UNIVERSITY OF CALIFORNIA
SANTA CRUZ

EXTREME ENVIRONMENTS AND SHARED SOLUTIONS: CONVERGENT ADAPTATION AND GENE FLOW IN SULFIDE SPRING FISHES

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ECOLOGY AND EVOLUTIONARY BIOLOGY

by

Kara Ryan

December 2025

The Dissertation of Kara Ryan is approved:

_____
Professor Joanna L. Kelley, Chair

_____
Professor Eric P. Palkovacs

_____
Professor Michael Tobler

_____
Professor Devon E. Pearse

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

EXTREME ENVIRONMENTS AND SHARED SOLUTIONS: CONVERGENT ADAPTATION AND GENE FLOW IN SULFIDE SPRING FISHES

by

KARA RYAN

A central question in evolutionary biology is whether evolution is predictable – driven by similar selective pressures leading to repeated outcomes – or contingent on historical and stochastic processes. While convergent traits are widespread across the tree of life, the underlying genetic mechanisms often vary, raising questions about how different sources of variation can influence evolutionary trajectories. In this dissertation, I use the *Poecilia mexicana* species complex, which includes multiple sulfide-adapted and non-sulfide-adapted populations across four river drainages, as a model for repeated adaptation to extreme environments. In Chapter One, I used targeted capture sequencing to test for selection in 250 candidate genes, identifying shared signatures of selection in key sulfidic detoxification genes across independent sulfidic lineages. In Chapter Two, I generated new genome assemblies for six *Poecilia* and *Gambusia* species, assessed synteny and demographic history, and tested hypotheses about population size changes and the genomic architecture in sulfidic lineages. Chapter Three investigated the role of structural variants in adaptation by integrating long- and short-read DNA sequencing with transcriptomic data. This work identified many structural variants under selection, including a deletion in *ethe1* associated with gene expression differences that was fixed in sulfidic individuals and at low frequency in nonsulfidic individuals. Finally, Chapter Four, using whole-

genome resequencing data from 172 individuals, revealed that gene flow has facilitated repeated adaptation to sulfide springs by shaping allele frequencies and patterns of divergence for both SNPs and structural variants. Collectively, this dissertation shows that while a defined set of genes and pathways involved in sulfide detoxification recurrently underpin adaptation, this is the result of the interplay of *de novo* mutation, gene flow and selection. This work highlights how predictable and contingent processes jointly shape evolutionary outcomes, proving new insight into convergent adaptation in extreme environments.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Joanna Kelley, for the unwavering support and the many opportunities she has provided for me to grow as a scientist. You have taught me to think critically, be curious, and to not be fearful of complexity. Thank you for believing in my potential and for leading by example. I feel lucky to have you as a mentor.

I am grateful to my committee members, who have shaped the way I think about science and the natural world. Thank you to Michi Tobler for sharing your expertise and philosophy on science and life -- your investment in my work has strengthen my research shaped how I think about evolution. Thank you to Wes Dowd and Michael Phelps for helping to shape my research questions and for helping me to grow beyond my expertise. I also thank Eric Palkovacs and Devon Pearse for welcoming me into the UCSC community, your genuine curiosity about my work, and for pushing me to think in more integrative ways.

A huge thank you goes to the members of the Kelley Lab for their friendship, adventures, feedback, code, dog sitting and being the best hype team. I thank Blair Perry for your mentorship and guidance – you have profoundly shaped the scientist I am today. I truly would not be here today without your help. I am also grateful to Kerry McGowan for your generosity with your time and knowledge when I first joined the lab. Thank you to the undergraduate researchers who have assisted in this work including Jake Landers, Jeremy Davis, and Ryan Alex.

To Alan -- none of this would be possible without your unwavering support. You have helped me through my hardest moments, been my loudest cheerleader, and have made life so full of joy. You inspire me every day, and I cannot thank you enough. And to Wilson for the endless laughs. Thank you to Mary Lefebre for the countless packages, cards, and coffee beans. They always seemed to show up at my door just when I needed them.

Thank you to my family and friends for always believing in me. To my friends, for the countless hikes, road trips, coffees, voice notes, game nights, dinners, and days in the garden. I feel lucky to have friends who love nature as much as I do.

