (n = 13 vs. n = 11). B) 99th percentile estimates of F_{st} across all SNPs between randomly sampled generalists and scale-eaters (n = 13 vs. n = 9). We took the 99th percentile of these distributions to set a threshold defining significantly high divergence (red dotted lines; $F_{st} > 0.36$ for generalists *vs.* snail-eaters; $F_{st} > 0.42$ for generalists *vs.* scale-eaters).

APPENDIX C: SUPPLEMENTARY MATERIAL FOR CHAPTER 3

C1. Supplemental Tables

Table C1.1. mRNA sequencing design.

round	sequencing date	pooled across *n* lanes	library prep kit
1	4/17	1	KAPA stranded mRNA
2	6/17	1	TruSeq stranded mRNA
3	5/18	1	TruSeq stranded mRNA
4	7/18	3	TruSeq stranded mRNA

Table C1.2. Read statistics for samples.

sample	species	stage	sequencing round	library prep kit	raw fastq reads	reads mapped	raw counts	normalized counts
1	hybrid	17-20dpf	2	truseq	41912228	39531780	14471030	6134242
2	hybrid	17-20dpf	2	truseq	18451756	17360214	6363816	6577234
3	hybrid	17-20dpf	2	truseq	33541230	31461875	11473464	6386857
4	hybrid	17-20dpf	2	truseq	27720328	26006609	9659367	6234856
5	generalist	8dpf	3	truseq	25656702	24407630	10461159	8815459
6	generalist	8dpf	3	truseq	22804982	21721330	9245634	8017330
7	generalist	8dpf	3	truseq	26313696	25476498	10757268	9390970
8	molluscivore	8dpf	4	truseq	38287748	36204014	15203504	7457667
9	molluscivore	8dpf	4	truseq	34288848	32578838	13434770	7536493
10	molluscivore	8dpf	4	truseq	33962768	32384092	13443902	9051218
11	generalist	8-10dpf	1	kapa	23172714	17847934	6880522	7213838
12	generalist	8-10dpf	1	kapa	20575374	19452261	7933158	8206676
13	generalist	8-10dpf	1	kapa	20631366	19202893	7750909	7161844
14	generalist	17-20dpf	1	kapa	20743782	18496992	6837539	8501245
15	generalist	17-20dpf	1	kapa	18728520	16361277	6040194	9051398
16	generalist	17-20dpf	1	kapa	21338994	19399691	6922698	7940092
17	molluscivore	8-10dpf	1	kapa	19100066	17430230	6789911	7921670
18	molluscivore	8-10dpf	1	kapa	19479052	17376013	6715924	7812869
19	molluscivore	8-10dpf	1	kapa	23224142	21581058	8519810	8476166
20	molluscivore	17-20dpf	1	kapa	21012680	18765633	7182366	7853844
21	molluscivore	17-20dpf	1	kapa	20996520	19096064	7507215	7831725
22	molluscivore	17-20dpf	1	kapa	20731964	17371497	6216825	7522140
23	generalist	8-10dpf	1	kapa	26283022	23001257	8649498	8749038
24	generalist	8-10dpf	1	kapa	29483942	27652542	11273682	7540908
25	generalist	8-10dpf	1	kapa	26094366	22722751	8639044	8982363
26	generalist	17-20dpf	1	kapa	23539660	21193066	8288540	9080255
27	generalist	17-20dpf	1	kapa	22989146	20041508	7630051	7855706
28	generalist	17-20dpf	1	kapa	24875424	21412781	7819750	7254103
29	molluscivore	8-10dpf	1	kapa	25828344	22266723	8306859	7798548
30	molluscivore	8-10dpf	1	kapa	25463686	22026757	7881773	7685499
31	molluscivore	8-10dpf	1	kapa	24912808	21994615	8135992	8278350
32	molluscivore	17-20dpf	1	kapa	24703694	21871287	8360480	10038049
33	molluscivore	17-20dpf	1	kapa	21852694	18695831	6892571	10248139
34	molluscivore	17-20dpf	1	kapa	22560226	19029742	6997928	9514273
35	generalist	8dpf	3	truseq	25934770	24751899	10559980	8480276
36	generalist	8dpf	3	truseq	24781078	23652972	10100401	7407245
37	generalist	8dpf	3	truseq	23199342	22179263	9546582	9933070
38	molluscivore	8dpf	3	truseq	25699038	24831966	10647567	9278752
39	molluscivore	8dpf	3	truseq	31456730	30017500	12699299	11025018
40	molluscivore	8dpf	3	truseq	27239292	26189352	11032426	10043336

41	hybrid	8dpf	4	truseq	27989988	26774703	11506670	8202298
42	hybrid	8dpf	4	truseq	26341200	25153435	10759990	8361452
43	hybrid	8dpf	4	truseq	41450864	39962577	16950855	8441280

Table C1.3. Quality control statistics for samples.

sample	species	stage	median TIN	average depth across features	proportion of duplicate reads	median GC content across reads
1	hybrid	17-20dpf	48.63	169.59	6.80	46.42
2	hybrid	17-20dpf	41.90	127.99	8.15	46.50
3	hybrid	17-20dpf	32.68	165.40	4.61	45.74
4	hybrid	17-20dpf	50.43	138.23	8.79	45.99
5	generalist	8dpf	82.94	131.20	10.19	46.82
6	generalist	8dpf	82.77	129.37	10.46	47.14
7	generalist	8dpf	83.55	125.86	10.03	47.30
8	molluscivore	8dpf	81.01	139.18	14.19	46.22
9	molluscivore	8dpf	82.25	128.50	14.18	46.91
10	molluscivore	8dpf	82.59	125.39	13.67	48.03
11	generalist	8-10dpf	72.56	157.17	13.08	46.25
12	generalist	8-10dpf	73.65	145.73	13.00	45.42
13	generalist	8-10dpf	73.53	140.65	13.40	46.28
14	generalist	17-20dpf	68.89	144.59	13.97	45.36
15	generalist	17-20dpf	70.57	134.99	14.22	46.27
16	generalist	17-20dpf	63.81	155.01	13.60	44.83
17	molluscivore	8-10dpf	73.53	132.25	13.88	46.28
18	molluscivore	8-10dpf	74.69	125.74	14.05	46.78
19	molluscivore	8-10dpf	74.43	142.56	12.79	45.92
20	molluscivore	17-20dpf	73.09	132.20	14.22	46.03
21	molluscivore	17-20dpf	73.17	128.74	15.12	46.81
22	molluscivore	17-20dpf	71.57	138.66	13.06	47.44
23	generalist	8-10dpf	76.01	140.15	12.42	46.50
24	generalist	8-10dpf	75.82	154.90	12.05	45.65
25	generalist	8-10dpf	74.11	146.22	12.72	46.21
26	generalist	17-20dpf	76.56	129.96	14.25	45.57
27	generalist	17-20dpf	75.39	136.84	13.92	45.89
28	generalist	17-20dpf	76.83	127.75	13.48	45.58
29	molluscivore	8-10dpf	75.34	132.93	13.50	45.96
30	molluscivore	8-10dpf	76.29	130.14	12.95	46.38
31	molluscivore	8-10dpf	75.54	131.94	13.25	46.49
32	molluscivore	17-20dpf	74.48	142.25	14.33	45.64
33	molluscivore	17-20dpf	74.08	138.28	13.73	45.90
34	molluscivore	17-20dpf	75.39	129.94	13.65	46.43
35	generalist	8dpf	82.43	132.68	9.94	47.27
36	generalist	8dpf	82.69	125.78	10.59	47.47
37	generalist	8dpf	81.58	136.72	9.71	46.98
38	molluscivore	8dpf	81.63	135.55	9.91	47.33
39	molluscivore	8dpf	84.49	125.89	10.69	47.31

40	molluscivore	8dpf	84.31	118.45	10.35	47.61
41	hybrid	8dpf	80.98	134.41	12.59	47.33
42	hybrid	8dpf	81.02	130.78	12.12	46.98
43	hybrid	8dpf	82.94	142.90	11.04	47.51

Table C1.4. Differentially expressed genes annotated for effects on skeletal system morphogenesis.

This ontology (GO:0048705) was the only enriched biological process for genes differentially expressed between generalists and molluscivores at 8 dpf ($P < 0.05$; geneontology.org).

gene symbol	$\log_2$ fold change	*P*
bmp3	0.511242	0.039704
chd7	0.423654	0.047135
foxe1	-0.63748	0.004896
gata3	0.369094	0.043925
gfpt1	-0.29543	0.039977
hand2	0.639402	0.012518
kat6a	0.55044	0.000901
matn1	1.144529	0.049159
matn4	0.447086	0.000203
mecom	0.552098	0.023904
polr1c	-0.68794	0.026325

Table C1.5. Misregulated genes annotated for effects on embryonic cranial skeleton morphogenesis.

This ontology (GO:0048701) was one of 210 enriched biological processes for 6,590 genes differentially expressed between hybrids and parental species in craniofacial tissue collected at 17-20 dpf ($P < 0.05$; geneontology.org).

gene symbol	$\log_2$ fold change	*P*
alcam	0.610547	0.000222
alx1	-0.85427	0.033755
bmp3	0.583685	0.022167
crispld2	-0.94461	0.000382
dcaf7	-0.81576	9.25E-09
egr1	-1.18846	0.039205
fam20b	0.893436	3.84E-05
fgf3	0.707254	0.026311
foxe1	-0.86424	0.023926
fst	-1.04804	0.000342
gfpt1	1.260497	1.00E-14
gnptab	0.847555	0.000209
hand2	-1.56118	1.71E-05
irf6	-1.14833	2.02E-09
itga8	-0.48604	0.037746
kat6a	-1.09119	7.80E-05
kdm6a	-0.57079	0.005865
kras	0.855153	2.61E-05
leo1	-0.47694	0.012396
mapre2	1.453761	2.81E-11
mecom	-1.08419	0.017119
med12	-1.45818	2.03E-15
med14	0.758638	0.000319
ocrl	1.025415	4.44E-05
pak1	0.960502	0.000228
pbx4	1.852615	2.31E-12
pdgfra	-0.48287	0.048917
phf8	0.493499	0.007445
pitx2	0.807376	0.014345
polr1d	0.721823	0.032707
rnf2	-1.12964	3.91E-05
runx3	1.718195	3.94E-18
s1pr2	0.415986	0.024007
scfd1	0.299537	0.028434
sec23a	-1.43718	3.77E-11

sec24d	0.910752	0.005705
sharpin	-1.43559	2.67E-06
shh	1.07417	0.004793
smo	-0.46599	0.039733
sphk2	0.967353	7.29E-07
tfap2a	0.81934	0.008711
tshz2	-0.68911	0.000809
wls	-1.49603	4.15E-07
wnt4	0.983625	0.020119
xylt1	0.661423	0.009892

Table C1.6. Gene ontologies enriched for 6,590 genes misregulated between hybrids and parental species in craniofacial tissue collected at 17-20 dpf.

GO:0002181	GO:0006417	GO:0044257	GO:0150063	GO:0071840
GO:0042255	GO:0006364	GO:0033554	GO:0009056	GO:0051246
GO:0042273	GO:0043603	GO:0034613	GO:1901575	GO:0006950
GO:0006402	GO:0034248	GO:0070727	GO:0048562	GO:0043412
GO:0051236	GO:0033365	GO:0016567	GO:0010605	GO:0036211
GO:0006412	GO:0006396	GO:0030163	GO:1901137	GO:0006464
GO:0000956	GO:0048701	GO:0046700	GO:0034622	GO:0022607
GO:0050658	GO:0010467	GO:0008104	GO:0009790	GO:0016043
GO:0050657	GO:1904888	GO:0046907	GO:0034654	GO:0051173
GO:0043043	GO:0034470	GO:0090304	GO:0044267	GO:0010604
GO:0015931	GO:0016072	GO:1901361	GO:0048598	GO:0009893
GO:0006401	GO:0048704	GO:0071705	GO:0044260	GO:0009888
GO:0034976	GO:0010498	GO:0044270	GO:0048568	GO:0048856
GO:0006403	GO:0070647	GO:0034641	GO:0048880	GO:0007275
GO:0006366	GO:0034660	GO:0006325	GO:0019438	GO:0048731
GO:0006605	GO:0043161	GO:0044249	GO:0018130	GO:0032502
GO:0007034	GO:0034655	GO:0051649	GO:0006082	GO:0019222
GO:0022618	GO:0044265	GO:1901576	GO:0044237	GO:0009653
GO:0072594	GO:0009059	GO:0001501	GO:0006807	GO:0010468
GO:0043604	GO:0034645	GO:0009058	GO:0009887	GO:0060255
GO:0090150	GO:0006886	GO:0006139	GO:0043170	GO:0031323
GO:0006518	GO:0006520	GO:0043009	GO:0031324	GO:0006810
GO:0071826	GO:0010608	GO:0006725	GO:0045935	GO:0051234
GO:0016197	GO:0043632	GO:0009792	GO:1901135	GO:0051171
GO:0016579	GO:0019941	GO:0046483	GO:0007423	GO:0080090
GO:0051169	GO:0048193	GO:0033036	GO:0006508	GO:0051179
GO:0006913	GO:1901566	GO:0018193	GO:0044238	GO:0009987
GO:0016570	GO:0015031	GO:0051641	GO:0044281	GO:0032501
GO:0042254	GO:0048705	GO:1901360	GO:0008152	GO:0008150
GO:0022613	GO:0048706	GO:0051276	GO:1901362	GO:0050896
GO:0016569	GO:0015833	GO:0006259	GO:0051172	GO:0007165
GO:0000398	GO:0044271	GO:0002520	GO:0071704	GO:0007154
GO:0000377	GO:0006511	GO:0044248	GO:0019538	GO:0023052
GO:0000375	GO:0045184	GO:0010629	GO:1901564	GO:0050877
GO:0070646	GO:0032446	GO:0048534	GO:0016192	GO:0006955
GO:0061919	GO:0009057	GO:0030097	GO:0044085	GO:0046777
GO:0006914	GO:0042886	GO:0019752	GO:0065003	GO:0099537
GO:0008380	GO:0016070	GO:0071702	GO:0006996	GO:0099536
GO:0017038	GO:0006974	GO:1901565	GO:0055114	GO:0007268
GO:0016071	GO:0006281	GO:0009892	GO:0032268	GO:0098916
GO:0006397	GO:0019439	GO:0043436	GO:0043933	GO:0007187
GO:0006457	GO:0051603	GO:0001654	GO:0048513	GO:0007186

C2. Supplemental Figures

Figure C2.1. 20 day old generalist (top) and molluscivore (bottom).

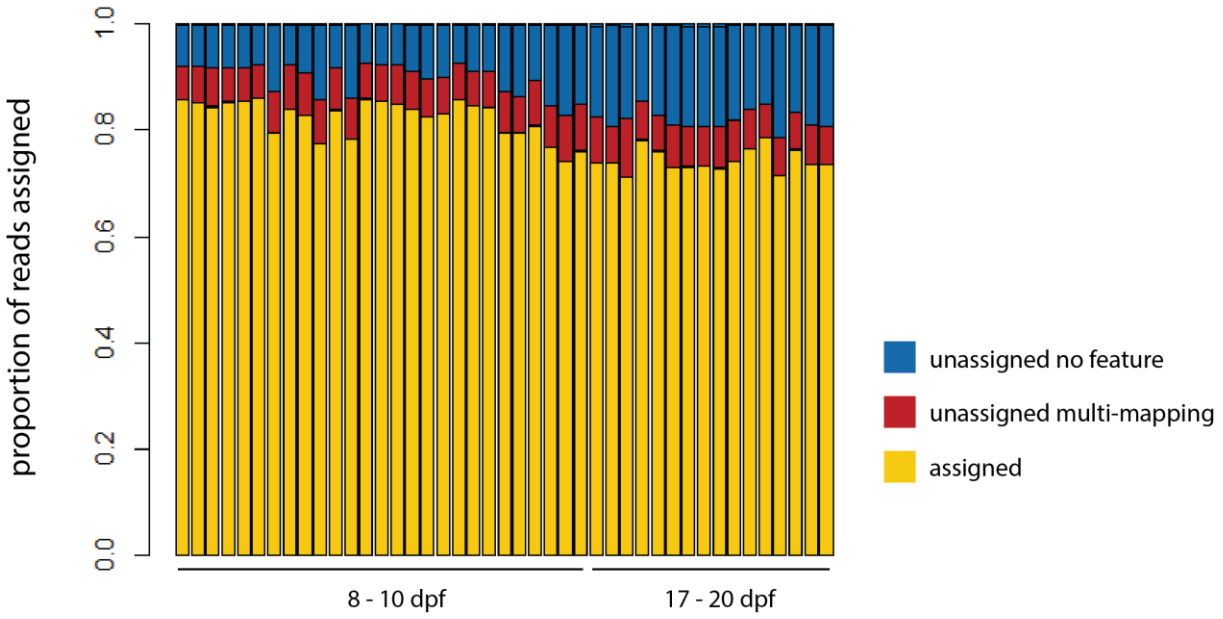

Figure C2.2. Read statistics for samples.

Proportion of reads assigned to features (yellow), unassigned due to multi-mapping (red), and unassigned due to no match to annotated features (blue) using STAR aligner.

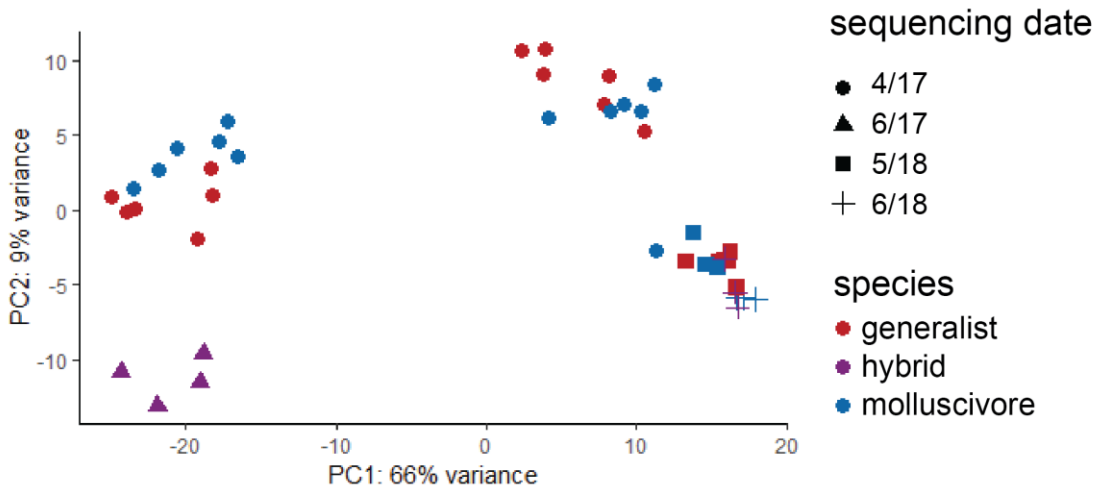

Figure C2.3. The first and second principal component axes accounting for a combined 75% of the total variation between generalist, molluscivore, and hybrid samples across reads mapped to annotated features.

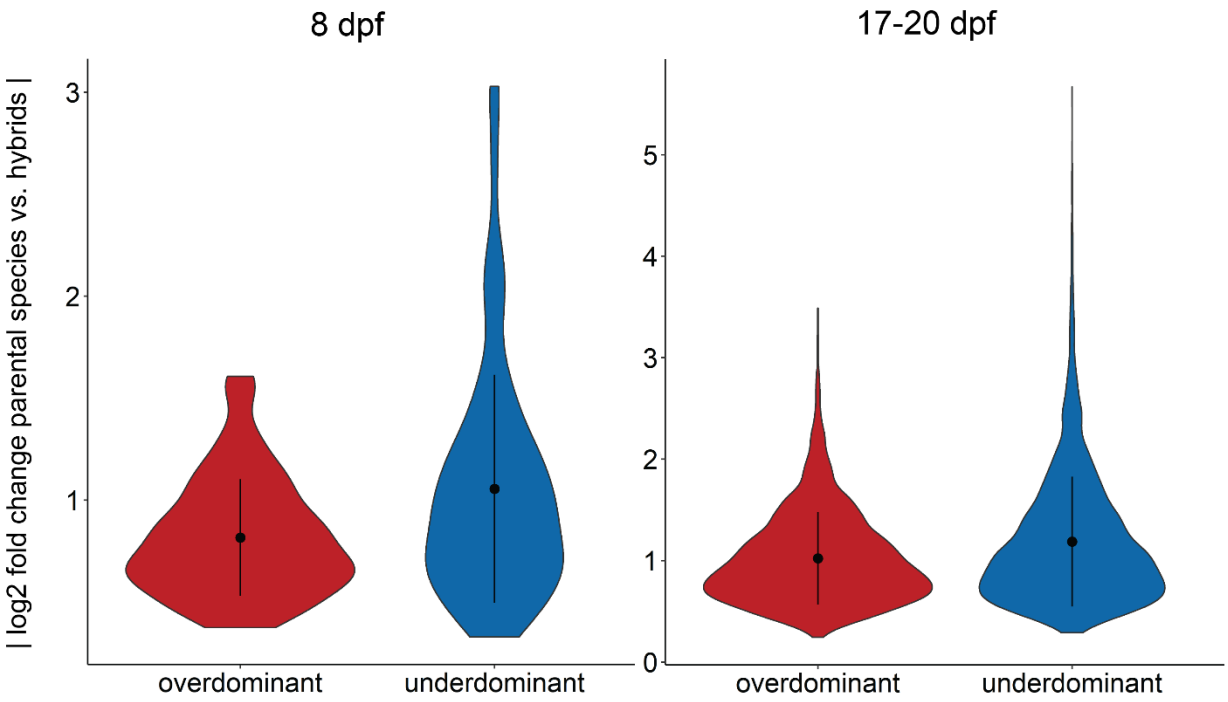

Figure C2.4. Genes showing underdominant expression in hybrids show a higher magnitude of misregulation than genes showing overdominance.

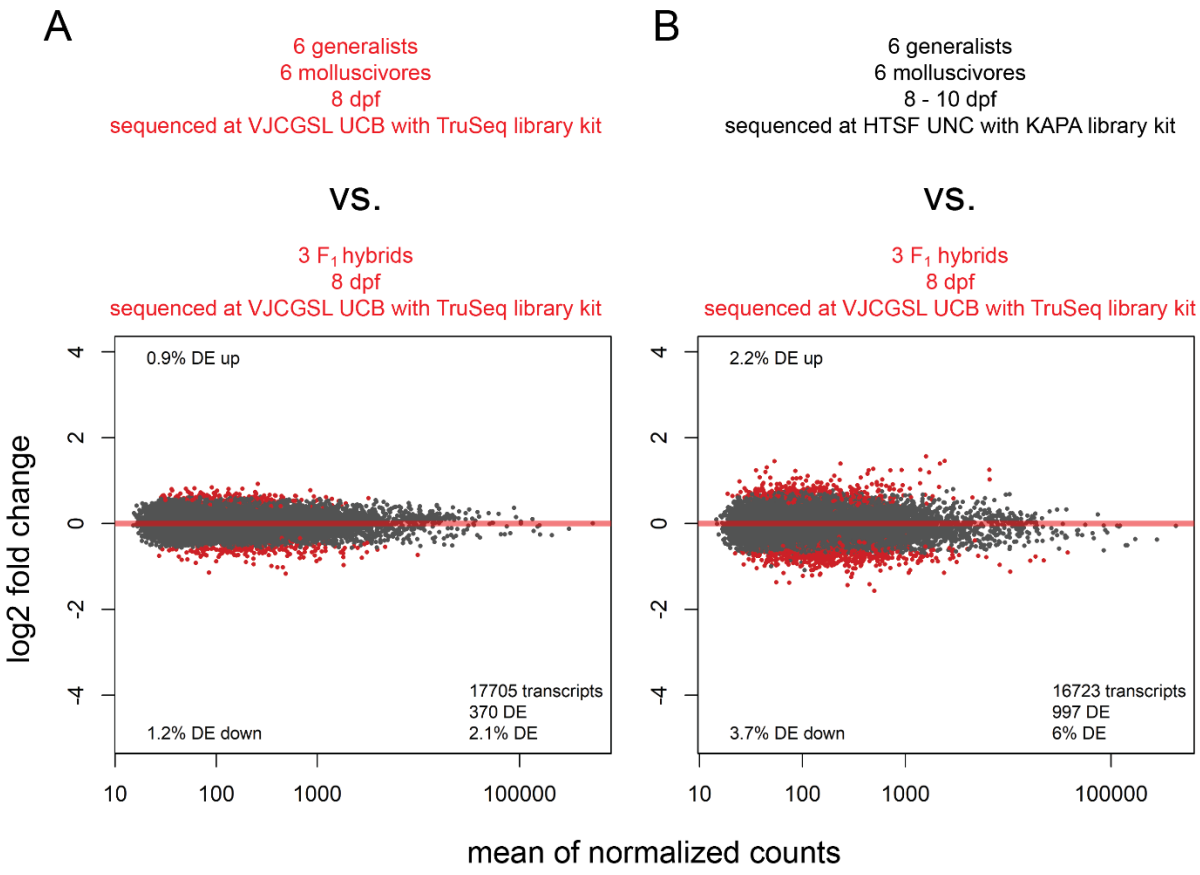

Figure C2.5. Estimating the effect of sequencing design on the proportion of genes misregulated in hybrids.

The 8 dpf hybrids were sequenced at the same facility with the same library kit as the 17-20 dpf hybrids, while the 8-10 dpf parental species were sequenced at the same facility with the same library kit as the 17-20 dpf parental species. A) The comparison between 8 dpf parental species and 8 dpf hybrids revealed 370 genes (2.1%) misregulated. B) The comparison between 8 dpf hybrids and 8-10 dpf parental species revealed 997 (6%) genes misregulated – a 37% increase. We used this inflated estimate to adjust our estimate of misregulation in 17-20 dpf hybrid craniofacial tissues. Red points indicate genes detected as differentially expressed at 5% false discovery rate with Benjamini-Hochberg multiple testing adjustment. Grey points indicate genes showing no significant difference in expression between groups.

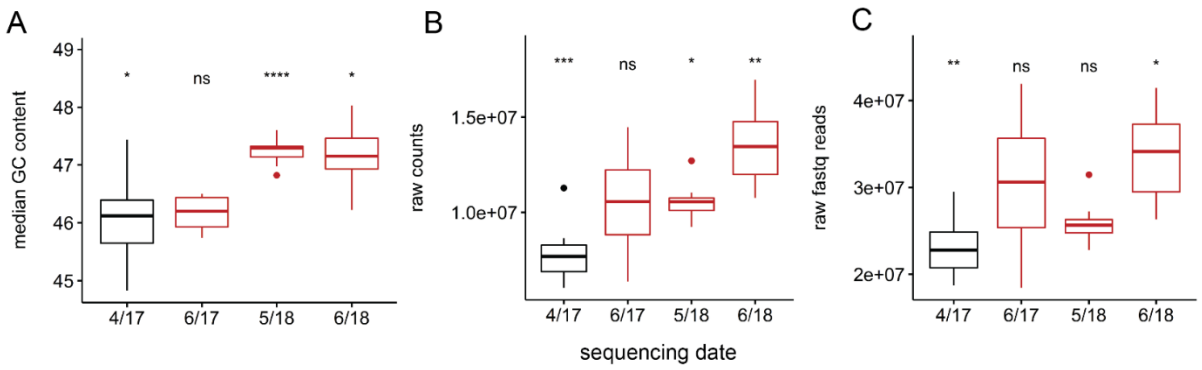

Figure C2.6. No significant differences between 17-20 dpf hybrid craniofacial samples and samples sequenced on other dates for quality control measures.

We did not find significant differences between 17-20 dpf hybrid craniofacial samples and samples sequenced on other dates for A) median percent GC content across reads, B) number of normalized read counts, or C) number of raw fastq reads. the proportion of duplicate reads for each sample (Pairwise Welch two sample t test; $P < 0.0001$ = ****, *** = 0.001, ** = 0.01, * = 0.05).

APPENDIX D: SUPPLEMENTARY MATERIAL FOR CHAPTER 4

D1. Supplemental Tables

Table D1.1. Cross design for 124 transcriptomes.

sample ID	stage	sequencing date	parents
CAE1	8dpf	May-18	Crescent Pond generalists
CAE2	8dpf	May-18	Crescent Pond generalists
CAE3	8dpf	May-18	Crescent Pond generalists
CAE4	8dpf	May-18	Crescent Pond generalists
CAE5	8dpf	May-18	Crescent Pond generalists
CME1	8dpf	Jul-18	Crescent Pond snail eaters
CME2	8dpf	Jul-18	Crescent Pond snail eaters
CME5	8dpf	Jul-18	Crescent Pond snail eaters
CPE1	8dpf	May-18	Crescent Pond scale eaters
CPE2	8dpf	May-18	Crescent Pond scale eaters
CPE3	8dpf	May-18	Crescent Pond scale eaters
CPE4	8dpf	May-18	Crescent Pond scale eaters
CPE5	8dpf	May-18	Crescent Pond scale eaters
CQE1	8dpf	Jul-18	New Providence female x New Providence generalist male
CQE2	8dpf	Jul-18	New Providence female x New Providence generalist male
CQE3	8dpf	Jul-18	New Providence female x New Providence generalist male
NCE1	8dpf	May-18	North Carolina generalists
NCE2	8dpf	May-18	North Carolina generalists
NCE3	8dpf	May-18	North Carolina generalists
NCE4	8dpf	May-18	North Carolina generalists
NCE5	8dpf	May-18	North Carolina generalists
OAE1	8dpf	May-18	Osprey Lake generalists
OAE2	8dpf	May-18	Osprey Lake generalists
OAE3	8dpf	May-18	Osprey Lake generalists
OAE4	8dpf	May-18	Osprey Lake generalists
OME1	8dpf	May-18	Osprey Lake snail eaters
OME2	8dpf	May-18	Osprey Lake snail eaters
OME3	8dpf	May-18	Osprey Lake snail eaters
OME4	8dpf	May-18	Osprey Lake snail eaters
OME5	8dpf	May-18	Osprey Lake snail eaters
OPE1	8dpf	May-18	Osprey Lake scale eaters
OPE2	8dpf	May-18	Osprey Lake scale eaters
OPE3	8dpf	May-18	Osprey Lake scale eaters
OPE4	8dpf	May-18	Osprey Lake scale eaters
OPE5	8dpf	May-18	Osprey Lake scale eaters
CPU1	8dpf	Jul-18	Crescent Pond generalist female x Crescent Pond snail eater male
CPU3	8dpf	Jul-18	Crescent Pond generalist female x Crescent Pond snail eater male

CPU5	8dpf	Jul-18	Crescent Pond generalist female x Crescent Pond snail eater male
CVE1	8dpf	Jul-18	Crescent Pond generalist female x Crescent Pond scale eater male
CVE2	8dpf	Jul-18	Crescent Pond generalist female x Crescent Pond scale eater male
CVE5	8dpf	Jul-18	Crescent Pond generalist female x Crescent Pond scale eater male
CWE2	8dpf	Jul-18	Crescent Pond snail eater female x Crescent Pond scale eater male
CWE3	8dpf	Jul-18	Crescent Pond snail eater female x Crescent Pond scale eater male
CWE4	8dpf	Jul-18	Crescent Pond snail eater female x Crescent Pond scale eater male
CXE2	8dpf	Jul-18	Crescent Pond snail eater female x Crescent Pond generalist male
CXE3	8dpf	Jul-18	Crescent Pond snail eater female x Crescent Pond generalist male
CXE4	8dpf	Jul-18	Crescent Pond snail eater female x Crescent Pond generalist male
NAE1	8dpf	Jul-18	North Carolina female x Crescent Pond generalist male
NAE2	8dpf	Jul-18	North Carolina female x Crescent Pond generalist male
NAE4	8dpf	Jul-18	North Carolina female x Crescent Pond generalist male
OUE1	8dpf	Jul-18	Osprey Lake generalist female x Osprey Lake snail eater male
OUE3	8dpf	Jul-18	Osprey Lake generalist female x Osprey Lake snail eater male
OUE4	8dpf	Jul-18	Osprey Lake generalist female x Osprey Lake snail eater male
OVE1	8dpf	Jul-18	Osprey Lake generalist female x Osprey Lake scale eater male
OVE4	8dpf	Jul-18	Osprey Lake generalist female x Osprey Lake scale eater male
OVE5	8dpf	Jul-18	Osprey Lake generalist female x Osprey Lake scale eater male
OXE2	8dpf	Jul-18	Osprey Lake snail eater female x Osprey Lake generalist male
OYE1	8dpf	May-18	Osprey Lake scale eater female x Osprey Lake generalist male
OYE2	8dpf	May-18	Osprey Lake scale eater female x Osprey Lake generalist male
OYE3	8dpf	May-18	Osprey Lake scale eater female x Osprey Lake generalist male
OYE4	8dpf	May-18	Osprey Lake scale eater female x Osprey Lake generalist male
OYE5	8dpf	May-18	Osprey Lake scale eater female x Osprey Lake generalist male
OZE2	8dpf	Jul-18	Osprey Lake scale eater female x Osprey Lake snail eater male
OZE4	8dpf	Jul-18	Osprey Lake scale eater female x Osprey Lake snail eater male
OZE5	8dpf	Jul-18	Osprey Lake scale eater female x Osprey Lake snail eater male
PAE1	8dpf	Jul-18	New Providence female x Crescent Pond generalist
PAE2	8dpf	Jul-18	New Providence female x Crescent Pond generalist
PAE5	8dpf	Jul-18	New Providence female x Crescent Pond generalist
CAT1	2dpf	May-18	Crescent Pond generalists
CAT2	2dpf	May-18	Crescent Pond generalists
CAT3	2dpf	May-18	Crescent Pond generalists
CMT1	2dpf	Jul-18	Crescent Pond snail eaters
CMT2	2dpf	Jul-18	Crescent Pond snail eaters
CMT3	2dpf	Jul-18	Crescent Pond snail eaters
CPT1	2dpf	May-18	Crescent Pond scale eaters
CPT2	2dpf	May-18	Crescent Pond scale eaters
CPT3	2dpf	Jul-18	Crescent Pond scale eaters
CQT1	2dpf	Jul-18	New Providence female x New Providence generalist male
CQT2	2dpf	Jul-18	New Providence female x New Providence generalist male
NCT1	2dpf	May-18	North Carolina generalists
NCT2	2dpf	May-18	North Carolina generalists
NCT3	2dpf	May-18	North Carolina generalists

OAT1	2dpf	May-18	Osprey Lake generalists
OAT2	2dpf	May-18	Osprey Lake generalists
OAT3	2dpf	Jul-18	Osprey Lake generalists
OMT1	2dpf	May-18	Osprey Lake snail eaters
OMT2	2dpf	May-18	Osprey Lake snail eaters
OMT3	2dpf	May-18	Osprey Lake snail eaters
OPT1	2dpf	May-18	Osprey Lake scale eaters
OPT2	2dpf	May-18	Osprey Lake scale eaters
OPT3	2dpf	May-18	Osprey Lake scale eaters
CUT1	2dpf	Jul-18	Crescent Pond generalist female x Crescent Pond snail eater male
CUT2	2dpf	Jul-18	Crescent Pond generalist female x Crescent Pond snail eater male
CUT3	2dpf	Jul-18	Crescent Pond generalist female x Crescent Pond snail eater male
CVT1	2dpf	May-18	Crescent Pond generalist female x Crescent Pond scale eater male
CVT2	2dpf	May-18	Crescent Pond generalist female x Crescent Pond scale eater male
CVT3	2dpf	May-18	Crescent Pond generalist female x Crescent Pond scale eater male
CWT1	2dpf	May-18	Crescent Pond snail eater female x Crescent Pond scale eater male
CWT2	2dpf	May-18	Crescent Pond snail eater female x Crescent Pond scale eater male
CWT3	2dpf	May-18	Crescent Pond snail eater female x Crescent Pond scale eater male
CXT1	2dpf	May-18	Crescent Pond snail eater female x Crescent Pond generalist male
CXT2	2dpf	May-18	Crescent Pond snail eater female x Crescent Pond generalist male
CXT3	2dpf	May-18	Crescent Pond snail eater female x Crescent Pond generalist male
NAT1	2dpf	May-18	North Carolina female x Crescent Pond generalist male
NAT2	2dpf	May-18	North Carolina female x Crescent Pond generalist male
NAT3	2dpf	May-18	North Carolina female x Crescent Pond generalist male
OUT1	2dpf	Jul-18	Osprey Lake generalist female x Osprey Lake snail eater male
OUT2	2dpf	Jul-18	Osprey Lake generalist female x Osprey Lake snail eater male
OUT3	2dpf	Jul-18	Osprey Lake generalist female x Osprey Lake snail eater male
OVT1	2dpf	May-18	Osprey Lake generalist female x Osprey Lake scale eater male
OVT2	2dpf	May-18	Osprey Lake generalist female x Osprey Lake scale eater male
OVT3	2dpf	May-18	Osprey Lake generalist female x Osprey Lake scale eater male
OXT1	2dpf	Jul-18	Osprey Lake snail eater female x Osprey Lake generalist male
OXT2	2dpf	Jul-18	Osprey Lake snail eater female x Osprey Lake generalist male
OXT3	2dpf	Jul-18	Osprey Lake snail eater female x Osprey Lake generalist male
OYT1	2dpf	May-18	Osprey Lake scale eater female x Osprey Lake generalist male
OYT2	2dpf	May-18	Osprey Lake scale eater female x Osprey Lake generalist male
OYT3	2dpf	May-18	Osprey Lake scale eater female x Osprey Lake generalist male
OZT1	2dpf	Jul-18	Osprey Lake scale eater female x Osprey Lake snail eater male
OZT2	2dpf	Jul-18	Osprey Lake scale eater female x Osprey Lake snail eater male
OZT3	2dpf	Jul-18	Osprey Lake scale eater female x Osprey Lake snail eater male
PAT1	2dpf	May-18	New Providence female x Crescent Pond generalist
PAT2	2dpf	May-18	New Providence female x Crescent Pond generalist
PAT3	2dpf	May-18	New Providence female x Crescent Pond generalist

Table D1.2. San Salvador Island within population genomic statistics measured across 13.8 million SNPs.

Statistics for the top three rows were calculated for all San Salvador individuals of each species (see Fig. D2.1). The remaining rows are comparisons separated by lake populations used to generate samples for RNAseq (CP = Crescent Pond, OL = Osprey Lake).

population	mean Tajima's D	Tajima's D 10th percentile	mean π
all generalists	0.704649	-0.90273	0.003029
all molluscivores	0.565385	-1.34112	0.002583
all scale-eaters	0.210182	-1.62616	0.002036
CP generalists	0.430683	-1.076	0.002806
CP molluscivores	0.097742	-1.44811	0.00194
CP scale-eaters	-0.01537	-1.53413	0.001385
OL generalists	0.338391	-0.77476	0.003022
OL molluscivores	0.227443	-1.37104	0.002458
OL scale-eaters	0.14957	-1.31009	0.00219

Table D1.3. San Salvador Island between population genomic statistics measured across 13.8 million SNPs.

Statistics for the top three rows were calculated for all San Salvador individuals of each species (see Fig. D2.1). The remaining rows are comparisons separated by lake populations used to generate samples for RNAseq (CP = Crescent Pond, OL = Osprey Lake).

population 1	**n**	**population 2**	**n**	**mean D_{xy}**	**D_{xy} 90th percentile**	**mean F_{st}**	**# fixed SNPs**
all generalists	8	all molluscivores	10	0.0047	0.0076	0.0564	179
all generalists	8	all scale-eaters	9	0.0047	0.0080	0.1065	5,331
all molluscivores	10	all scale-eaters	9	0.0049	0.0085	0.1357	36,335
CP generalists	5	CP molluscivores	5	0.0042	0.0075	0.0740	11,015
CP generalists	5	CP scale-eaters	5	0.0046	0.0082	0.1356	109,072
CP molluscivores	5	CP scale-eaters	5	0.0048	0.0093	0.1839	559,728
OL generalists	3	OL molluscivores	5	0.0049	0.0084	0.0964	47,356
OL generalists	3	OL scale-eaters	4	0.0049	0.0084	0.1130	108,813
OL molluscivores	5	OL scale-eaters	4	0.0049	0.0087	0.1347	168,192
CP generalists	5	OL generalists	3	0.0049	0.0082	0.0759	19,582
CP molluscivores	5	OL molluscivores	5	0.0045	0.0082	0.1169	92,317
CP scale-eaters	5	OL scale-eaters	4	0.0035	0.0073	0.0983	86,367

Table D1.4. Percentage of genes controlled by different regulatory mechanisms for each hybrid cross.

Informative genes are those containing heterozygous sites in hybrids that were alternatively homozygous in parents. The final column is the percentage of misregulated genes showing no difference in expression between parental populations and allele-specific expression in F1 hybrids, consistent with compensatory regulatory divergence. NC = North Carolina, NP = New Providence, CP = Crescent Pond, OL = Osprey Lake.

mother	father	stage	conserved	*cis*	*trans*	compensatory	misregulated	misregulated showing compensatory
NC generalist	CP generalist	2dpf	61.18	2.66	0.37	19.98	15.81	32.75
NP generalist	CP generalist	2dpf	79.57	0.34	0.42	16.45	3.22	11.84
CP generalist	CP molluscivore	2dpf	60.82	0.18	0.26	33.70	5.03	37.50
CP generalist	CP scale-eater	2dpf	83.50	0.17	0.17	15.87	0.28	40.00
CP molluscivore	CP scale-eater	2dpf	69.79	1.64	0.55	26.38	1.64	33.33
CP molluscivore	CP generalist	2dpf	62.79	0.14	0.05	36.07	0.96	57.14
OL generalist	OL molluscivore	2dpf	46.66	0.03	0.06	34.20	19.04	38.95
OL generalist	OL scale-eater	2dpf	62.77	0.05	0.52	22.75	13.91	18.96
OL scale-eater	OL molluscivore	2dpf	74.09	1.03	0.69	23.39	0.80	21.43
OL molluscivore	OL generalist	2dpf	59.79	0.03	0.03	37.29	2.85	38.55
OL scale-eater	OL generalist	2dpf	57.72	0.21	0.59	29.66	11.82	31.32
NC generalist	CP generalist	8dpf	60.13	1.40	0.07	8.41	29.98	13.47
NP generalist	CP generalist	8dpf	93.24	0.21	0.71	5.41	0.43	50.00
CP generalist	CP molluscivore	8dpf	87.06	0.12	0.12	9.77	2.93	20.83
CP generalist	CP scale-eater	8dpf	81.17	0.26	0.44	6.63	11.51	13.64
CP molluscivore	CP scale-eater	8dpf	78.87	1.85	2.04	4.48	12.76	8.40
CP molluscivore	CP generalist	8dpf	88.55	0.08	0.15	10.55	0.68	33.33
OL generalist	OL molluscivore	8dpf	75.26	0.45	0.61	7.19	16.49	9.63
OL generalist	OL scale-eater	8dpf	85.62	0.24	1.57	3.38	9.19	13.68
OL scale-eater	OL molluscivore	8dpf	90.24	0.81	0.61	6.20	2.13	14.29
OL scale-eater	OL generalist	8dpf	73.60	0.18	1.10	2.21	22.91	5.62

Table D1.5. Number of genes showing differential expression between species and misregulation in F1 hybrids.

Lines separate cross type (top: specialists, middle: generalist and scale-eater, bottom: generalist and molluscivore).

maternal population	paternal population	genes	DE between species	misregulated in F1	DE and misregulated	stage
CP molluscivore	CP scale-eater	11718	862	88	10	2dpf
OL scale-eater	OL molluscivore	11820	1900	150	32	2dpf
CP molluscivore	CP scale-eater	13013	4141	1208	320	8dpf
OL scale-eater	OL molluscivore	13225	2020	158	18	8dpf
CP generalist	CP scale-eater	11671	335	7	0	2dpf
OL generalist	OL scale-eater	11650	1455	1453	362	2dpf
CP generalist	CP scale-eater	13300	716	1009	87	8dpf
OL generalist	OL scale-eater	13254	3918	1088	244	8dpf
OL scale-eater	OL generalist	11650	1455	1283	38	2dpf
OL scale-eater	OL generalist	13254	3918	2016	72	8dpf
CP generalist	CP molluscivore	12202	606	536	37	2dpf
OL generalist	OL molluscivore	12207	97	2142	4	2dpf
CP generalist	CP molluscivore	13594	371	168	13	8dpf
OL generalist	OL molluscivore	13697	1945	1780	194	8dpf
CP molluscivore	CP generalist	11814	606	69	4	2dpf
OL molluscivore	OL generalist	12099	97	256	0	2dpf
CP molluscivore	CP generalist	13768	371	31	0	8dpf
OL molluscivore	OL generalist	13694	1945	443	25	8dpf

Table D1.6. Genes differentially expressed between species and misregulated in hybrids that were common to both 8dpf Crescent Pond (CP) and Osprey Lake (OL) comparisons.

cross	transcript	gene	log2 fold change CP mother vs CP father	log2 fold change OL mother vs OL father	log2 fold change CP parents vs. CP hybrids	log2 fold change OL parents vs. OL hybrids
generalist × scale-eater	XM_015396529.1	*trim47*	-1.332	0.547	-1.332	-1.278
generalist × scale-eater	XM_015405031.1	*krt13*	-1.184	-1.181	-1.183	-1.229
generalist × scale-eater	XM_015380548.1	*s100a1*	-1.176	0.466	-1.176	-0.905
scale-eater × molluscivore	XM_015396195.1	*elovl7*	0.784	-0.641	-0.978	-0.996

Table D1.7. 360 significantly enriched gene ontology terms for 125 genes showing differential expression between species and misregulation in F1 hybrids found within highly differentiated regions of the genome.

GO term	Enrichment FDR	Genes in list
Muscle structure development	0.000347	16
Muscle organ development	0.000673	12
Neuron projection development	0.000673	19
Cellular component biogenesis	0.000673	39
Neuron development	0.002059	19
Response to stress	0.002071	43
Response to abiotic stimulus	0.002071	19
Anatomical structure morphogenesis	0.002071	31
Animal organ development	0.002071	38
System development	0.002071	47
Cellular response to organic cyclic compound	0.002071	13
Tissue development	0.002589	26
Hindbrain structural organization	0.002632	2
Cerebellum structural organization	0.002632	2
Cellular response to stress	0.002632	26
Negative regulation of neuron differentiation	0.002632	8
Response to external stimulus	0.002697	29
Striated muscle tissue development	0.002697	10
Neuron differentiation	0.002697	20
Cellular response to nutrient levels	0.002697	8
Organic substance transport	0.002996	32
Generation of neurons	0.003242	21
Muscle tissue development	0.003242	10
Cell development	0.003307	26
Regulation of neuron projection development	0.00339	11
Cardiac muscle contraction	0.003875	6
Negative regulation of cell development	0.003926	9
Cellular response to external stimulus	0.003926	9
Cellular response to extracellular stimulus	0.004413	8
Cellular component assembly	0.005139	33
Nitrogen compound transport	0.005139	28
Neurogenesis	0.005335	21
Regulation of anatomical structure morphogenesis	0.005335	16
Cell differentiation	0.005335	40
Protein-containing complex subunit organization	0.005335	27
Anatomical structure arrangement	0.005335	3
Regulation of multicellular organismal development	0.005335	24
Negative regulation of neuron projection development	0.005397	6
Response to organic cyclic compound	0.005695	15
Negative regulation of neurogenesis	0.005782	8
Regulation of neuron differentiation	0.005898	12
Lateral motor column neuron migration	0.005898	2
Response to oxygen-containing compound	0.005898	21
Regulation of plasma membrane bounded cell projection organization	0.006627	12
Regulation of cell projection organization	0.007269	12
Striated muscle cell development	0.007269	6
Ribosome biogenesis	0.007398	8
Negative regulation of nervous system development	0.007398	8
Striated muscle contraction	0.00753	6

Fructose catabolic process	0.007713	2
Positive regulation of metabolic process	0.007713	35
Spinal cord development	0.007713	5
Cellular protein-containing complex assembly	0.007713	17
Fructose catabolic process to hydroxyacetone phosphate and glyceraldehyde-3-phosphate	0.007713	2
Spinal cord motor neuron migration	0.007713	2
Actin-mediated cell contraction	0.007955	5
Regulation of cellular response to heat	0.007955	4
Ribonucleoprotein complex biogenesis	0.008043	10
Regulation of nervous system development	0.008242	14
Muscle cell development	0.00842	6
Negative regulation of cell projection organization	0.008537	6
Cellular developmental process	0.008827	40
Regulation of neurogenesis	0.009003	13
Plasma membrane bounded cell projection organization	0.009559	19
Regulation of cell development	0.009846	14
Skeletal muscle organ development	0.009846	6
Cellular response to heat	0.010074	5
Chaperone-mediated protein folding	0.010116	4
RRNA metabolic process	0.010443	7
Negative regulation of intracellular signal transduction	0.010511	10
Regulation of developmental process	0.010661	27
Protein-containing complex assembly	0.010772	23
Cell projection organization	0.011215	19
Muscle cell differentiation	0.011641	8
Motor neuron migration	0.011641	2
Movement of cell or subcellular component	0.011646	23
Muscle fiber development	0.012324	4
Response to nitrogen compound	0.012511	15
Response to organic substance	0.012613	32
Nervous system development	0.013067	25
Neuron projection morphogenesis	0.013067	11
Cellular response to nitrogen compound	0.013067	11
Striated muscle cell differentiation	0.013121	7
Response to organonitrogen compound	0.013435	14
Actin filament-based movement	0.013435	5
Anterior/posterior axon guidance	0.013435	2
Cardiac muscle cell development	0.013962	4
Plasma membrane bounded cell projection morphogenesis	0.014449	11
Cell projection morphogenesis	0.014635	11
Response to mechanical stimulus	0.014808	6
Regulation of biological quality	0.014808	36
Monosaccharide metabolic process	0.015408	7
Regulation of cell-substrate adhesion	0.015572	6
G1 to G0 transition	0.01575	2
Cardiac cell development	0.016095	4
Cellular response to organonitrogen compound	0.016709	10
Cell part morphogenesis	0.016796	11
Positive regulation of developmental process	0.017118	17
Muscle filament sliding	0.01717	3
Actin-myosin filament sliding	0.01717	3
Regulation of microtubule polymerization or depolymerization	0.017257	4
Desmosome organization	0.01743	2
RRNA processing	0.01743	6

Response to wounding	0.01743	11
Regulation of neuron maturation	0.01743	2
Aggrephagy	0.01743	2
Cellular response to chemical stimulus	0.018149	31
Regulation of keratinocyte differentiation	0.018299	3
Circulatory system development	0.018299	14
Cellular response to starvation	0.018748	5
Endonucleolytic cleavage involved in rRNA processing	0.019353	2
Endonucleolytic cleavage of tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	0.019353	2
Protein folding	0.019353	6
Post-embryonic development	0.019353	4
Cerebellum morphogenesis	0.019353	3
Monocarboxylic acid metabolic process	0.019353	10
Regulation of cell differentiation	0.019353	20
Axon development	0.019353	9
Regulation of response to stress	0.019353	18
Regulation of protein modification by small protein conjugation or removal	0.019353	6
Intracellular receptor signaling pathway	0.01975	7
Cellular response to epidermal growth factor stimulus	0.020014	3
Heart contraction	0.020237	6
Dendrite development	0.020502	6
Microtubule depolymerization	0.02085	3
Cellular response to nitrogen starvation	0.021155	2
Cellular response to nitrogen levels	0.021155	2
Negative regulation of cell morphogenesis involved in differentiation	0.02142	4
Organic acid biosynthetic process	0.02142	8
Carboxylic acid biosynthetic process	0.02142	8
Regulation of response to stimulus	0.02142	38
Regulation of developmental growth	0.02142	7
Regulation of multicellular organismal process	0.02142	29
Cellular response to abiotic stimulus	0.02142	7
Cellular response to environmental stimulus	0.02142	7
Response to cAMP	0.021439	4
Heart process	0.021812	6
Purine nucleoside diphosphate metabolic process	0.021812	4
Purine ribonucleoside diphosphate metabolic process	0.021812	4
Response to heat	0.021812	5
Hexose metabolic process	0.021812	6
Hindbrain morphogenesis	0.021812	3
Positive regulation of organ growth	0.021812	3
Response to epidermal growth factor	0.021812	3
Ribonucleoside diphosphate metabolic process	0.02307	4
Regulation of response to external stimulus	0.02307	12
Negative regulation of cell differentiation	0.02314	11
RNA processing	0.023278	13
Response to peptide hormone	0.023278	8
Skeletal muscle tissue development	0.023395	5
Embryo implantation	0.023395	3
Positive regulation of developmental growth	0.024272	5
Muscle contraction	0.024451	7
Heart development	0.024451	9
Response to acid chemical	0.026233	7
Positive regulation of cellular metabolic process	0.026233	30

Fructose metabolic process	0.026609	2
Animal organ morphogenesis	0.026609	13
Skeletal muscle thin filament assembly	0.026609	2
Positive regulation of protein ubiquitination	0.026643	4
Cell-cell adhesion	0.027027	12
Response to inorganic substance	0.02784	9
Macromolecule localization	0.02784	29
Regulation of axonogenesis	0.02784	5
Cellular macromolecule localization	0.02784	20
Myotube differentiation	0.027946	4
Hexose catabolic process	0.027946	3
Cellular component morphogenesis	0.027946	14
Cellular localization	0.027946	27
Mesenchyme development	0.027946	6
Cellular response to endogenous stimulus	0.027946	16
Cellular response to organic substance	0.028515	26
Axonogenesis	0.029032	8
Tube development	0.029032	13
Response to drug	0.029032	13
Positive regulation of neuron differentiation	0.029032	7
Cellular response to oxygen-containing compound	0.029032	14
Carboxylic acid metabolic process	0.029103	13
Regulation of cellular component organization	0.029382	24
Cardiac muscle cell differentiation	0.029515	4
Response to starvation	0.029555	5
Cellular response to steroid hormone stimulus	0.029555	6
Positive regulation of neuron projection development	0.02975	6
Head development	0.02975	11
Response to insulin	0.030109	6
NAD biosynthetic process	0.030452	3
Coenzyme metabolic process	0.031917	7
Nucleoside diphosphate metabolic process	0.031917	4
Skeletal myofibril assembly	0.032288	2
Supramolecular fiber organization	0.032357	10
Anion transmembrane transport	0.032357	6
Polyol metabolic process	0.033638	4
Microtubule polymerization or depolymerization	0.033638	4
Regulation of epidermal cell differentiation	0.033638	3
Positive regulation of cell projection organization	0.033969	7
Female pregnancy	0.034504	5
Response to muscle stretch	0.034504	2
Neural retina development	0.03529	3
Carbohydrate metabolic process	0.03529	9
Glucose metabolic process	0.03529	5
Protein localization to nucleus	0.03529	6
Nucleic acid transport	0.03529	5
RNA transport	0.03529	5
Membrane organization	0.03529	11
Negative regulation of metabolic process	0.035338	27
Negative regulation of cell-substrate adhesion	0.035338	3
Regulation of protein ubiquitination	0.035338	5
Response to nutrient levels	0.035338	8
Monosaccharide catabolic process	0.035338	3
Intracellular transport	0.035338	19
Cardiac muscle fiber development	0.035338	2
Maternal process involved in female pregnancy	0.035338	3

Positive regulation of protein modification by small protein conjugation or removal	0.035338	4
Establishment of RNA localization	0.035735	5
Negative regulation of cell adhesion	0.036136	6
Regulation of cell morphogenesis	0.036136	8
Lipoprotein metabolic process	0.036136	4
Organic acid transmembrane transport	0.036136	4
Carboxylic acid transmembrane transport	0.036136	4
Regulation of nitric oxide biosynthetic process	0.03672	3
Cardiac muscle tissue development	0.03672	5
Cleavage involved in rRNA processing	0.036849	2
Glyceraldehyde-3-phosphate metabolic process	0.036849	2
Muscle cell cellular homeostasis	0.036849	2
Negative regulation of cellular component organization	0.036849	10
Regulation of cell morphogenesis involved in differentiation	0.037187	6
Cellular response to nutrient	0.037187	3
Maturation of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	0.037741	2
Glycerol metabolic process	0.037741	2
Cytoskeleton organization	0.037741	15
Cell adhesion	0.037741	16
Detection of external stimulus	0.037741	4
Negative regulation of signal transduction	0.037741	14
Biological adhesion	0.037741	16
Establishment of mitochondrion localization, microtubule-mediated	0.037741	2
Amide transport	0.037741	21
Regulation of mRNA stability	0.037741	4
Mitochondrion transport along microtubule	0.037741	2
Negative regulation of axonogenesis	0.037741	3
Negative regulation of ERK1 and ERK2 cascade	0.037741	3
Cellular response to amino acid stimulus	0.037741	3
Cardiac muscle cell action potential	0.037741	3
Response to peptide	0.037741	8
Detection of abiotic stimulus	0.038322	4
Negative regulation of cellular metabolic process	0.038322	24
Cellular protein localization	0.038322	19
Positive regulation of cell differentiation	0.038322	12
Response to organophosphorus	0.038322	4
Regulation of cell adhesion	0.038658	10
Retina layer formation	0.03906	2
Response to steroid hormone	0.03906	7
Developmental cell growth	0.03906	5
Positive regulation of mesonephros development	0.03906	2
Regulation of cellular response to stress	0.03906	10
Oxoacid metabolic process	0.040117	13
Response to endogenous stimulus	0.040319	17
Response to extracellular stimulus	0.040785	8
Small molecule biosynthetic process	0.040785	10
Brain development	0.041395	10
Regulation of cellular component movement	0.041395	12
Regulation of cell maturation	0.041395	2
Developmental growth	0.041884	9
Establishment of protein localization	0.041903	21
Regulation of neurotransmitter levels	0.042553	6
Muscle system process	0.042553	7

Organic acid metabolic process	0.042553	13
Cellular protein modification process	0.042553	34
Glutamine metabolic process	0.042553	2
NADH regeneration	0.042553	2
Nitric oxide biosynthetic process	0.042553	3
Carbohydrate transport	0.042553	4
Response to temperature stimulus	0.042553	5
Response to hormone	0.042553	12
Regulation of signal transduction	0.042553	28
Endomembrane system organization	0.042553	7
Regulation of cell communication	0.042553	30
Response to purine-containing compound	0.042553	4
Protein transport	0.042553	20
Protein import	0.042553	5
Alditol metabolic process	0.042553	2
NAD metabolic process	0.042553	3
Regulation of rhodopsin mediated signaling pathway	0.042553	2
Regulation of epithelial cell differentiation	0.042553	4
Membrane raft organization	0.042553	2
Regulation of response to extracellular stimulus	0.042553	2
Regulation of response to nutrient levels	0.042553	2
Maintenance of protein location in cell	0.042553	3
Cardiocyte differentiation	0.042553	4
Protein modification process	0.042553	34
Regulation of locomotion	0.042553	12
Ribosomal large subunit biogenesis	0.042553	3
Regulation of RNA stability	0.042553	4
Multi-multicellular organism process	0.042553	5
Decidualization	0.042553	2
Reproductive structure development	0.042553	7
Positive regulation of multicellular organismal process	0.042553	18
Nucleus localization	0.042553	2
Establishment of localization in cell	0.042553	21
Establishment of mitochondrion localization	0.042553	2
Positive regulation of nervous system development	0.042553	8
Regulation of ryanodine-sensitive calcium-release channel activity	0.042553	2
Canonical glycolysis	0.042553	2
Glucose catabolic process to pyruvate	0.042553	2
Regulation of anion transmembrane transport	0.042553	2
Heterotypic cell-cell adhesion	0.043292	3
Cellular response to lipid	0.043292	9
Reproductive system development	0.043576	7
Cardiac myofibril assembly	0.043663	2
Regulation of mesonephros development	0.043663	2
Glycolytic process through fructose-6-phosphate	0.043663	2
Glycolytic process through glucose-6-phosphate	0.043663	2
Cellular response to hypoxia	0.044341	4
Protein localization	0.044752	25
Transport along microtubule	0.044752	4
Nitric oxide metabolic process	0.044752	3
Maintenance of location	0.044752	6
Microtubule-based transport	0.044752	4
Regulation of signaling	0.044901	30
Keratinocyte differentiation	0.045069	6
Maturation of 5.8S rRNA	0.045277	2

Cell morphogenesis	0.045277	12
Neuron migration	0.045277	4
RNA localization	0.045277	5
Intracellular protein transport	0.045277	13
Cell death	0.045277	21
Posttranscriptional regulation of gene expression	0.045277	8
Peptide transport	0.045277	20
Regulation of fatty acid metabolic process	0.045277	3
N-terminal protein amino acid modification	0.045277	2
Regulation of protein modification process	0.045277	18
Homotypic cell-cell adhesion	0.045277	3
Cholesterol homeostasis	0.045277	3
Macromolecule modification	0.045277	35
Positive regulation of molecular function	0.045277	18
Regulation of fatty acid oxidation	0.045277	2
Positive regulation of lipid biosynthetic process	0.045277	3
MRNA transport	0.045277	4
Sterol homeostasis	0.045277	3
Oxidation-reduction process	0.045277	12
Regulation of mRNA catabolic process	0.045277	4
Response to oxygen levels	0.045277	6
Cellular response to vitamin	0.045277	2
Positive regulation of animal organ morphogenesis	0.045277	3
Regulation of cell motility	0.045277	11
Reactive nitrogen species metabolic process	0.045277	3
Positive regulation of macromolecule metabolic process	0.046773	28
Skin development	0.047322	7
Regulation of keratinocyte proliferation	0.047462	2
Cerebellar Purkinje cell layer development	0.047462	2
Regulation of microtubule depolymerization	0.047462	2
Regulation of epidermis development	0.047462	3
Cell-substrate adhesion	0.047838	6
Cellular response to decreased oxygen levels	0.048446	4
Muscle organ morphogenesis	0.048446	3
Nucleobase-containing compound transport	0.049275	5
Gluconeogenesis	0.049438	3
Adult walking behavior	0.049438	2
Rhodopsin mediated signaling pathway	0.049438	2
Regulation of axon extension involved in axon guidance	0.049438	2
Wound healing	0.049598	8

Table D1.8. 26 genes showing differential expression between species and misregulation in F1 hybrids found within highly differentiated regions of the that also show strong signs of a hard selective sweep in specialists.

26 genes showing differential expression between species and misregulation in F1 hybrids found within highly differentiated regions of the genome (F_{st} = 1; D_{xy} ≥ genome-wide 90th percentile (values in bold; range = 0.0031 – 0.0075; see table D1.3 for all population thresholds)) that also show strong signs of a hard selective sweep in specialists (negative Tajima's D < genome-wide 10th percentile (values in bold; range = -1.62 – -0.77 (see table D1.2 for all population thresholds); SweeD composite likelihood ratio > 90th percentile for scaffold (values in bold)).

maternal population	paternal population	stage	gene	fixed SNPs within 20kb	Tajima's D maternal population	Tajima's D paternal population	CLR maternal population	CLR paternal population
OL generalist	OL scale-eater	2dpf	*pak3*	11	-0.45	**-1.33**	530.3	**1241.6**
OL generalist	OL scale-eater	2dpf	*mttp*	111	-0.28	**-1.31**	315.9	**1011.5**
OL generalist	OL scale-eater	2dpf	*phgdh*	8	0.33	**-1.48**	383.9	**1076.1**
OL generalist	OL scale-eater	2dpf	*svil*	6	**-0.97**	**-1.53**	3136.0	**4458.7**
OL generalist	OL scale-eater	2dpf	*dscam*	8	**-1.03**	**-1.34**	923.7	**2663.9**
OL generalist	OL scale-eater	2dpf	*dab1*	24	-0.04	**-1.51**	1285.5	**2755.9**
CP generailst	CP scale-eater	8dpf	*dbi*	3	0.39	**-1.66**	337.5	**1121.7**
OL scale-eater	OL molluscivore	2dpf	*lctl*	42	**-1.75**	0.99	**962.1**	202.8
CP molluscivore	CP scale-eater	8dpf	*pdcd11*	52	**-1.62**	-1.41	**2351.7**	**2208.3**
CP molluscivore	CP scale-eater	8dpf	*nup205*	50	**-1.56**	-0.87	**1747.5**	206.1
CP molluscivore	CP scale-eater	8dpf	*107098071*	3	**-1.95**	-0.68	**1289.4**	754.8
CP molluscivore	CP scale-eater	8dpf	*ttn*	52	**-1.68**	**-1.66**	**5370.8**	2041.6
CP molluscivore	CP scale-eater	8dpf	*nup155*	4	0.99	**-1.74**	201.4	**1929.8**
CP molluscivore	CP scale-eater	8dpf	*cabp7*	8	-0.14	**-1.61**	**1480.7**	161.9
CP molluscivore	CP scale-eater	8dpf	*ppp5c*	301	**-1.64**	**-1.66**	**163.2**	**130.4**
CP molluscivore	CP scale-eater	8dpf	*unc45a*	66	**-1.68**	**-1.66**	**5369.8**	2042.5
CP molluscivore	CP scale-eater	8dpf	*polr2b*	183	-1.27	**-1.71**	807.3	**2203.0**
CP molluscivore	CP scale-eater	8dpf	*dusp3*	21	**-1.54**	0.14	17.0	60.9
CP molluscivore	CP scale-eater	8dpf	*ndufa4l2*	19	-1.39	**-1.77**	**3031.1**	**2809.3**
CP molluscivore	CP scale-eater	8dpf	*psmd11*	13	**-1.58**	0.94	**135.8**	**125.8**
CP molluscivore	CP scale-eater	8dpf	*pde6g*	30	0.24	**-1.77**	**1530.2**	**1261.4**
CP molluscivore	CP scale-eater	8dpf	*map1s*	7	0.16	**-1.75**	457.8	**1523.2**
CP molluscivore	CP scale-eater	8dpf	*ptprn2*	29	**-1.61**	**-1.82**	**2211.6**	1392.6
CP molluscivore	CP scale-eater	8dpf	*slc43a1*	362	**-1.64**	-1.49	**809.6**	**662.4**
OL scale-eater	OL molluscivore	8dpf	*slc38a8*	62	**-1.48**	-0.13	**3749.1**	2435.3
OL scale-eater	OL molluscivore	8dpf	*sema6c*	64	-0.82	**-1.82**	2253.9	**3918.3**

Table D1.9. Ecological DMI candidate genes associated with jaw size.

Nine genes showing differential expression between species and misregulation in F1 hybrids found within highly differentiated regions of the genome (F_{st} = 1; $D_{xy} \geq$ genome-wide 90th percentile (values in bold; range = 0.0075 – 0.0031; see table D1.3 for all population thresholds)) were also in a 20 kb regions significantly associated with oral jaw size variation across our Caribbean pupfish samples (GEMMA PIP > 99th percentile (0.00175)). Genes in bold are discussed in the main text. The genes *sema6c* and *dbi* also show signs of a hard selective sweep in specialists (negative Tajima's D < genome-wide 10th percentile; range = -1.62 – -0.77 (see table D1.2 for all population thresholds); SweeD composite likelihood ratio > 90th percentile by scaffold (values in bold)).

maternal population	paternal population	stage	gene	fixed SNPs within 20kb	Tajima's D maternal population	Tajima's D paternal population	CLR maternal population	CLR paternal population	PIP
CP generailst	CP scale-eater	8	**mpp1**	170	0.824871	-0.57836	1181.48	1364.328	0.00255
CP generailst	CP scale-eater	8	dbi	3	0.390309	**-1.65859**	337.5028	**1121.688**	0.00198
CP molluscivore	CP scale-eater	8	rcl1	9	-0.59334	-1.19039	1028.911	433.5589	0.00379
CP molluscivore	CP scale-eater	8	prpf39	325	-1.03984	-1.14611	137.0623	2474.137	0.0025
CP molluscivore	CP scale-eater	8	107082296	2	-1.07899	-0.35454	289.3542	1000.216	0.00175
OL generalist	OL scale-eater	8	rcl1	3	-0.46693	-1.19082	654.403	1471.226	0.00379
OL scale-eater	OL molluscivore	8	**sema6c**	64	-0.81823	**-1.81724**	2253.855	**3918.334**	0.00213
OL scale-eater	OL molluscivore	8	mid1ip1	1	0.594817	-0.27379	32.32544	1237.023	0.00185
CP molluscivore	CP scale-eater	48	hbae	29	-1.3977	1.87904	1218.031	41.76962	0.00191
OL generalist	OL scale-eater	48	ak3	4	-0.79556	-1.19082	797.5076	1471.226	0.00379

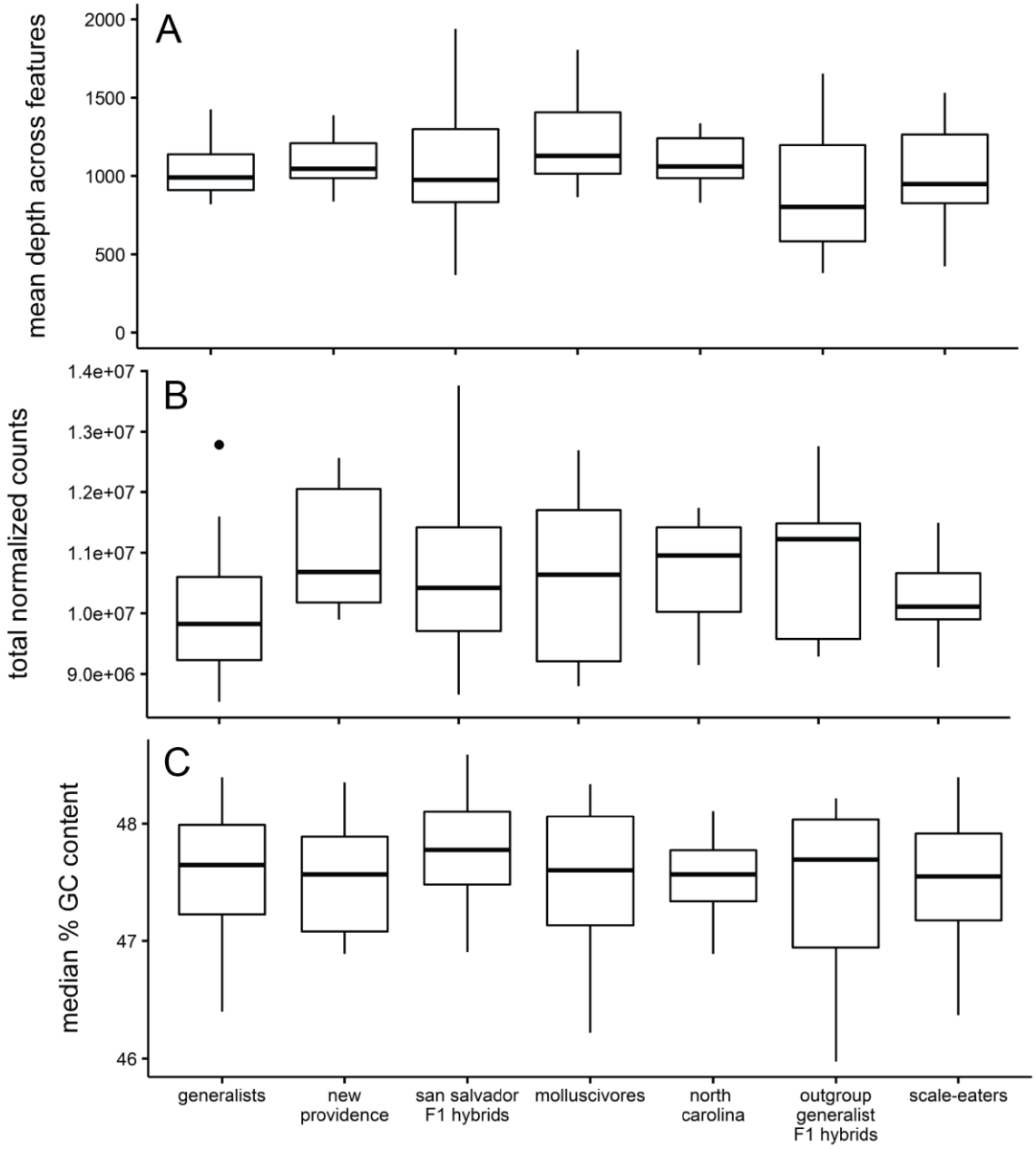

Figure D2.1. No significant difference among F1 purebred and F1 hybrid samples for quality control measures.

No significant difference among F1 purebred and F1 hybrid samples for A) mean read depth across annotated features (ANOVA; $P = 0.32$), B) total normalized read counts (ANOVA; $P = 0.16$), C) median percent GC content of reads (ANOVA; $P = 0.32$).

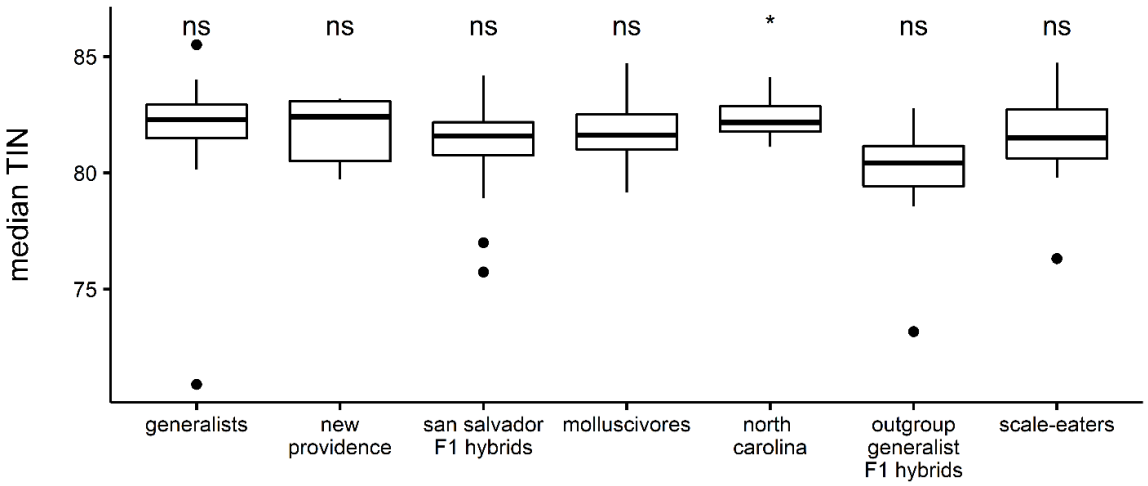

Figure D2.2. Median transcript integrity numbers for each species and generalist population.

Tukey post-hoc test: $P < 0.05$ = *; $P > 0.05$ = ns.

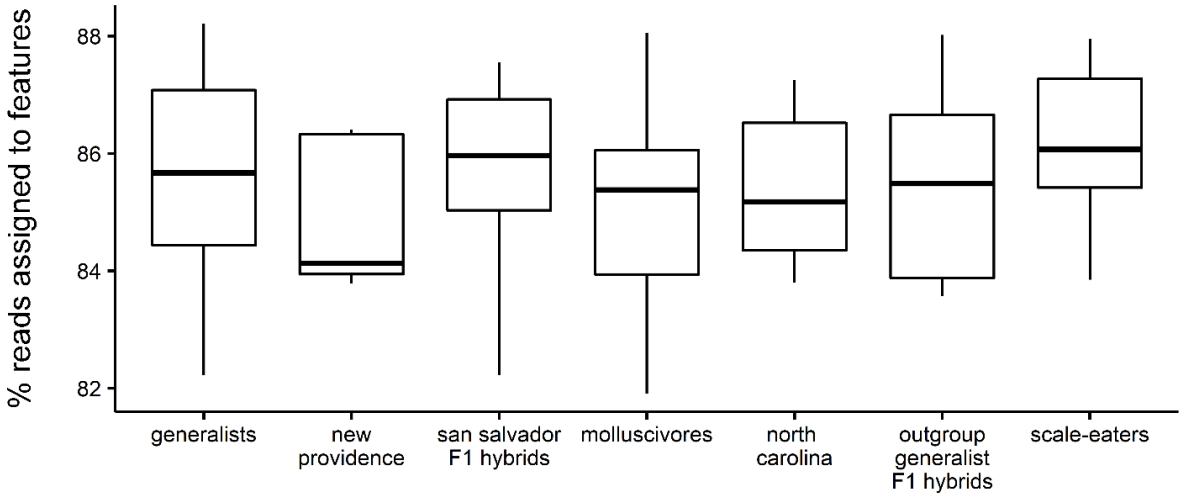

Figure D2.3. No significant difference in the percentage of reads mapping to annotated features of the Cyprinodon reference genome among F1 purebred and F1 hybrid samples.

ANOVA; $P = 0.17$

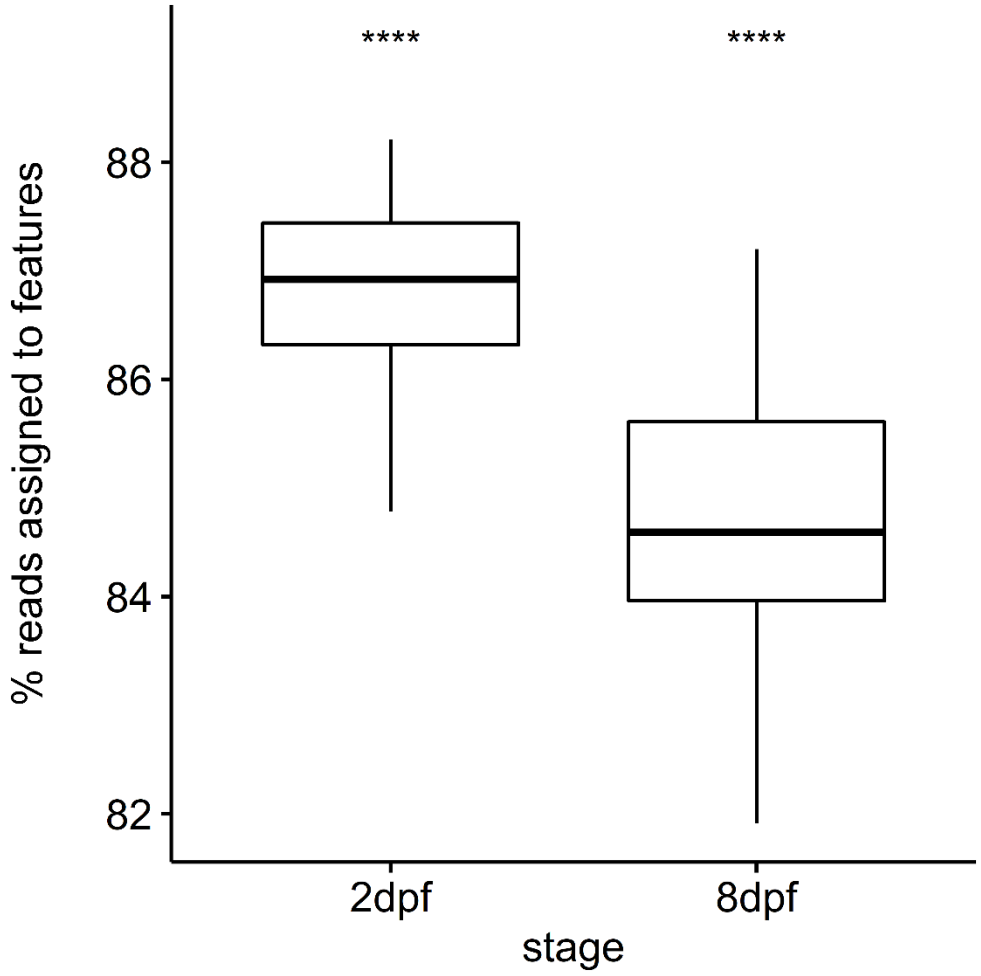

Figure D2.4. More reads assigned to features for 2 dpf samples than 8 dpf samples.

Student's *t*-test; $P < 2.2 \times 10^{-16}$

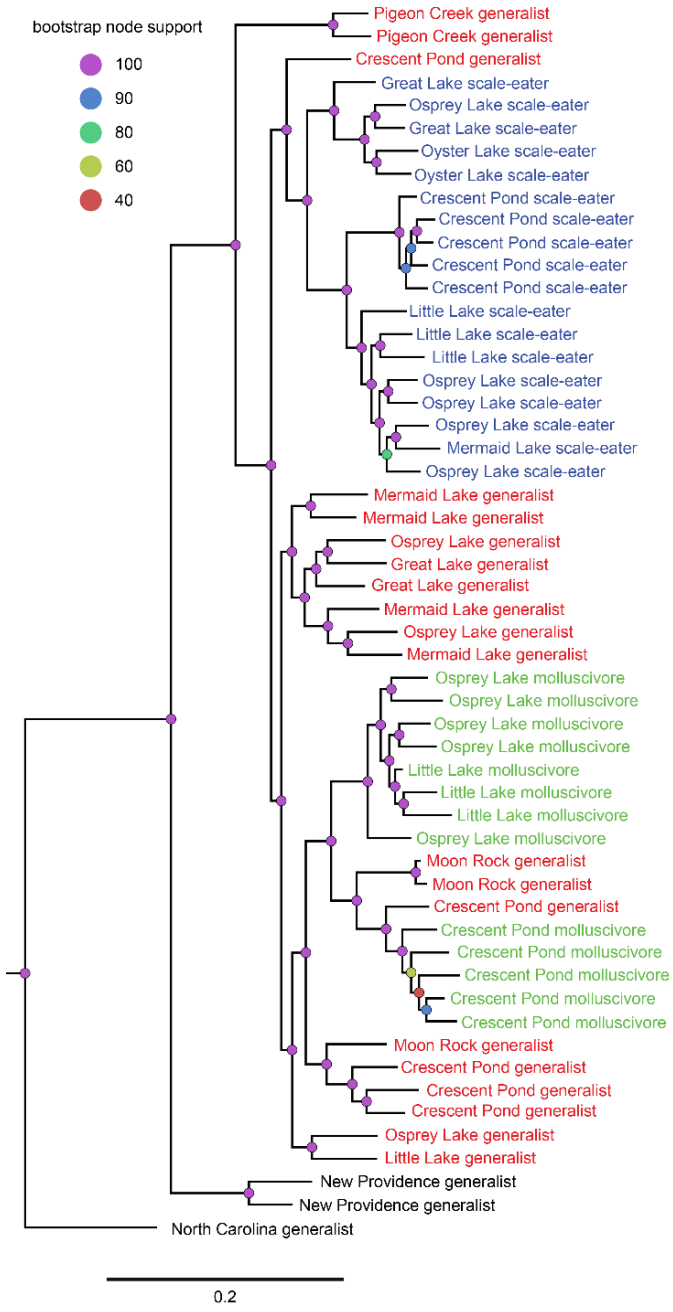

Figure D2.5. Maximum likelihood tree generated using RAxML with 1.7 million SNPs showing phylogenetic relationships between 55 Cyprinodon individuals.

Relationships for three outgroup individuals that were included in the genomic dataset are not shown. Red = San Salvador generalist, green = molluscivore, blue = scale-eater, black = outgroup generalist.

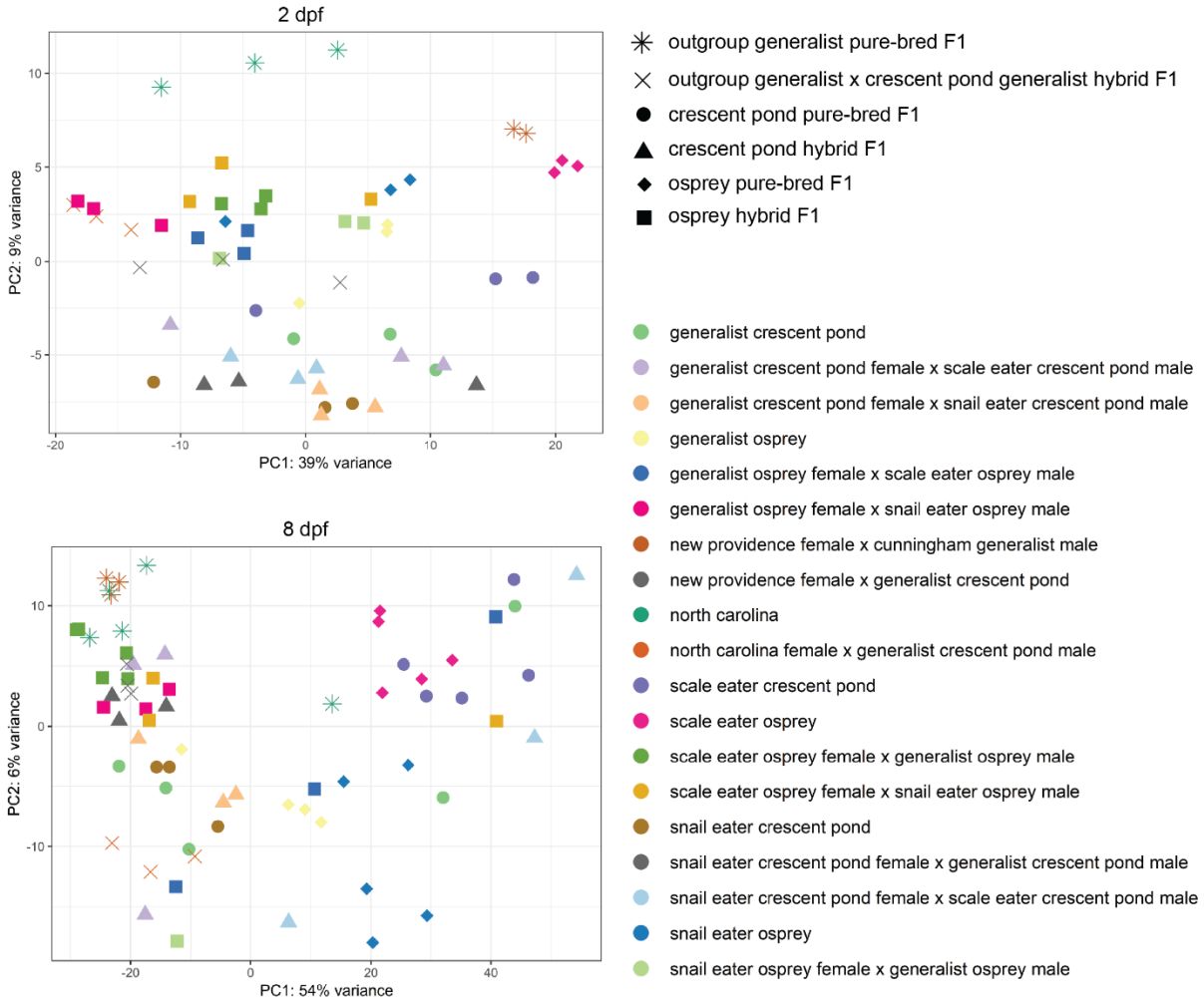

Figure D2.6. First two principal components explaining 48% (2 dpf) and 60% (8 dpf) of the variance across normalized read counts.

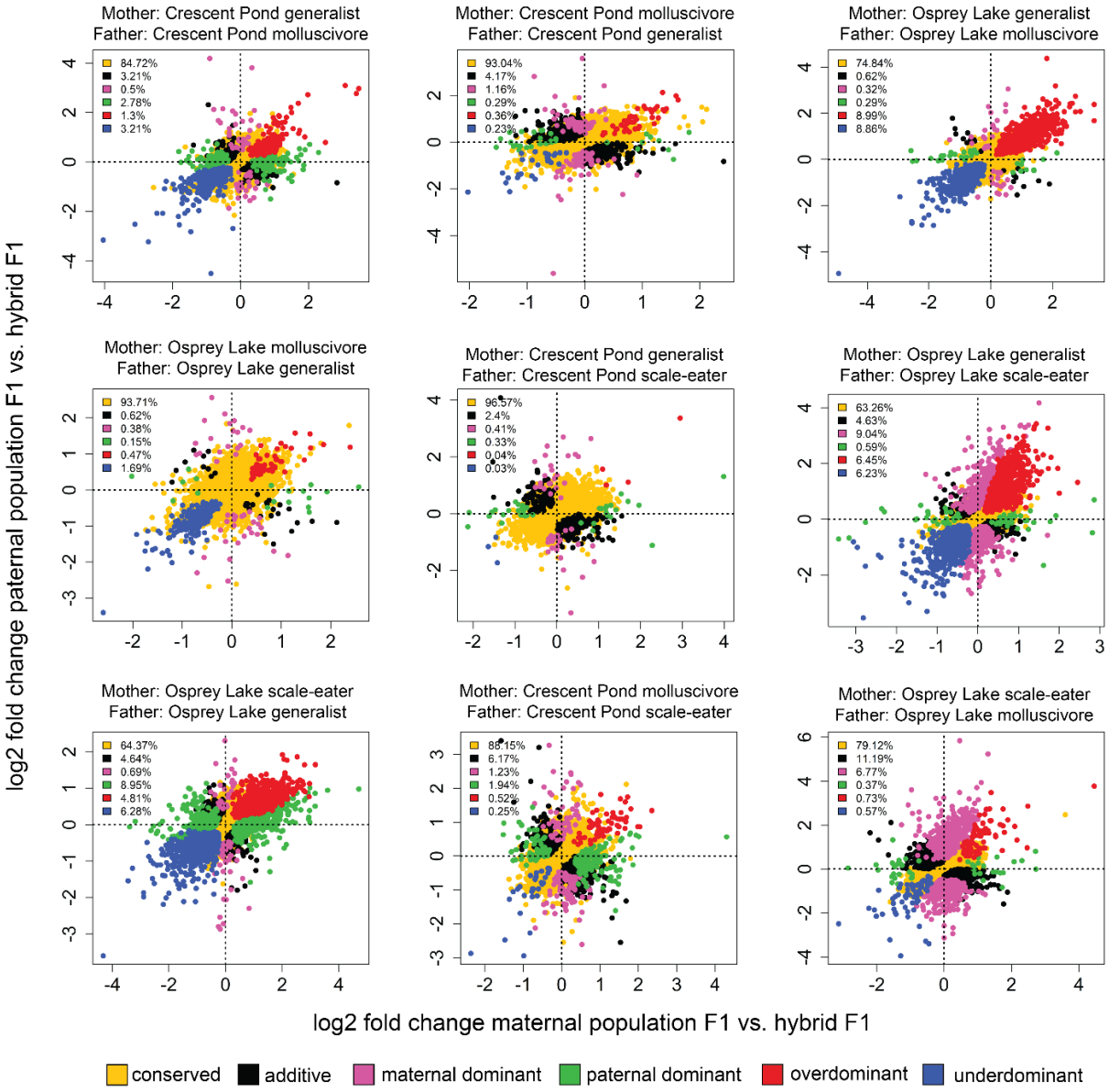

Figure D2.7. Gene expression inheritance for 2 dpf San Salvador hybrid crosses.

Yellow = conserved (no difference in expression between groups or ambiguous expression patterns), black = additive (differential expression between purebred F1 and intermediate expression levels in hybrid F1), pink = maternal dominant (differential expression between purebred F1, differential expression between paternal population purebred F1 and F1 hybrids, no differential expression between maternal population purebred F1 and F1 hybrids), green = paternal dominant (differential expression between purebred F1, differential expression between maternal population purebred F1 and F1 hybrids, no differential expression between paternal population purebred F1 and F1 hybrids), red = overdominant (F1 hybrid gene expression significantly higher than parental population purebred F1), blue = underdominant (F1 hybrid gene expression significantly lower than parental population purebred F1).

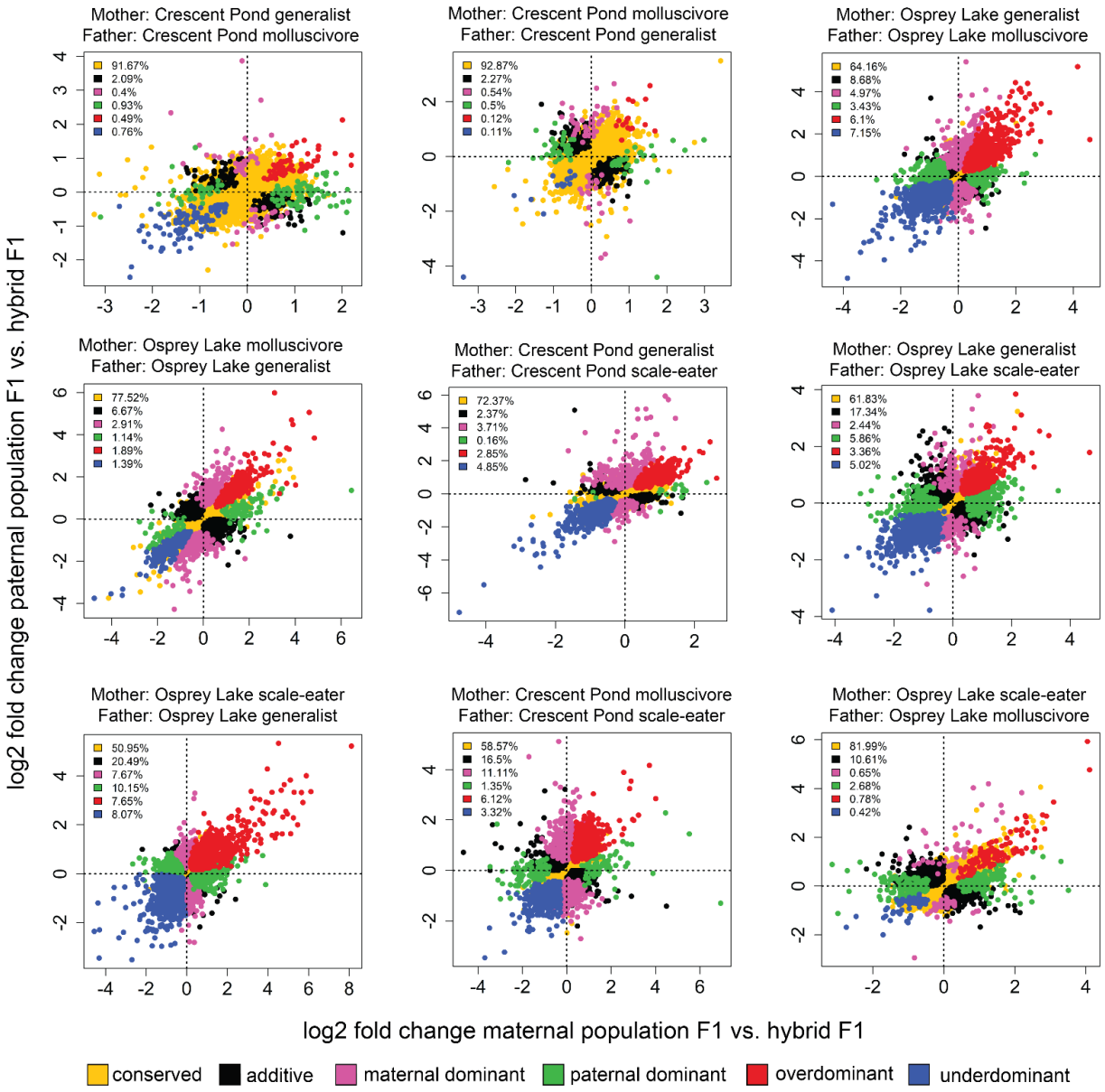

Figure D2.8. Gene expression inheritance for 8 dpf San Salvador hybrid crosses.

Yellow = conserved (no difference in expression between groups or ambiguous expression patterns), black = additive (differential expression between purebred F1 and intermediate expression levels in hybrid F1), pink = maternal dominant (differential expression between purebred F1, differential expression between paternal population purebred F1 and F1 hybrids, no differential expression between maternal population purebred F1 and F1 hybrids), green = paternal dominant (differential expression between purebred F1, differential expression between maternal population purebred F1 and F1 hybrids, no differential expression between paternal population purebred F1 and F1 hybrids), red = overdominant (F1 hybrid gene expression significantly higher than parental population purebred F1), blue = underdominant (F1 hybrid gene expression significantly lower than parental population purebred F1).

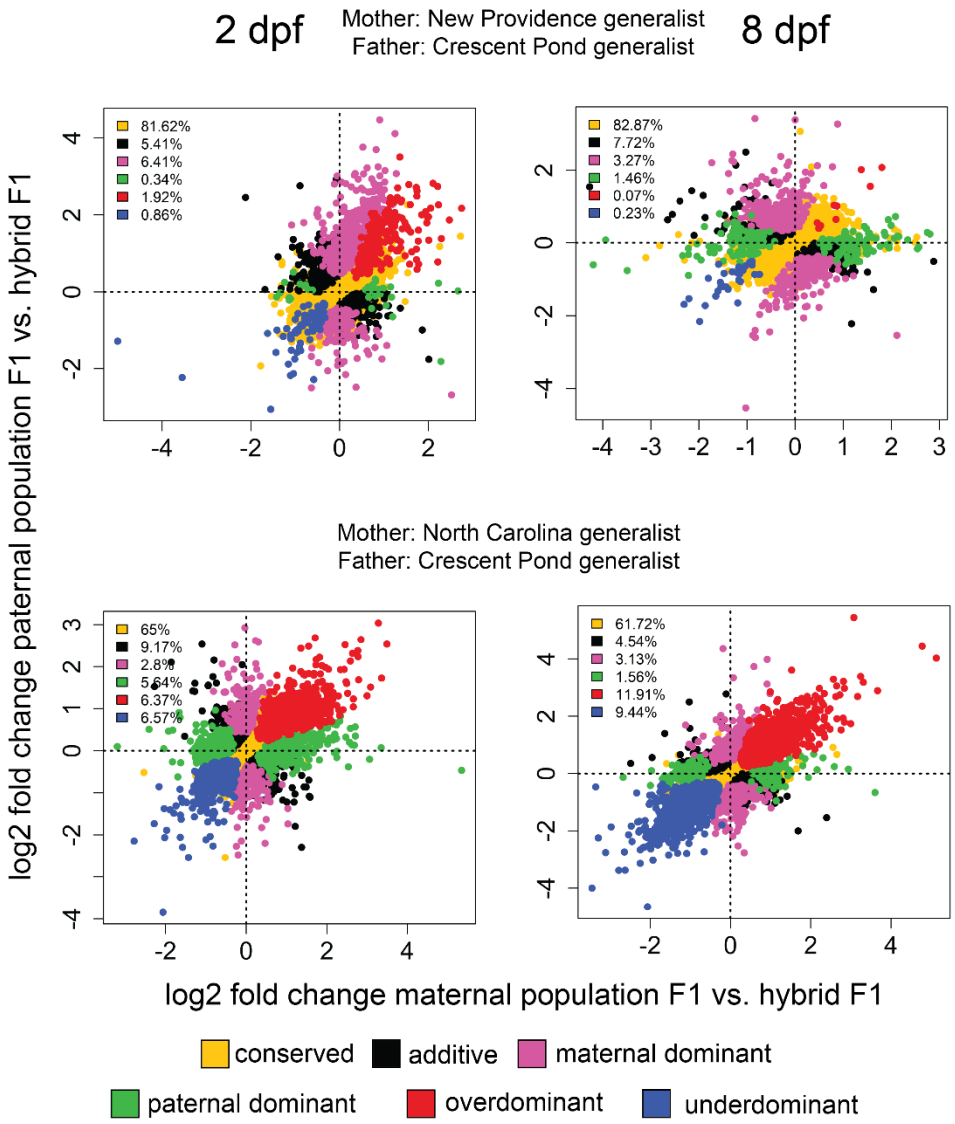

Figure D2.9. Gene expression inheritance for outgroup generalist population hybrid crosses.

Yellow = conserved (no difference in expression between groups or ambiguous expression patterns), black = additive (differential expression between purebred F1 and intermediate expression levels in hybrid F1), pink = maternal dominant (differential expression between purebred F1, differential expression between paternal population purebred F1 and F1 hybrids, no differential expression between maternal population purebred F1 and F1 hybrids), green = paternal dominant (differential expression between purebred F1, differential expression between maternal population purebred F1 and F1 hybrids, no differential expression between paternal population purebred F1 and F1 hybrids), red = overdominant (F1 hybrid gene expression significantly higher than parental population purebred F1), blue = underdominant (F1 hybrid gene expression significantly lower than parental population purebred F1).

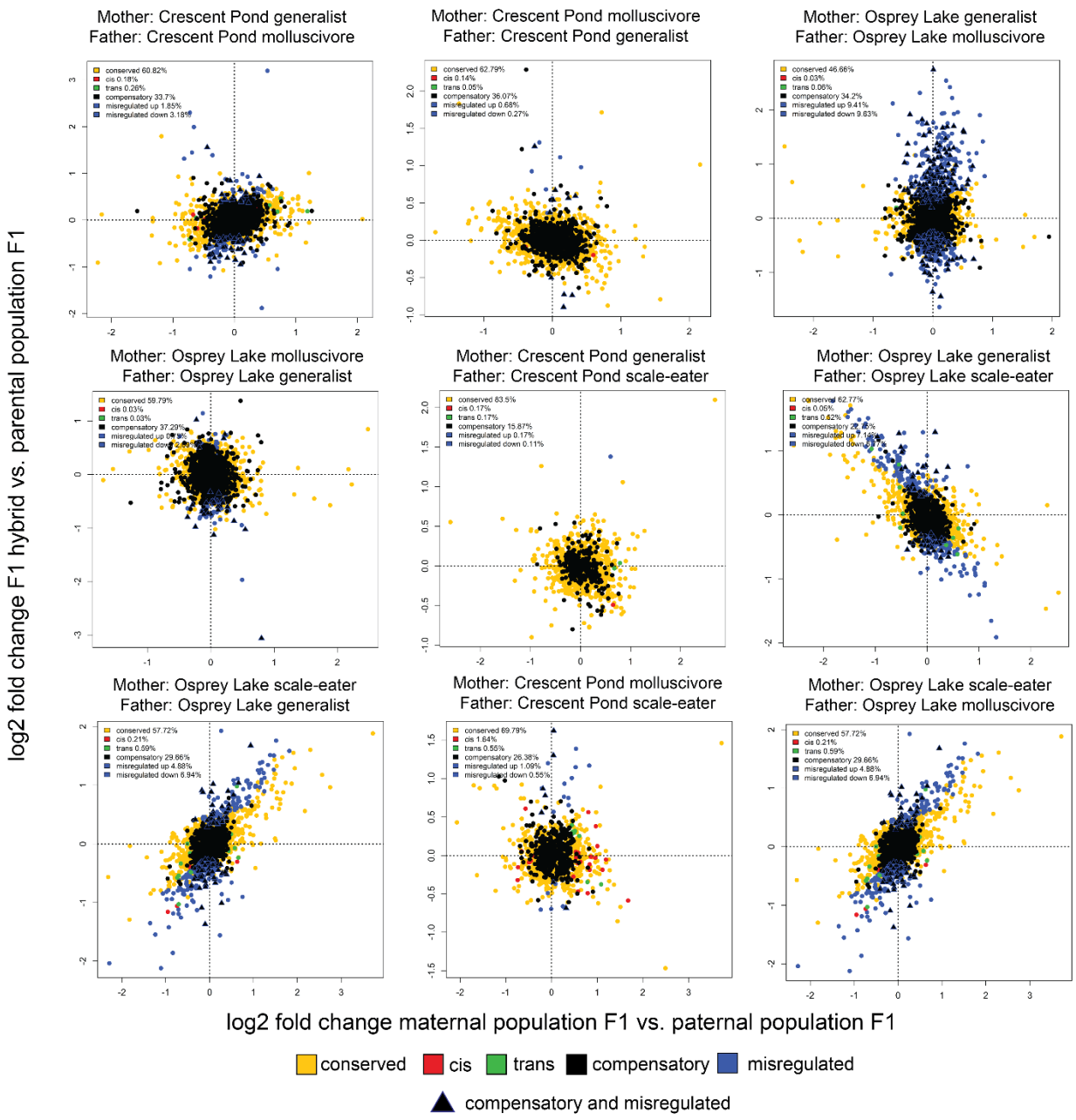

Figure D2.10. Regulatory mechanisms underlying expression divergence at 2 dpf in San Salvador crosses.

Yellow = conserved (no difference in expression between any group or ambiguous expression patterns), red = *cis* (significant ASE in hybrids, significant differential expression between parental populations of purebred F1 offspring, and no significant *trans*- contribution), green = *trans* (significant ASE in hybrids, significant differential expression between parental populations of purebred F1 offspring, and significant *trans*- contribution), black = compensatory (significant ASE in hybrids, no significant differential expression between parental populations of purebred F1 offspring), blue = misregulated (significant differential expression between purebred F1 and hybrid F1), triangle = compensatory and misregulated.

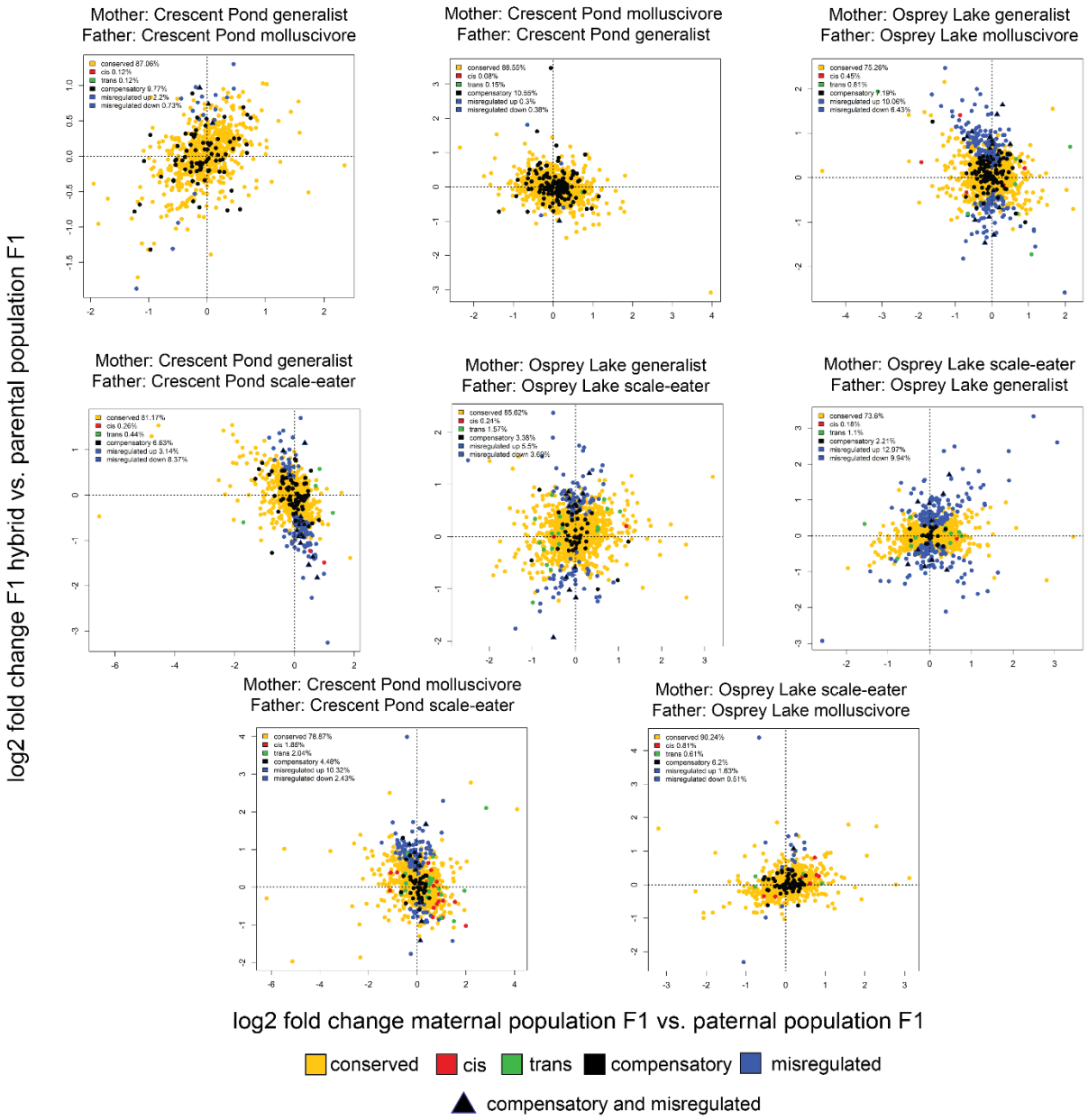

Figure D2.11. Regulatory mechanisms underlying expression divergence at 8 dpf in San Salvador crosses.

Yellow = conserved (no difference in expression between any group or ambiguous expression patterns), red = *cis* (significant ASE in hybrids, significant differential expression between parental populations of purebred F1 offspring, and no significant *trans*- contribution), green = *trans* (significant ASE in hybrids, significant differential expression between parental populations of purebred F1 offspring, and significant *trans*- contribution), black = compensatory (significant ASE in hybrids, no significant differential expression between parental populations of purebred F1 offspring), blue = misregulated (significant differential expression between purebred F1 and hybrid F1), triangle = compensatory and misregulated.

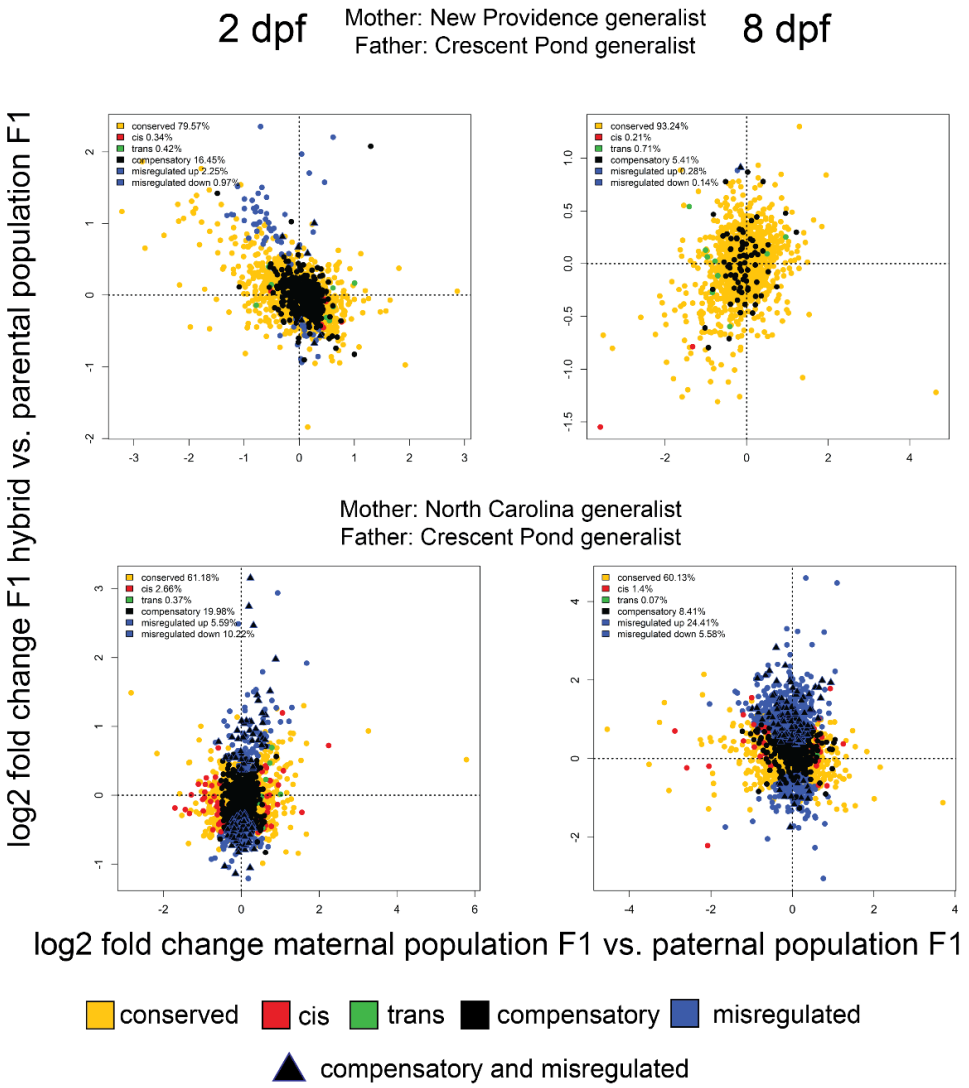

Figure D2.12. Regulatory mechanisms underlying expression divergence in outgroup generalist population crosses.

Yellow = conserved (no difference in expression between any group or ambiguous expression patterns), red = *cis* (significant ASE in hybrids, significant differential expression between parental populations of purebred F1 offspring, and no significant *trans*- contribution), green = *trans* (significant ASE in hybrids, significant differential expression between parental populations of purebred F1 offspring, and significant *trans*- contribution), black = compensatory (significant ASE in hybrids, no significant differential expression between parental populations of purebred F1 offspring), blue = misregulated (significant differential expression between purebred F1 and hybrid F1), triangle = compensatory and misregulated.

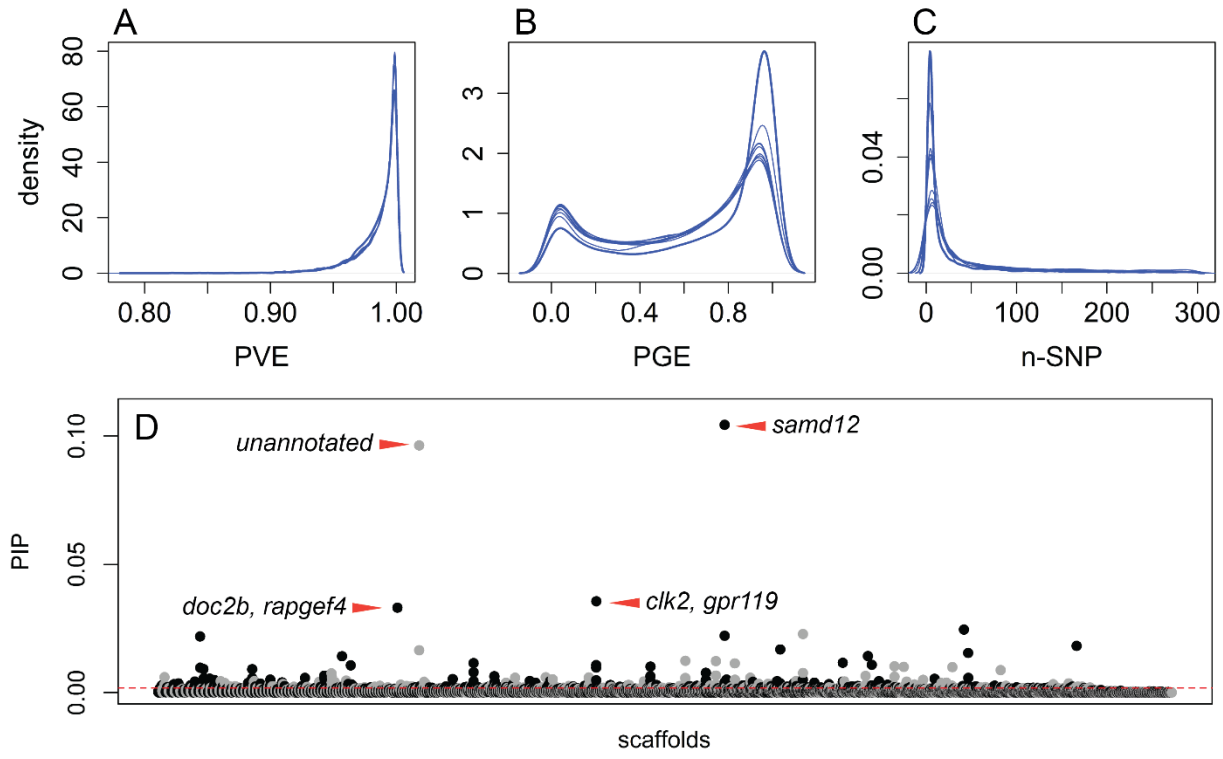

Figure D2.13. Genome-wide association mapping.

GEMMA implements a Bayesian sparse linear mixed model (BSLMM) that uses MCMC to estimate the proportion of phenotypic variation explained by every SNP included in the analysis (A; PVE), the proportion of phenotypic variation explained by SNPs of large effect (B; PGE), which are defined as SNPs with a non-zero effect on the phenotype, and the number of large-effect SNPs needed to explain PGE (C; nSNPs). Each blue line represents one of ten independent runs of the BSLMM. D) Posterior inclusion probability for 20 kb windows across all scaffolds (alternating black and grey for each scaffold). Windows that showed PIP values above the 99th percentile (0.00175; dotted red line) were considered to have a significant effect on jaw size variation. Red arrows indicate genes within top four windows (*samd12*, *clk2*, *gpr119*, *doc2b*, *rapgef4*).

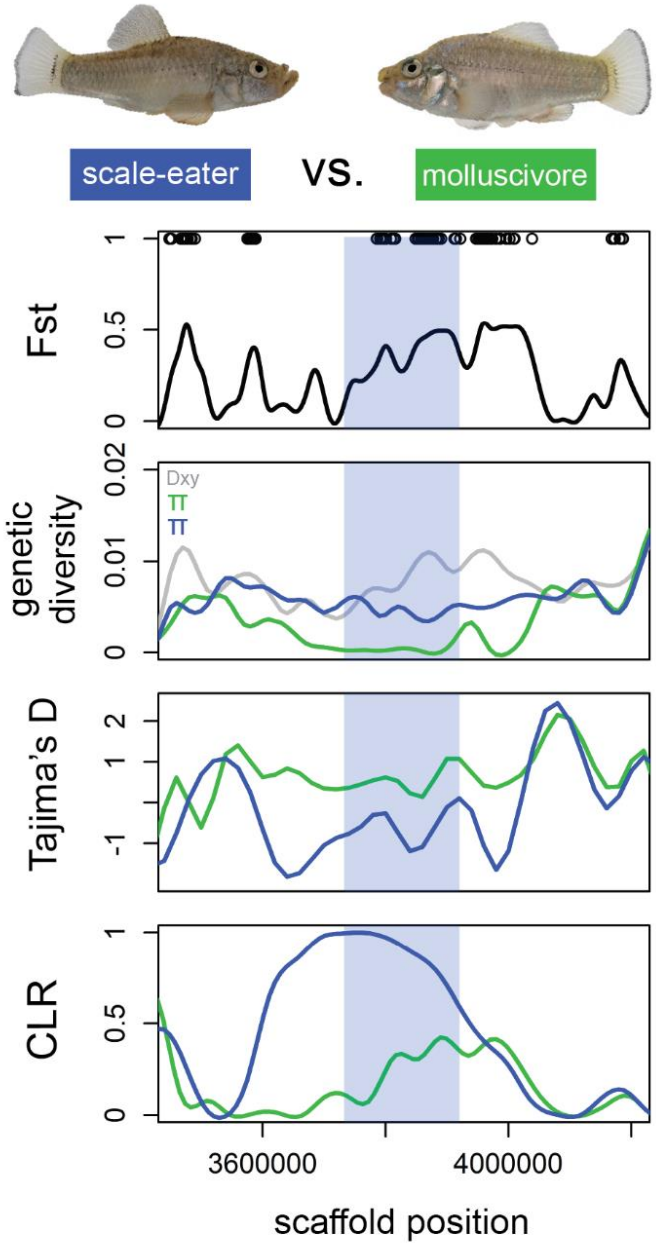

Figure D2.14. The sema6c gene region.

The *sema6c* gene region (light blue) contains 64 SNPs fixed between Osprey Lake scale-eaters (blue) vs. molluscivores (green), shows strong between-population divergence and low within-population diversity, shows strong signs of a hard selective sweep, and is significantly associated with oral jaw length variation in a genome-wide association analysis using GEMMA (Table D1.8).

APPENDIX E: SUPPLEMENTARY MATERIAL FOR CHAPTER 5

E1. Supplemental Tables

Table E1.1. Protein coding genes near 157 SNPs and 87 deletions fixed between molluscivores and scale-eaters.

Protein coding genes within 10 kb of the first or last exon

near fixed SNP	near fixed deletion
cckar	*acat2*
cdc14ab	*acvr1c*
cdk5r1	*adra2db*
cxcr1	*cckar*
dapk2	*cep170*
derl1	*col12a1*
dysf	*ctnnb1*
eef1d	*dph5*
fev	*dync2li1*
gimap2	*eef1a1*
nabp1	*fam219a*
nat14	*fgfr2*
nsmce2	*gm11992*
polg	*gpa33*
prpf4b	*hint1*
pxk	*hlf*
pycr3	*hlh-13*
sbk2	*irf1*
sgk1	*kcnq5*
slc25a29	*lyrm7*
slc38a2	*med25*
vrtn	*mprip*
washc5	*ncl1*
wdr78	*odf3l2*
wnt7b	*pdhb*
zhx2	*pld5*
znf628	*pxk*
	rabgap1l
	sh3pxd2a
	shisa2
	slc30a7
	u2af2
	upp2
	znf865

Table E1.2. Cross design used to produce RNA sequencing libraries for F1 offspring sampled at 2 days post fertilization (dpf), 8 dpf, and 20 dpf.

CP = Crescent Pond, OL = Osprey Lake, and LL = Little Lake.

Mother	Father	Stage	Libraries	F1
CP molluscivore	CP scale-eater	2 dpf	3	hybrid
OL scale-eater	OL molluscivore	2 dpf	3	hybrid
CP molluscivore	CP scale-eater	8 dpf	3	hybrid
OL scale-eater	OL molluscivore	8 dpf	3	hybrid
CP molluscivore	CP molluscivore	2 dpf	3	purebred
CP scale-eater	CP scale-eater	2 dpf	3	purebred
OL molluscivore	OL molluscivore	2 dpf	3	purebred
OL scale-eater	OL scale-eater	2 dpf	3	purebred
CP molluscivore	CP molluscivore	8 dpf	3	purebred
CP scale-eater	CP scale-eater	8 dpf	5	purebred
OL molluscivore	OL molluscivore	8 dpf	5	purebred
OL scale-eater	OL scale-eater	8 dpf	5	purebred
CP molluscivore	CP molluscivore	20 dpf	3	purebred
CP scale-eater	CP scale-eater	20 dpf	2	purebred
LL molluscivore	LL molluscivore	20 dpf	3	purebred

Table E1.3. Predicted transcription factor binding sites altered by genetic variants fixed between species.

Binding motifs from JASPAR database.

gene region	allele	transcription factor	matrix ID	binding sequence	relative profile score
dync2li1	reference	NFIC	MA0161.1	TTGGCA	1.00000
dync2li1	reference	NFIA	MA0670.1	AATGCCAAGT	0.98703
dync2li1	reference	NFIX	MA0671.1	AATGCCAAG	0.98551
dync2li1	reference	ZNF384	MA1125.1	TCAGAAAAAAAA	0.96526
dync2li1	reference	HOXA5	MA0158.1	CTGTAATT	0.96148
dync2li1	reference	Gata1	MA0035.1	TGATGC	0.95591
dync2li1	reference	MYB	MA0100.3	CACAACTGGC	0.95232
dync2li1	reference	Prrx2	MA0075.1	AATTA	1.00000
dync2li1	reference	Stat5a	MA1624.1	GTTCCAAGAATT	0.98454
dync2li1	alternate	Prrx2	MA0075.1	AATTA	1.00000
dync2li1	alternate	Stat5a	MA1624.1	GTTCCAAGAATT	0.98454
pycr	reference	GATA2	MA0036.1	AGATA	0.97565
pycr	reference	MZF1	MA0056.1	GGGGGA	0.96199
pycr	alternate	PLAGL2	MA1548.1	TGGGCCCCCA	0.98454
pycr	alternate	GATA2	MA0036.1	AGATA	0.97565

E2. Supplemental Figures

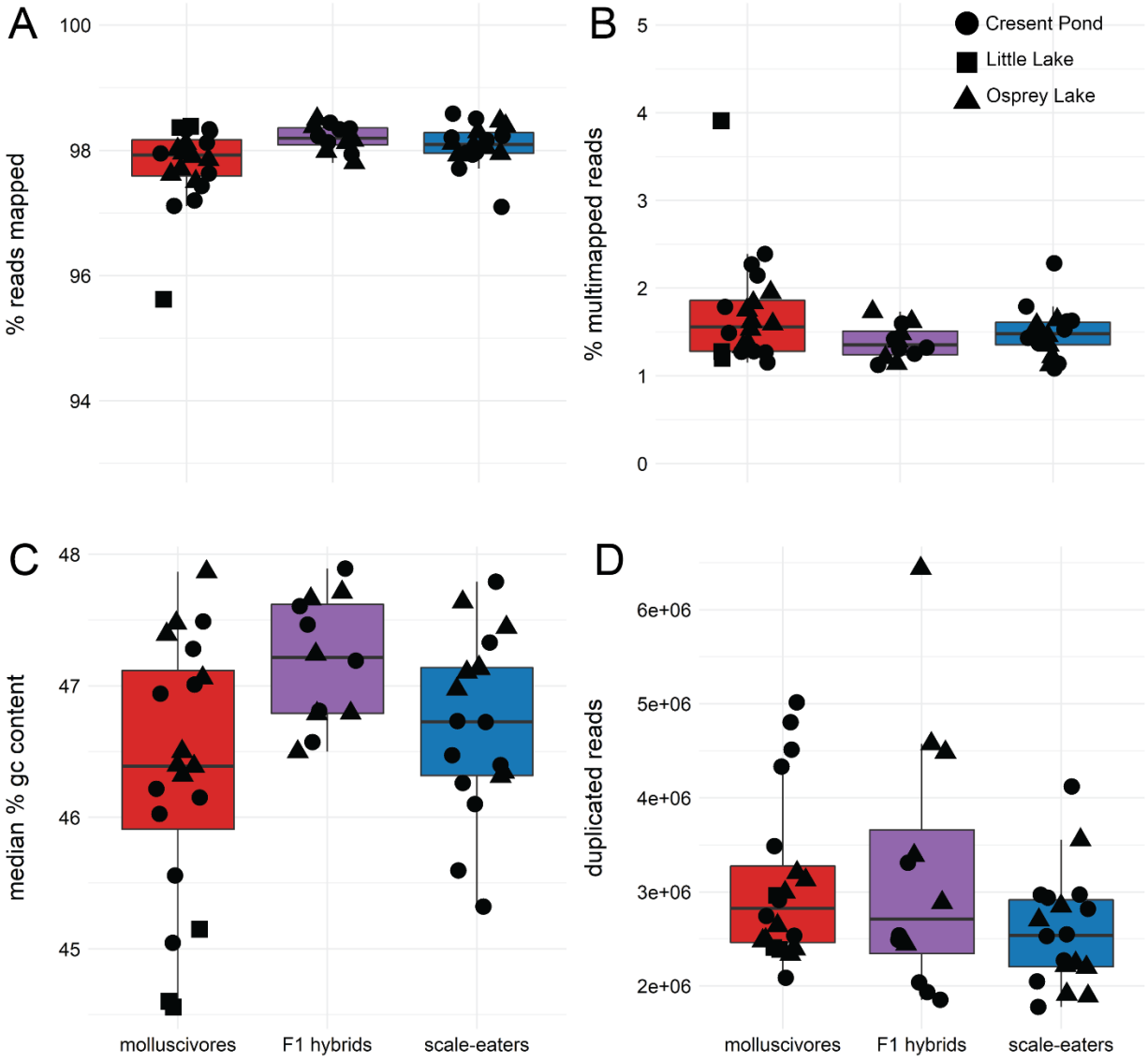

Figure E2.1. Quality control measures for 50 RNAseq libraries.

We did not find a difference between scale-eaters and mollusciores in A) the proportion of reads uniquely mapped to the molluscivore reference genome (Student's t-test, $P = 0.061$), B) the proportion of multimapped reads (Student's t-test, $P = 0.14$) ,C) the median GC content of aligned reads (Student's t-test, $P = 0.22$), or D) the number of duplicate reads (Student's t-test, $P = 0.05$).

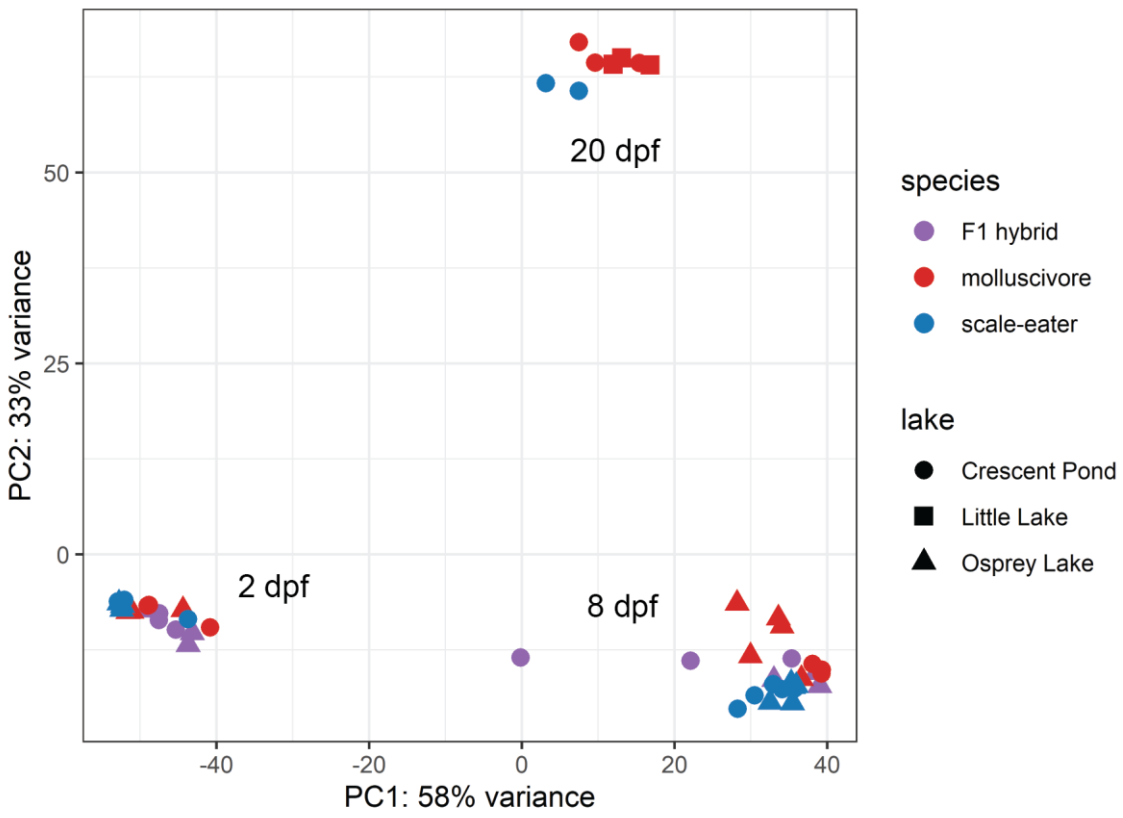

Figure E2.2. Principal component analysis for 50 transcriptomes.

Principal component analysis for 50 transcriptomes showing first two axes accounting for a combined 91% of the total variation in read counts normalized for library size.

REFERENCES

Abzhanov, A., M. Protas, B. R. Grant, P. R. Grant, and C. J. Tabin. 2011. Variation of Beaks in Darwin's Finches. Science 80:1462–1466.

Ahi, E. P., K. H. Kapralova, A. Pálsson, V. H. Maier, J. Gudbrandsson, S. S. Snorrason, Z. O. Jónsson, and S. R. Franzdóttir. 2014. Transcriptional dynamics of a conserved gene expression network associated with craniofacial divergence in Arctic charr. EvoDevo 5:1–19.

Alberti, A., C. Belser, S. Engelen, L. Bertrand, C. Orvain, L. Brinas, C. Cruaud, L. Giraut, C. Da Silva, C. Firmo, J. M. Aury, and P. Wincker. 2014. Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. BMC Genomics 15:1–13.

Albertson, R. C., W. Cresko, H. W. D. Iii, and J. H. Postlethwait. 2008. Evolutionary mutant models for human disease. Trends. Genet. 25:74-81.

Albertson, R. C., J. T. Streelman, and T. D. Kocher. 2003. Directional selection has shaped the oral jaws of Lake Malawi cichlid fishes. Proc. Natl. Acad. Sci. U. S. A. 100:5252–57.

Albertson, R. C., J. T. Streelman, T. D. Kocher, and P. C. Yelick. 2005. Integration and evolution of the cichlid mandible: The molecular basis of alternate feeding strategies. Proc. Natl. Acad. Sci. 102:16287–92.

Andolfatto, P. 2001. Adaptive hitchhiking effects on genome variability. Curr. Opin. Genet. Dev. 11:635–41.

Arnegard, M. E., M. D. McGee, B. Matthews, K. B. Marchinko, G. L. Conte, S. Kabir, N. Bedford, S. Bergek, Y. F. Chan, F. C. Jones, D. M. Kingsley, C. L. Peichel, and D. Schluter. 2014. Genetics of ecological divergence during speciation. Nature 511:307–311.

Barreto, F. S., R. J. Pereira, and R. S. Burton. 2015. Hybrid dysfunction and physiological compensation in gene expression. Mol. Biol. Evol. 32:613–622.

Barrett, R. D. H., and D. Schluter. 2008. Adaptation from standing genetic variation. Trends Ecol. Evol. 23:38–44.

Bastian, F., G. Parmentier, J. Roux, S. Moretti, U. De Lyon, I. De Génomique, F. De Lyon, and E. N. S. Lyon. 2008. Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. 124–131.

Baxter, S. W., S. E. Johnston, and C. D. Jiggins. 2009. Butterfly speciation and the distribution of gene effect sizes fixed during adaptation. Heredity 102:57–65.

Bedford, T., and D. L. Hartl. 2009. Optimization of gene expression by natural selection. Proc. Natl. Acad. Sci. 106:1133–38.

Bell, G. D. M., N. C. Kane, L. H. Rieseberg, and K. L. Adams. 2013. RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. Genome Biol. Evol. 5:1309–1323.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B. 57:289-300.

Biosystems, K. 2014. Technical Data Sheet KAPA Stranded mRNA-Seq Kit. 1–16.

Bono, J. M., E. C. Olesnicky, and L. M. Matzkin. 2015. Connecting genotypes, phenotypes and fitness: Harnessing the power of CRISPR/Cas9 genome editing. Mol. Ecol. 24:3810–22.

Boyle, E. A., Y. I. Li, and J. K. Pritchard. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169:1177–86.

Brian K. Hall. 2009. The Neural Crest and Neural Crest Cells in Vertebrate Development and Evolution. Springer US, New York.

Brueton, L. A., M. J. Dillon, and R. M. Winter. 1990. Ellis-van Creveld syndrome, Jeune Syndrome, and renal-hepatic-pancreatic dysplasia: Separate entities or disease spectrum? J. Med. Genet. 27:252–55.

Byers, K. J. R. P., S. Xu, and P. M. Schlüter. 2017. Molecular mechanisms of adaptation and speciation: why do we need an integrative approach? Mol. Ecol. 26:277-90.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST + : architecture and applications. BMC Bioinformatics. 10:421.

Carneiro, M., F. W. Albert, S. Afonso, R. J. Pereira, H. Burbano, R. Campos, J. Melo-Ferreira, J. A. Blanco-Aguiar, R. Villafuerte, M. W. Nachman, J. M. Good, and N. Ferrand. 2014. The genomic architecture of population divergence between subspecies of the European Rabbit. PLoS Genet. 10:1.

Carroll, S. B. 2008. Perspective Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell. 134:25–36.

Chan, Y. F., M. E. Marks, F. C. Jones, G. V. Jr, M. D. Shapiro, S. D. Brady, A. M. Southwick, D. M. Absher, J. Grimwood, J. Schmutz, R. M. Myers, D. Petrov, B. Jónsson, D. Schluter, M. A. Bell, and D. M. Kingsley. 2010. Adaptive Evolution of Pelvic Reduction of a Pitx1 Enhancer. Science 327:302–6.

Chaves, J. A., E. A. Cooper, A. P. Hendry, J. Podos, J. Albert, and C. Uy. 2016. Genomic variation at the tips of the adaptive radiation of Darwin's finches. Mol. Ecol. 25:5282-95.

Chen, D. H., Q. W. Wu, X. D. Li, S. J. Wang, and Z. M. Zhang. 2017. SYPL1 overexpression predicts poor prognosis of hepatocellular carcinoma and associates with epithelial-mesenchymal transition. Oncol. Rep. 38:1533–42.

Cleves, P. A., N. A. Ellis, M. T. Jimenez, S. M. Nunez, D. Schluter, D. M. Kingsley, and C. T. Miller. 2014. Evolved tooth gain in sticklebacks is associated with a cis-regulatory allele of Bmp6. Proc. Natl. Acad. Sci. 111:13912–17.

Cole, D. G. 2003. The intraflagellar transport machinery of Chlamydomonas reinhardtii. Traffic 4:435–42.

Comeault, A. A., C. F. Carvalho, S. Dennis, and P. Nosil. 2016. Color phenotypes are under similar genetic control in two distantly related species of Timema stick insect. Evolution 70:1283–96.

Comeault, A. A., A. Serrato-capuchina, D. A. Turissini, P. J. Mclaughlin, J. R. David, and D. R. Matute. 2017. A nonrandom subset of olfactory genes is associated with host preference in the fruit fly Drosophila orena. Evol. Letters 1:1–13.

Comeault, A. a, V. Soria-Carrasco, Z. Gompert, T. E. Farkas, C. A. Buerkle, T. L. Parchman, and P. Nosil. 2014. Genome-wide association mapping of phenotypic traits subject to a range of intensities of natural selection in Timema cristinae. Am. Nat. 183:711–27.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-cabrero, A. Cervera, A. Mcpherson, W. Szcze, D. J. Gaffney, L. L. Elo, and X. Zhang. 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 17:13.

Conte, G. L., M. E. Arnegard, C. L. Peichel, and D. Schluter. 2012. The probability of genetic parallelism and convergence in natural populations. Proc. R. Soc. B Biol. Sci. 279:5039–47.

Conway, J. R., A. Lex, and N. Gehlenborg. 2017. UpSetR: An R package for the visualization of intersecting sets and their properties. Bioinformatics 33:2938–40.

Coolon, J. D., C. J. Mcmanus, K. R. Stevenson, J. D. Coolon, C. J. Mcmanus, K. R. Stevenson, B. R. Graveley, and P. J. Wittkopp. 2014. Tempo and mode of regulatory evolution in Drosophila Tempo and mode of regulatory evolution in Drosophila. Genome Research 24:797–808.

Cooper, T. F., D. E. Rozen, and R. E. Lenski. 2003. Parallel changes in gene expression after 20,000 generations of evolution in Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 100:1072-77.

Corbett-detig, R. B., J. Zhou, A. G. Clark, D. L. Hartl, and J. F. Ayroles. 2013. Genetic incompatibilities are widespread within species. Nature 504:135–137.

Cowles, C. R., J. N. Hirschhorn, D. Altshuler, and E. S. Lander. 2002. Detection of regulatory variation in mouse genes. Nat. Genet. 32:432–37.

Coyne, J. A. & Orr. 2004. Speciation. Sunderland, MA Sinauer Assoc.

Coyne, J. A., and H. A. Orr. 1989. Patterns of Speciation in Drosophila. Evolution. 43:362–81.

Cruickshank, T. E., and M. W. Hahn. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23:3133–57.

Cutter, A. D. 2012. The polymorphic prelude to Bateson – Dobzhansky – Muller incompatibilities. Trends Ecol. Evol. 27:210–19.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–8.

Davidson, J. H., and C. N. Balakrishnan. 2016. Gene Regulatory Evolution During Speciation in a Songbird. 6:1357–64.

Degner, J. F., J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25:3207–12.

Denver, D. R., K. Morris, J. T. Streelman, S. K. Kim, M. Lynch, and W. K. Thomas. 2005. The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat. Genet. 37:544-48.

DePristo, M. A., E. Banks, R. Poplin, K. V Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–98.

Derome, N., and L. Bernatchez. 2006. The Transcriptomics of Ecological Convergence between 2 Limnetic Coregonine Fishes (Salmonidae). Mol. Biol. Evol. 2006 23:2370-78

Dittmar, E. L., C. G. Oakley, J. K. Conner, B. A. Gould, and D. W. Schemske. 2016. Factors influencing the effect size distribution of adaptive substitutions. Proc. Biol. Sci. 283:1828.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21.

Echelle, A. A., E. W. Carson, A. F. Echelle, R. A. Van Den Bussche, T. E. Dowling, and A. Meyer. 2005. Historical Biogeography of the New-World Pupfish Genus Cyprinodon (Teleostei: Cyprinodontidae). Copeia 2005:320–39.

Enard, W., P. Khaitovich, J. Klose, F. Heissig, P. Giavalisco, K. Nieselt-struwe, E. Muchmore, A. Varki, R. Ravid, G. M. Doxiadis, and R. E. Bontrop. 2002. Intra- and Interspecific Variation in Primate Gene Expression Patterns. Science 296:340–344.

Erickson, P. A., J. Baek, J. C. Hart, P. A. Cleves, and C. T. Miller. 2018. Genetic dissection of a supergene implicates tfap2a in craniofacial evolution of threespine sticklebacks. Genetics 209:591–605.

Ewels, P., S. Lundin, and K. Max. 2016. Data and text mining MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048.

Fang, S., R. Yukilevich, Y. Chen, D. A. Turissini, K. Zeng, I. A. Boussy, and C. I. Wu. 2012. Incompatibility and competitive exclusion of genomic segments between sibling Drosophila species. PLoS Genet. 8:6.

Felsenstein, J. 1985. Phylogenies and the Comparative Method. Am. Nat., doi: 10.1086/284325.

Ferna, A., J. J. Tena, C. Gonza, H. Parra-acero, J. W. Cross, P. W. J. Rigby, J. J. Carvajal, J. Wittbrodt, and J. R. Martı. 2014. Comparative epigenomics in distantly related teleost species identifies conserved cis -regulatory nodes active during the vertebrate phylotypic period. Genome Res. 24:1075–85.

Fornes, O., J. A. Castro-Mondragon, A. Khan, R. van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, W. Santana-Garcia, G. Tan, J. Chèneby, B. Ballester, F. Parcy, A. Sandelin, B. Lenhard, W. W. Wasserman, and A. Mathelier. 2019. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 48:87-92.

Fournier-Level, A., A. Korte, M. D. Cooper, M. Nordborg, J. Schmitt, and A. M. Wilczek. 2011. A map of local adaptation in Arabidopsis thaliana. Science 334:86–89.

Fritz, D. I., J. M. Johnston, and A. H. Chishti. 2014. MPP1/p55 gene deletion in a hemophilia A patient with ectrodactyly and severe developmental defects. Am. J. Hematol. 94:29–32.

Furutani-Seiki, M., and J. Wittbrodt. 2004. Medaka and zebrafish, an evolutionary twin study. Mech. Dev. 121:629–37.

Gaboli, M., P. A. Kotsi, C. Gurrieri, G. Cattoretti, S. Ronchetti, C. Cordon-Cardo, H. E. Broxmeyer, R. Hromas, and P. P. Pandolfi. 2001. Mzf1 controls cell proliferation and tumorigenesis. Genes Dev. 15:1625–30.

Galtier, N., F. Depaulis, and N. H. Barton. 2000. Detecting Bottlenecks and Selective Sweeps From DNA Sequence Polymorphism. Genetics 155: 981–987.

Garfield, D. A., D. E. Runcie, C. C. Babbitt, R. Haygood, W. J. Nielsen, and G. A. Wray. 2013. The Impact of Gene Expression Variation on the Robustness and Evolvability of a Developmental Gene Regulatory Network. PLoS Biol. 11:e1001696.

Garud, N. R., P. W. Messer, E. O. Buzbas, and D. A. Petrov. 2015. Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. PLoS Genet. 11:1–32.

Ge, S. X., and D. Jung. 2018. ShinyGO: a graphical enrichment tool for animals and plants. Bioinformatics 1:1-2.

Gene Ontology Consortium. 2017. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 45:331-38.

Gompert, Z., L. K. Lucas, C. C. Nice, and C. A. Buerkle. 2012. Genome divergence and the genetic architecture of barriers to gene flow between Lycaeides idas and L. melissa. Evolution 67:2498–2514.

Goncalves, A., S. Leigh-Brown, D. Thybert, K. Stefflova, E. Turro, P. Flicek, A. Brazma, D. T. Odom, and J. C. Marioni. 2012. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome Res. 22:2376–2384.

Gorlin, R. J., M. M. Cohen, and R. Hennekam. 1990. Syndromes of the head and neck. Oxford University Press, Oxford, UK.

Griswold, C. K. 2006. Gene flow's effect on the genetic architecture of a local adaptation and its consequences for QTL analyses. Heredity 96:445–453.

Gross, J. B., and A. K. Powers. 2018. A Natural Animal Model System of Craniofacial Anomalies: The Blind Mexican Cavefish. Anat. Rec. 303:24-29.

Guerrero, R. F., A. L. Posto, L. C. Moyle, and M. W. Hahn. 2016. Genome-wide patterns of regulatory divergence revealed by introgression lines. Evolution 70:696-706.

Haerty, W., and R. S. Singh. 2006. Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of Drosophila. Mol. Biol. Evol. 23:1707–14.

Hansen, T. F., J. Pienaar, and S. H. Orzack. 2008. A comparative method for studying adaptation to a randomly evolving environment. Evolution 62:1965–77.

He, X., and J. Zhang. 2006. Toward a Molecular Understanding of Pleiotropy. Genetics. 1891:1885–91.

Helms, J. A., D. Cordero, and M. D. Tapadia. 2005. New insights into craniofacial morphogenesis. Development 132:851–861.

Helms, J. A., and R. A. Schneider. 2003. Cranial skeletal biology. Nature. 423:326-31.

Hermisson, J., and P. S. Pennings. 2005. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335–52.

Hernandez, L. P., D. Adriaens, C. H. Martin, P. C. Wainwright, B. Masschaele, and M. Dierick. 2018. Building trophic specializations that result in substantial niche partitioning within a young adaptive radiation. J. Anat. 232:173–85.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. 106:9362–67.

Ho, S. S., A. E. Urban, and R. E. Mills. 2019. Structural variation in the sequencing era. Nat. Rev. Genet. 21:171-89.

Hoban, S., J. L. Kelley, K. E. Lotterhos, M. F. Antolin, G. Bradburd, D. B. Lowry, M. L. Poss, L. K. Reed, A. Storfer, and M. C. Whitlock. 2016. Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. Am. Nat. 188:379–97.

Hodgins-Davis, A., D. P. Rice, J. P. Townsend, and J. Novembre. 2015. Gene expression evolves under a house-of-cards model of stabilizing selection. Mol. Biol. Evol. 32:2130–40.

Holtmeier, C. L. 2001. Heterochrony, maternal effects, and phenotypic variation among sympatric pupfishes. Evolution 55:330–338.

Huang, S., W. Zhang, X. Chang, and J. Guo. 2019. Overlap of periodic paralysis and paramyotonia congenita caused by SCN4A gene mutations. Channels 13:110-19.

Huber, C., and V. Cormier-Daire. 2012. Ciliary disorder of the skeleton. Am. J. Med. Genet. Part C Semin. Med. Genet. 160:165–74.

Hufnagel, S. B., K. N. Weaver, R. B. Hufnagel, P. I. Bader, E. K. Schorry, and R. J. Hopkin. 2014. A novel dominant COL11A1 mutation resulting in a severe skeletal dysplasia. Am. J. Med. Genet. Part A 164:2607–12.

Hughes, K. A., J. F. Ayroles, M. M. Reedy, J. M. Drnevich, K. C. Rowe, E. A. Ruedi, C. E. Cáceres, and K. N. Paige. 2006. Segregating variation in the transcriptome: Cis regulation and additivity of effects. Genetics 173:1347–1355.

Humphries, J. M., R. R. Miller. 1981. A Remarkable Species Flock of Pupfishes Genus Cyprinodon from Yucatán, México. Copeia 1:52–64.

Hunter, D. J., P. Kraft, K. B. Jacobs, D. G. Cox, M. Yeager et al. 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat. Genet. 39:870-74.

Indjeian, V. B., G. A. Kingman, F. C. Jones, C. A. Guenther, J. Grimwood, J. Schmutz, R. M. Myers, and D. M. Kingsley. 2016. Evolving New Skeletal Traits by cis-Regulatory Changes in Bone Morphogenetic Proteins. Cell 164:45–56.

Invernizzi, F., M. Tigano, C. Dallabona, C. Donnini, I. Ferrero, M. Cremonte, D. Ghezzi, C. Lamperti, and M. Zeviani. 2013. A Homozygous Mutation in LYRM7 / MZM1L Associated with Early Onset Encephalopathy, Lactic Acidosis, and Severe Reduction of Mitochondrial Complex III Activity. Hum Mutat. 34:1619-22.

Irwin, D. E., M. Alcaide, K. E. Delmore, J. H. Irwin, and G. L. Owens. 2016. Recurrent selection explains parallel evolution of genomic regions of high relative but low absolute differentiation in a ring species. Mol. Ecol. 25:4488–4507.

Jensen, J. D. 2014. On the unfounded enthusiasm for soft selective sweeps. Nat. Commun. 5:5281.

Johnson, N. A., and A. H. Porter. 2000. Rapid speciation via parallel, directional selection on regulatory genetic pathways. J. Theor. Biol. 205:527–42.

Johnson, R. C., G. W. Nelson, J. L. Troyer, J. A. Lautenberger, B. D. Kessing, C. A. Winkler, and S. J. O'Brien. 2010. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics 11:724.

Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55-61

Kalinka, A. T., K. M. Varga, D. T. Gerrard, S. Preibisch, D. L. Corcoran, J. Jarrells, U. Ohler, C. M. Bergman, and P. Tomancak. 2010. developmental hourglass model. Nature 468:811–14.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42:348–54.

Kang, P., and K. K. H. Svoboda. 2005. Epithelial-mesenchymal transformation during craniofacial development. J. Dental Res., 84:678–90.

Kerwin, R. E., and A. L. Sweigart. 2020. Rampant misexpression in a Mimulus (monkeyflower) introgression line caused by hybrid sterility, not regulatory divergence. Mol. Biol. Evol. doi:10.1093/molbev/msaa071

Kessler, K., I. Wunderlich, S. Uebe, N. S. Falk, A. Gießl, J. Helmut Brandstätter, B. Popp, P. Klinger, A. B. Ekici, H. Sticht, H. G. Dörr, A. Reis, R. Roepman, E. Seemanova, and C. T. Thiel. 2015. DYNC2LI1 mutations broaden the clinical spectrum of dynein-2 defects. Sci. Rep. 5:1–12.

Kratochwil, C. F., and A. Meyer. 2015. Closing the genotype-phenotype gap: Emerging technologies for evolutionary genetics in ecological model vertebrate systems. BioEssays 37:213–26.

Kronforst, M. R., G. S. Barsh, A. Kopp, J. Mallet, A. Monteiro, S. P. Mullen, M. Protas, E. B. Rosenblum, C. J. Schneider, and H. E. Hoekstra. 2012. Unraveling the thread of nature's tapestry: the genetics of diversity and convergence in animal pigmentation. Pigment Cell Melanoma Res. 25:411-33.

Kulmuni, J., and A. M. Westram. 2017. Intrinsic incompatibilities evolving as a by-product of divergent ecological selection: Considering them in empirical studies on divergence with gene flow. Mol. Ecol. 26:3093–3103

Lamichhaney, S., J. Berglund, M. S. Almén, K. Maqbool, M. Grabherr, A. Martinez-Barrio, M. Promerová, C.-J. Rubin, C. Wang, N. Zamani, B. R. Grant, P. R. Grant, M. T. Webster, and L. Andersson. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature 518:371–375.

Lamichhaney, S., F. Han, J. Berglund, C. Wang, M. S. Almen, M. T. Webster, B. R. Grant, P. R. Grant, and L. Andersson. 2016. A beak size locus in Darwins finches facilitated character displacement during a drought. Science 352:470–74.

Landry, C. R., D. L. Hartl, and J. M. Ranz. 2007. Genome clashes in hybrids: Insights from gene expression. Heredity. 99:483–93.

Landry, C. R., P. J. Wittkopp, C. H. Taubes, J. M. Ranz, A. G. Clark, and D. L. Hartl. 2005. Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of drosophila. Genetics 171:1813–22.

Larson, E. L., D. Vanderpool, B. A. J. Sarver, C. Callahan, S. Keeble, L. L. Provencio, M. D. Kessler, V. Stewart, E. Nordquist, M. D. Dean, and J. M. Good. 2018. The Evolution of Polymorphic Hybrid Incompatibilities in House Mice. Genetics 209:845–59.

Lee, S., D. Cook, and M. Lawrence. 2019. Plyranges: A grammar of genomic data transformation. Genome Biol. 20:1–10.

Lencer, E. S., and A. R. McCune. 2018. An embryonic staging series up to hatching for Cyprinodon variegatus: An emerging fish model for developmental, evolutionary, and ecological research. J. Morphol. 279:1559–78.

Lencer, E. S., M. L. Riccio, and A. R. McCune. 2016. Changes in growth rates of oral jaw elements produce evolutionary novelty in bahamian pupfish. J. Morphol. 277:935–47.

Lencer, E. S., W. C. Warren, R. Harrison, and A. R. Mccune. 2017. The Cyprinodon variegatus genome reveals gene expression changes underlying differences in skull morphology among closely related species. BMC Genomics 18:424.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–60.

Li, H., and R. Durbin. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493–96.

Liao, Y., G. K. Smyth, and W. Shi. 2014. Sequence analysis featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. 30:923–30.

Lin, Y., K. Golovnina, Z. Chen, H. N. Lee, Y. L. S. Negron, H. Sultana, B. Oliver, and S. T. Harbison. 2016. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. BMC Genomics 17:28.

Losos, J. B. 2011. Convergence, adaptation, and constraint. Evolution. 65:1827–40.

Losos, J. B., and R. E. Ricklefs. 2009. Adaptation and diversification on islands. Nature 457:830–36.

Love, M. I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 15:550.

Mabee, P., J. P. Balhoff, W. M. Dahdul, H. Lapp, P. E. Midford, T. J. Vision, and M. Westerfield. 2012. 500,000 fish phenotypes: The new informatics landscape for evolutionary and developmental biology of the vertebrate skeleton. J. Appl. Ichthyol. 28:300–5.

Mack, K. L., P. Campbell, and M. W. Nachman. 2016. Gene regulation and speciation in house mice. Genome Res. 26:451–61.

Mack, K. L., and M. W. Nachman. 2017. Gene Regulation and Speciation. Trends Genet. 33:68–80.

Madeira, F., Y. M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A. R. N. Tivey, S. C. Potter, R. D. Finn, and R. Lopez. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47:W636–W641.

Maheshwari, S., and D. A. Barbash. 2012. Cis-by-trans regulatory divergence causes the asymmetric lethal effects of an ancestral hybrid incompatibility gene. PLoS Genet. 8:3.

Malinsky, M., R. J. Challis, A. M. Tyers, S. Schiffels, Y. Terai, B. P. Ngatunga, E. A. Miska, R. Durbin, M. J. Genner, and G. F. Turner. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. Science 350:1493–98.

Malone, J. H., and P. Michalak. 2008. Gene expression analysis of the ovary of hybrid females of Xenopus laevis and X. muelleri. BMC Evol. Biol. 8:1.

Manceau, M., V. S. Domingues, R. Mallarino, and H. E. Hoekstra. 2011. The developmental role of Agouti in color pattern evolution. Science 331:1062-65.

Manda, P., J. P. Balhoff, H. Lapp, P. Mabee, and T. J. Vision. 2015. Using the Phenoscape Knowledgebase to Relate Genetic Perturbations to Phenotypic Evolution. 571:561–71.

Manousaki, T., P. M. Hull, H. Kusche, G. Machado-schiaffino, P. Franchini, and C. Harrod. 2013. Parsing parallel evolution: ecological divergence and differential gene expression in the adaptive radiations of thick-lipped Midas cichlid fishes from Nicaragua. Mol. Ecol. 22:650–69.

Marchini, J., L. R. Cardon, M. S. Phillips, and P. Donnelly. 2004. The effects of human population structure on large genetic association studies. Nat. Genet. 36:512–17.

Marsden, C. D., Y. Lee, K. Kreppel, A. Weakley, A. Cornel, H. M. Ferguson, E. Eskin, and G. C. Lanzaro. 2014. Diversity, differentiation, and linkage disequilibrium: prospects for association mapping in the malaria vector Anopheles arabiensis. G3. 4:121–31.

Martin, A., and V. Orgogozo. 2013. The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. Evolution. 67:1235-50.

Martin, C. H., and P. C. Wainwright. 2011. Trophic novelty is linked to exceptional rates of morphological diversification in two adaptive radiations of Cyprinodon pupfish. Evolution 65:2197–212.

Martin, C. H., and P. C. Wainwright. 2013a. A remarkable species flock of Cyprinodon pupfishes endemic to San Salvador Island, Bahamas. Bull. Peabody Museum Nat. Hist. 54:231–41.

Martin, C. H., and P. C. Wainwright. 2013b. Multiple fitness peaks on the adaptive landscape drive adaptive radiation in the wild. Science 339:208–11.

Martin, C. H., and P. C. Wainwright. 2013c. On the measurement of ecological novelty: scale-eating pupfish are separated by 168 my from other scale-eating fishes. PLoS One 8:e71164.

Martin, C. H., and L. C. Feinstein. 2014. Novel trophic niches drive variable progress towards ecological speciation within an adaptive radiation of pupfishes. Mol. Ecol. 23:1846–1862.

Martin, C. H. 2016a. Context-dependence in complex adaptive landscapes: frequency and trait-dependent selection surfaces within an adaptive radiation of Caribbean pupfishes. Evolution 70:1265-1282.

Martin, C. H. 2016b. The cryptic origins of evolutionary novelty: 1000-fold faster trophic diversification rates without increased ecological opportunity or hybrid swarm. Evolution. 70:2504-19.

Martin, C. H., P. A. Erickson, and C. T. Miller. 2017. The genetic architecture of novel trophic specialists: larger effect sizes are associated with exceptional oral jaw diversification in a pupfish adaptive radiation. Mol. Ecol. 26:624–38.

Martin, C. H., J. A. McGirr, E. J. Richards, and M. St. John. 2019a. How to investigate the origins of novelty: insights gained from genetic, behavioral, and fitness perspectives. Integr. Org. Biol., doi: 10.1093/iob/obz018.

Martin, C. H., and E. J. Richards. 2019b. The Paradox Behind the Pattern of Rapid Adaptive Radiation: How Can the Speciation Process Sustain Itself Through an Early Burst? Annu. Rev. Ecol. Evol. Syst. 50:1–25.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet, and C. D. Jiggins. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 23:1817–28.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. Science 337:1190–95.

Mayba, O., H. N. Gilbert, J. Liu, P. M. Haverty, S. Jhunjhunwala, Z. Jiang, C. Watanabe, and Z. Zhang. 2014. MBASED: Allele-specific expression detection in cancer tissues and cell lines. Genome Biol. 15:1–21.

Maynared Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genet. Res. 23:23–35.

McDonald, A. C., J. A. Schuijers, A. L. Gundlach, and B. L. Grills. 2007. Galanin treatment offsets the inhibition of bone formation and downregulates the increase in mouse

calvarial expression of TNF and GalR2 mRNA induced by chronic daily injections of an injurious vehicle. Bone 40:895–903.

McGirr, J. A., and C. H. Martin. 2017. Novel candidate genes underlying extreme trophic specialization in caribbean pupfishes. Mol. Biol. Evol. 34:873–88.

McGirr, J. A., and C. H. Martin. 2018. Parallel evolution of gene expression between trophic specialists despite divergent genotypes and morphologies. Evol. Lett. 2:62–75.

McGirr, J. A., and C. H. Martin. 2019a. Hybrid misregulation in multiple developing tissues within a recent adaptive radiation of Cyprinodon pupfishes. PLoS One 14:e0218899.

McGirr, J.A, and C. H. Martin. 2019b. Ecological divergence in sympatry causes gene misregulation in hybrids. bioRxiv 717025:1–71.

McGowan, H. W., J. A. Schuijers, B. L. Grills, S. J. McDonald, and A. C. McDonald. 2014. Galnon, a galanin receptor agonist, improves intrinsic cortical bone tissue properties but exacerbates bone loss in an ovariectomised rat model. J. Musculoskelet. Neuronal Interact. 14:162–72.

McManus, C. J., J. D. Coolon, M. O. Duff, J. Eipper-Mains, B. R. Graveley, and P. J. Wittkopp. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res. 20:816–25.

McVean, G. 2009. A genealogical interpretation of principal components analysis. PLoS Genet. 5:e1000686.

Michalak, P., and M. A. F. Noor. 2004. Association of misexpression with sterility in hybrids of Drosophila simulans and D. mauritiana. J. Mol. Evol. 59:277–82.

Midford, P. E., T. A. Dececchi, J. P. Balhoff, W. M. Dahdul, N. Ibrahim, H. Lapp, J. G. Lundberg, P. M. Mabee, P. C. Sereno, M. Westerfield, T. J. Vision, and D. C. Blackburn. 2013. The vertebrate taxonomy ontology: a framework for reasoning across model organism and species phenotypes. J. Biomed. Semantics 4:34.

Miller, C. T., S. Beleza, A. A. Pollen, D. Schluter, R. A. Kittles, M. D. Shriver, and D. M. Kingsley. 2007a. cis -Regulatory Changes in Kit Ligand Expression and Parallel Evolution of Pigmentation in Sticklebacks and Humans. Cell 131:1179–89.

Miller, C. T., M. E. Swartz, P. A. Khuu, M. B. Walker, J. K. Eberhart, and C. B. Kimmel. 2007b. mef2ca is required in cranial neural crest to effect Endothelin1 signaling in zebrafish. Dev. Biol. 308:144–157.

Moczek, A. P. 2008. On the origins of novelty in development and evolution. BioEssays 30:432–447.

Moriwaki, K., K. Noda, Y. Furukawa, K. Ohshima, A. Uchiyama, T. Nakagawa, N. Taniguchi, Y. Daigo, Y. Nakamura, N. Hayashi, and E. Miyoshi. 2009. Deficiency of GMDS Leads

to Escape from NK Cell-Mediated Tumor Surveillance Through Modulation of TRAIL Signaling. Gastroenterology 137:188-98.

Mylroie, J.E, Hagey, F. M. 1995. Terrestrial and Shallow Marine Geology of the Bahamas and Bermuda. Geological Society of America, Boulder, CO.

Nachman, M. W., and B. A. Payseur. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos. Trans. R. Soc. B Biol. Sci. 367:409–21.

Nadachowska-brzyska, K., R. Burri, and E. Linn. 2016. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. Mol. Ecol. 25:1058–72.

Nagai, H., Y. Terai, T. Sugawara, H. Imai, H. Nishihara, M. Hori, and N. Okada. 2011. Reverse Evolution in RH1 for Adaptation of Cichlids to Water Depth in Lake Tanganyika. Mol. Biol. Evol. 28:1769–76.

Nguyen, D., and X. Tian. 2008. The expanding role of mouse genetics for understanding human biology and disease. DMM Dis. Model. Mech. 1:56–66.

Niceta, M., K. Margiotti, M. C. Digilio, V. Guida, A. Bruselles, S. Pizzi, A. Ferraris, L. Memo, N. Laforgia, M. L. Dentici, F. Consoli, I. Torrente, V. L. Ruiz-Perez, B. Dallapiccola, B. De Luca, and M. Tartaglia. 2018. Biallelic mutations in DYNC2LI1 are a rare cause of Ellis-van Creveld syndrome. Clin. Genet. 93:632–39.

Nielsen, R. 2005. Molecular signatures of natural selection. Annu. Rev. Genet. 39:197–218.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. 2005a. Genomic scans for selective sweeps using SNP data. Genome Res. 15:1566–1575.

Noor, M. a F., and S. M. Bennett. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity 103:439–44.

Noor, M. a F., and J. L. Feder. 2006. Speciation genetics: evolving approaches. Nat. Rev. Genet. 7:851–61.

Nosil, P. 2012. Ecological Speciation and its alternatives. Oxford University Press.

Novembre, J., and M. Stephens. 2008. Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. 40:646–49.

O'Quin, K. E., C. M. Hofmann, H. A. Hofmann, and K. L. Carleton. 2010. Parallel Evolution of Opsin Gene Expression in African Cichlid Fishes. Mol Biol Evol. 27:2839-54.

Oliver, G. R., P. R. Blackburn, M. S. Ellingson, E. Conboy, F. Pinto e Vairo, M. Webley, E. Thorland, M. Ferber, E. Van Hul, I. M. van der Werf, W. Wuyts, D. Babovic-Vuksanovic, and E. W. Klee. 2019. RNA-Seq detects a SAMD12-EXT1 fusion transcript

and leads to the discovery of an EXT1 deletion in a child with multiple osteochondromas. Mol. Genet. Genomic Med. 7:1–13.

Orr, H. A. 1996. Anecdotal, Historical and Critical Commentaries on Genetics. Genetics 144:1331–35.

Orr, H. A. 1998. The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution. Evolution. 52:935–49.

Orr, H. A., and A. J. Betancourt. 2001. Haldane's sieve and adaptation from the standing genetic variation. Genetics 157:875–84.

Orr, H. A. 2005. The genetic theory of adaptation: a brief history. Nat. Rev. Genet. 6:119–27.

Ortíz-Barrientos, D., B. A. Counterman, and M. A. F. Noor. 2007. Gene expression divergence and the origin of hybrid dysfunctions. Genetica 129:71–81.

Otto, G. P., M. Y. Wu, N. Kazgan, O. R. Anderson, and R. H. Kessin. 2004. Dictyostelium Macroautophagy Mutants Vary in the Severity of Their Developmental Defects. J. Biol. Chem. 279:15621–29.

Pallares, L. F., B. Harr, L. M. Turner, and D. Tautz. 2014. Use of a natural hybrid zone for genomewide association mapping of craniofacial traits in the house mouse. Mol. Ecol. 23:5756–70.

Papakostas, S., L. A. Vøllestad, M. Bruneaux, T. Aykanat, J. Vanoverbeke, M. Ning, C. R. Primmer, and E. H. Leder. 2014. Gene pleiotropy constrains gene expression changes in fish adapted to different thermal conditions. Nature. Comm. 5:4071.

Paradis, E., and K. Schliep. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. 35:526–28.

Parekh, S., C. Ziegenhain, B. Vieth, W. Enard, and I. Hellmann. 2016. The impact of amplification on differential expression analyses by RNA-seq. Sci. Rep. 6:1–11.

Parry, J. W. L., K. L. Carleton, T. Spady, A. Carboo, D. M. Hunt, and J. K. Bowmaker. 2005. Mix and match color vision: Tuning spectral sensitivity by differential opsin gene expression in Lake Malawi cichlids. Curr. Biol. 15:1734–39.

Pavey, S. A., P. Nosil, and S. M. Rogers. 2010. The role of gene expression in ecological speciation. Ann. N. Y. Acad. Sci. 1206:110–29.

Pavlidis, P., D. Živković, A. Stamatakis, and N. Alachiotis. 2013. SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. Mol. Biol. Evol. 30:2224–34.

Pfister, K. K., P. R. Shah, H. Hummerich, A. Russ, J. Cotton, A. A. Annuar, S. M. King, and E. M. C. Fisher. 2006. Genetic analysis of the cytoplasmic dynein subunit families. PLoS Genet. 2:11–26.

Poelstra, J. W., N. Vijay, C. M. Bossu, H. Lantz, B. Ryll, I. Müller, V. Baglione, P. Unneberg, M. Wikelski, M. G. Grabherr, and J. B. W. Wolf. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science 344:1410–14.

Porto, A., R. Schmelter, J. L. Vandeberg, G. Marroig, and J. M. Cheverud. 2016. Evolution of the genotype-to-phenotype map and the cost of pleiotropy in mammals. Genetics 204:1601–12.

Powder, K. E., and R. C. Albertson. 2016. Cichlid fi shes as a model to understand normal and clinical craniofacial variation. Dev. Biol. 415:338–46.

Presgraves, D. C. 2003. A fine-scale genetic analysis of hybrid incompatibilities in Drosophila. Genetics 163:955–72.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38:904–09.

Pritchard, J. K., J. K. Pickrell, and G. Coop. 2010. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Curr. Biol. 20:R208–R215.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–75.

Puzey, J. R., J. H. Willis, and J. K. Kelly. 2017. Population structure and local selection yield high genomic variation in Mimulus guttatus. Mol. Ecol. 26:519-35.

Ranz, J. M., K. Namgyal, G. Gibson, D. L. Hartl, J. M. Ranz, K. Namgyal, G. Gibson, and D. L. Hartl. 2004. Anomalies in the Expression Profile of Interspecific Hybrids of Drosophila melanogaster and Drosophila simulans Anomalies in the Expression Profile of Interspecific Hybrids of Drosophila melanogaster and Drosophila simulans. Genome Res. 14:373–79.

Rausch, T., T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:333–39.

Rausher, M. D., and L. F. Delph. 2015. Commentary: When does understanding phenotypic evolution require identification of the underlying genes? Evolution 69:1655–64.

Recknagel, H., K. R. Elmer, and A. Meyer. 2013. A Hybrid Genetic Linkage Map of Two Ecologically and Morphologically Divergent Midas Cichlid Fishes (Amphilophus spp.) Obtained by Massively Parallel DNA Sequencing. G3 3:65–74.

Reddiex, A. J., T. P. Gosden, R. Bonduriansky, and S. F. Chenoweth. 2013. Sex-Specific Fitness Consequences of Nutrient Intake and the Evolvability of Diet Preferences. Am. Nat. 182:91-102.

Reed, L. K., and T. A. Markow. 2004. Early events in speciation: Polymorphism for hybrid male sterility in Drosophila. Proc. Nat. Acad. Sci. 101:9009–9012.

Reed, R. D., R. Papa, A. Martin, H. M. Hines, M. R. Kronforst, R. Chen et al. 2011. optix drives the repeated convergent evolution of butterfly wing pattern mimicry. Science. 333:1137-41

Reid, N. M., D. A. Proestou, B. W. Clark, W. C. Warren, J. K. Colbourne, J. R. Shaw, S. I. Karchner, M. E. Hahn, D. Nacci, M. F. Oleksiak, D. L. Crawford, and A. Whitehead. 2016. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. Science 354:1305-08

Renaut, S., and L. Bernatchez. 2011. Transcriptome-wide signature of hybrid breakdown associated with intrinsic reproductive isolation in lake whitefish species pairs (Coregonus spp. Salmonidae). Heredity 106:1003–11.

Renaut, S., A. W. Nolte, and L. Bernatchez. 2009. Gene expression divergence and hybrid misexpression between lake whitefish species pairs (Coregonus spp. Salmonidae). Mol. Biol. Evol. 26:925–36.

Richards, E. J., and C. H. Martin. 2017. Adaptive introgression from distant Caribbean islands contributed to the diversification of a microendemic adaptive radiation of trophic specialist pupfishes. PLoS Genet. 13:e1006919.

Riehle, M. M., A. F. Bennett, R. E. Lenski, A. D. Long, M. Michelle, A. F. Bennett, E. Richard, and A. D. Long. 2003. Evolutionary changes in heat-inducible gene expression in lines of Escherichia coli adapted to high temperature. Physiol. Genomics. 14:47-58.

Roberge, C., É. Normandeau, S. Einum, H. Guderley, and L. Bernatchez. 2008. Genetic consequences of interbreeding between farmed and wild Atlantic salmon: Insights from the transcriptome. Mol. Ecol. 17:314–24.

Roberts, R. B., Y. Hu, R. C. Albertson, and T. D. Kocher. 2011. Craniofacial divergence and ongoing adaptation via the hedgehog pathway. Proc. Natl. Acad. Sci. 108:13194–99.

Rockman, M. V., and L. Kruglyak. 2006. Genetics of global gene expression. Nat. Rev. Genet. 7:862–72.

Rogers, S. M., P. Tamkee, B. Summers, S. Balabahadra, M. Marks, D. M. Kingsley, and D. Schluter. 2012. Genetic Signature of Adaptive Peak Shift in Threespine Stickleback. Evolution. 66:2439–50.

Romero, I. G., A. A. Pai, J. Tung, and Y. Gilad. 2014. Impact of RNA degradation on measurements of gene expression. BMC Biol. 12:42.

Rosenblum, E. B., C. E. Parent, and E. E. Brandt. 2014. The Molecular Basis of Phenotypic Convergence. Ann. Rev. Eco. Evo. Sys. 45:203-26.

Rottscheidt, R., and B. Harr. 2007. Extensive additivity of gene expression differentiates subspecies of the house mouse. Genetics 177:1553–67.

Ruiz-Perez, V. L., and J. A. Goodship. 2009. Ellis-van Creveld syndrome and Weyers acrodental dysostosis are caused by cilia-mediated diminished response to Hedgehog ligands. Am. J. Med. Genet. Part C Semin. Med. Genet. 151:341–51.

Sadkowski, T., M. Jank, L. Zwierzchowski, J. Oprzadek, and T. Motyl. 2009. Comparison of skeletal muscle transcriptional profiles in dairy and beef breeds bulls. J. Appl. Genet. 50:109–23.

Safari-alighiarloo, N., M. Taghizadeh, M. Rezaei-tavirani, B. Goliaei, and A. A. Peyvandi. 2014. Protein-protein interaction networks (PPI) and complex diseases. Gastroenterol. Hepatol. Bed. Bench. 7:17–31.

Schartl, M. 2014. Beyond the zebrafish: Diverse fish species for modeling human disease. DMM Dis. Model. Mech. 7:181–92.

Schaub, M. A., A. P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder. 2012. Linking disease associations with regulatory information in the human genome. Genome Res. 22:1748–59.

Schiffels, S., and R. Durbin. 2014. Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46:919–25.

Schilling, T. F., and C. B. Kimmel. 1994. Segment and cell type lineage restrictions during pharyngeal arch development in the zebrafish embryo. Development 120:483–94.

Schluter, D. 2000. The Ecology of Adaptive Radiation. Oxford University Press, Oxford.

Schluter, D., E. A. Clifford, M. Nemethy, and J. S. Mckinnon. 2004. Parallel Evolution and Inheritance of Quantitative Traits. Am. Nat. 163:809-22.

Schrider, D. R., F. K. Mendes, M. W. Hahn, and A. D. Kern. 2015. Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. Genetics 200:267–84.

Schumer, M., and Y. Brandvain. 2016. Determining epistatic selection in admixed populations. Mol. Ecol. 25:2577-91.

Schumer, M., R. Cui, D. L. Powell, R. Dresner, G. G. Rosenthal, and P. Andolfatto. 2014. High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish species. Elife 2014:1–21.

Seehausen, O. 2006. African cichlid fish: a model system in adaptive radiation research. Proc. R. Soc. B. 273:1987–98.

Sella, G., G. Sella, and N. H. Barton. 2019. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. Annu. Rev. Genom. Hum. G. 20:461-493.

Servedio, M. R., G. S. Van Doorn, M. Kopp, A. M. Frame, and P. Nosil. 2011. Magic traits in speciation: “magic” but not rare? Trends Ecol. Evol. 26:389–97.

Shapiro, M. D., M. A. Bell, and D. M. Kingsley. 2006. Parallel genetic origins of pelvic reduction in vertebrates. Proc. Natl. Acad. Sci. 103:13753–58.

Shen, Y., L. Liang, G. Li, R. W. Murphy, and Y. Zhang. 2012. Parallel Evolution of Auditory Genes for Echolocation in Bats and Toothed Whales. PLoS Genet. 8(6):e1002788.

Signor, S. A., and S. V. Nuzhdin. 2018. The Evolution of Gene Expression in cis and trans. Trends Genet. 34:532-44.

Soria-Carrasco, V., Z. Gompert, A. A. Comeault, T. E. Farkas, T. L. Parchman, J. S. Johnston, C. A. Buerkle, J. L. Feder, J. Bast, T. Schwander, S. P. Egan, B. J. Crespi, and P. Nosil. 2014. Stick insect genomes reveal natural selection’s role in parallel speciation. Science 344:738–42.

St. John, M. E., R. Holzman, and C. H. Martin. 2020. Rapid adaptive evolution of scale-eating kinematics to a novel ecological niche. J. Exp. Biol. jeb.217570.

St. John, M. E., J. A. McGirr, and C. H. Martin. 2019. The behavioral origins of novelty: Did increased aggression lead to scale-eating in pupfishes? Behav. Ecol. 30:557–69.

Stadel, D., V. Millarte, K. D. Tillmann, J. Huber, B. C. Tamin-Yecheskel, M. Akutsu, A. Demishtein, B. Ben-Zeev, Y. Anikster, F. Perez, V. Dötsch, Z. Elazar, V. Rogov, H. Farhan, and C. Behrends. 2015. TECPR2 Cooperates with LC3C to Regulate COPII-Dependent ER Export. Mol. Cell 60:89–104.

Stapley, J., J. Reger, P. G. D. Feulner, C. Smadja, J. Galindo, R. Ekblom, C. Bennison, A. D. Ball, A. P. Beckerman, and J. Slate. 2010. Adaptation genomics: The next generation. Trends Ecol. Evol. 25:705–12.

Stern, D. L., V. Orgogozo. 2008. The loci of evolution: How predictable is genetic evolution? Evolution 62:2155–77.

Strecker, U. 2006. Genetic differentiation and reproductive isolation in a Cyprinodon fish species flock from Laguna Chichancanab, Mexico. Mol. Phylogenet. Evol. 39:865–72.

Svensson, A., R. Libelius, and S. Tågerud. 2008. Semaphorin 6C expression in innervated and denervated skeletal muscle. J. Mol. Histol. 39:5–13.

Sweigart, A. L., L. Fishman, and J. H. Willis. 2006. A simple genetic incompatibility causes hybrid male sterility in mimulus. Genetics 172:2465–79.

Szklarczyk, D., A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. Von Mering. 2015. STRING v10: protein – protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43:447–452.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–95.

Takahasi, K. R., T. Matsuo, and T. Takano-Shimizu-Kouno. 2011. Two types of cis-trans compensation in the evolution of transcriptional regulation. Proc. Natl. Acad. Sci. 108:15276–81.

Taylor SP, Dantas TJ, Duran I, Wu S, Lachman RS, Consortium G, Nelson SF, Cohn DH, Vallee RB, Krakow D. 2015. Mutations in DYNC2LI1 disrupt cilia function and cause short rib polydactyly syndrome. Nat. Commun. 6:1–11.

Thompson, A. C., T. D. Capellini, C. A. Guenther, Y. F. Chan, C. R. Infante, D. B. Menke, and D. M. Kingsley. 2018. A novel enhancer near the Pitx1 gene influences development and evolution of pelvic appendages in vertebrates. Elife 7:1–21.

Tomislav, S. 2011. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLoS One. 6(7):e21800.

True, J. R., and E. S. Haag. 2001. Developmental system drift and flexibility in evolutionary trajectories. Evol. Dev. 3:109–19.

Tulchinsky, A. Y., N. A. Johnson, W. B. Watt, and A. H. Porter. 2014. Hybrid incompatibility arises in a sequence-based bioenergetic model of transcription factor binding. Genetics 198:1155–1166.

Turissini, D. A., J. A. McGirr, S. S. Patel, J. R. David, and D. R. Matute. 2018. The rate of evolution of postmating-prezygotic reproductive isolation in drosophila. Mol. Biol. Evol. 35:312-34.

Turner, B. J., D. D. Duvernell, T. M. Bunt, and M. G. Barton. 2008. Reproductive isolation among endemic pupfishes (Cyprinodon) on San Salvador Island, Bahamas: Microsatellite evidence. Biol. J. Linn. Soc. 95:566–82.

Turner, J. R. G. 1976. Adaptive radiation and convergence in subdivisions of the butterfly genus Heliconius (Lepidoptera: Nymphalidae). Zool. J. Linn. Soc. 58:297–308.

Uebbing, S., A. Künstner, H. Mäkinen, N. Backström, P. Bolivar, R. Burri et al. 2016. Divergence in gene expression within and between two closely related flycatcher species. Mol Ecol. 25:2015-28.

Van De Geijn, B., G. Mcvicker, Y. Gilad, and J. K. Pritchard. 2015. WASP: Allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods 12:1061–63.

Van Dijk, E. L., Y. Jaszczyszyn, and C. Thermes. 2014. Library preparation methods for next-generation sequencing: Tone down the bias. Exp. Cell Res. 322:12–20.

Verta, J. P., and F. C. Jones. 2019. Predominance of cis-regulatory changes in parallel expression divergence of sticklebacks. Elife 8:1–30.

Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang. 2012. Five years of GWAS discovery. Am. J. Hum. Genet. 90:7–24.

Wainwright, P. C., and B. A. Richard. 1995. Predicting patterns of prey use from morphology of fishes. Environ. Biol. Fishes 44:97–113.

Wang, L., J. Nie, H. Sicotte, Y. Li, J. E. Eckel-Passow, S. Dasari, P. T. Vedell, P. Barman, L. Wang, R. Weinshiboum, J. Jen, H. Huang, M. Kohli, and J. P. A. Kocher. 2016. Measure transcript integrity using RNA-seq data. BMC Bioinformatics 17:1–16.

Wang, L., S. Wang, and W. Li. 2012. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 28:2184-85

Webling, K. E. B., J. Runesson, T. Bartfai, and Ü. Langel. 2012. Galanin receptors and ligands. Front. Endocrinol. 3:1–14.

Wei, K. H. C., A. G. Clark, and D. A. Barbash. 2014. Limited gene misregulation is exacerbated by allele-specific upregulation in lethal hybrids between drosophila melanogaster and drosophila simulans. Mol. Biol. Evol. 31:1767–78.

Wellenreuther, M., C. Mérot, E. Berdan, and L. Bernatchez. 2019. Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. Mol. Ecol. 28:1203–09.

West, R. J. D., and A. Kodric-Brown. 2015. Mate Choice by Both Sexes Maintains Reproductive Isolation in a Species Flock of Pupfish (Cyprinodon spp) in the Bahamas. Ethology 121:793–800.

Whitehead, A., and D. L. Crawford. 2006. Variation within and among species in gene expression: Raw material for evolution. Mol. Ecol. 15:1197–1211.

Whiteley, A. R., S. M. Rogers, N. Derome, S. Renaut, J. Jeukens, L. Bernatchez, A. W. Nolte, K. Østbye, J. St-Cyr, G. Lu, and L. Landry. 2010. On the origin of species: insights from the ecological genomics of lake whitefish. Philos. Trans. R. Soc. B Biol. Sci. 365:1783–1800.

Williams, B. B., N. C. Tebbutt, M. Buchert, T. L. Putoczki, K. Doggett, S. Bao, C. N. Johnstone, F. Masson, F. Hollande, A. W. Burgess, A. M. Scott, M. Ernst, and J. K. Heath. 2015. Glycoprotein A33 deficiency: A new mouse model of impaired intestinal epithelial barrier function and inflammatory disease. DMM Dis. Model. Mech. 8:805–815.

Wittkopp, P. J., B. K. Haerum, and A. G. Clark. 2004. Evolutionary changes in cis and trans gene regulation. Nature 430:85-88.

Wittkopp, P. J., and G. Kalay. 2011. Cis -regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. 13:59–69.

Wolf, J. B. W., J. Lindell, and N. Backstrom. 2010. Speciation genetics: current status and evolving approaches. Philos. Trans. R. Soc. B Biol. Sci. 365:1717–33.

Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. 20:1377-419

Wright, S. 1988. Surfaces of Selective Value Revisited. Am. Nat. 131:115–123.

Xia, J. H., G. Lin, X. He, P. Liu, F. Liu, F. Sun, R. Tu, and G. H. Yue. 2013. Whole genome scanning and association mapping identified a significant association between growth and a SNP in the IFABP-a gene of the Asian seabass. BMC Genomics 14:295.

Ye, J., G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, and T. L. Madden. 2012. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 13:134.

Yeaman, S., and M. C. Whitlock. 2011. The genetic architecture of adaptation under migration-selection balance. Evolution 65:1897–1911.

Zhao, K., C.-W. Tung, G. C. Eizenga, M. H. Wright, M. L. Ali, A. H. Price, G. J. Norton, M. R. Islam, A. Reynolds, J. Mezey, A. M. McClung, C. D. Bustamante, and S. R. McCouch. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat. Commun. 2:467.

Zhao, L., J. Wit, N. Svetec, and D. J. Begun. 2015. Parallel Gene Expression Differences between Low and High Latitude Populations of Drosophila melanogaster and D. simulans. PLoS Genet. 11(5):e1005184.

Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44:821–24.

Zhou, X., P. Carbonetto, and M. Stephens. 2013. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 9:e1003264.