I thank the communities of Teapa and Tapijulapa, including local landowners and Villa Luz Nature Park providing access to the studies sites. I thank the Centro de Investigación E Innovación Para la Enseñanza y El Aprendizaje (CIIEA) and Universidad Juárez Autónoma de Tabasco for their hospitality and support.

Finally, I am indebted to the extraordinary sulfide fishes. I am grateful that you made toxic streams your home, and to those individuals sacrificed for this research – I thank you.

STATEMENT OF CONTRIBUTION

This dissertation includes the reprint of the following previously published material:

Chapter 1: Ryan, K., Greenway, R., Landers, J., Arias-Rodriguez, L., Tobler, M., & Kelley, J. L. (2023). Selection on standing genetic variation mediates convergent evolution in extremophile fish. Molecular Ecology, 32, 5042–5054. https://doi.org/10.1111/mec.17081

Chapter 2: Ryan, K., De-Kayne, R., Davis, J., Arias-Rodriguez, L., Tobler, M., & Kelley, J. L. (2025). New Genome Assemblies for Poeciliidae: A Foundation for Adaptation Studies. *Genome biology and evolution*, *17*(6), evaf111. https://doi.org/10.1093/gbe/evaf111

INTRODUCTION

Evolutionary biologists have long wondered if evolution is predictable or if it is instead shaped by chance (Darwin, 1859; Jonathan B. Losos, 2011, 2017; Rosenblum et al., 2014). This debate centers on whether repeated evolution is inevitable – driven by similar selective pressures leading to predictable, convergent outcomes – or whether it is contingent on historical and stochastic processes. Both theoretical work (Gould, 1990; Jonathan B. Losos, 2017), and empirical studies of convergence ranging from *E. coli* (Blount et al., 2008; Lenski et al., 1991), to anoles (J. B. Losos et al., 1998; Mahler et al., 2013), to stickleback (Colosimo et al., 2005; Jones et al., 2012) and cavefish (Borowsky, 2018), have provided evidence for both sides of this debate, underscoring that evolutionary outcomes are often both repeatable and idiosyncratic.

Despite decades of work addressing this question, an important gap remains: much of this debate has treated repeated adaptation as the product of either inevitability (i.e. selection finding the same solutions) or contingency (unique histories and accidents) (Blount et al., 2018) while paying less attention to how the sources of adaptive variation may influence evolutionary outcomes. Convergent phenotypes are widespread across the tree of life. Importantly, convergence can occur across multiple levels of the biological hierarchy – from protein function to physiological pathways to whole organism traits. In many cases, similar phenotypes arise through distinct molecular mechanisms, suggesting that while natural selection favors similar trait solutions, the genomic basis of adaptation may often differ across

lineages (Stern, 2013). This variation between repeatable phenotypic convergence and heterogeneous genotypic underpinnings complicates this debate about the predictability of evolution.

Convergent phenotypes can be underpinned by variation in the same or different genes. When the same gene is implicated, the source of this adaptive variation can be classic independent *de novo* mutation, but it could also exist as standing variation, either as ancestral variation that is maintained beyond speciation events or introduced via gene flow (Stern, 2013). Gene flow provides a pathway to convergence that blurs the dichotomy of inevitability or contingency (Slatkin, 1987). Gene flow can make outcomes appear more predictable – because multiple populations find similar adaptive solutions – while also retaining a strong element of contingency, since the opportunity for alleles to move depends on the historical and demographic context of population connectivity.

Extreme environments impose intense, well-defined selective pressures that few lineages can tolerate (Maček et al., 2016; Rappaport & Oliverio, 2024). These environments act as long-term natural experiments that test the limits of adaptation and thus provide powerful opportunities for investigating whether evolutionary responses are constrained and repeatable or idiosyncratic and contingent. An example of such a system is the *Poecilia mexicana* species complex in the Grijalva River Basin in southern Mexico. This system contains multiple lineages that have independently adapted to hydrogen sulfide levels orders of magnitude higher than what is typically toxic to animals. The presence of closely related nonsulfidic

populations within the same river drainages make this system particularly valuable for investigating convergent evolution (Tobler et al., 2018).

In Chapter One (Ryan et al., 2023), I used a population level targeted capture approach to test for selection in 250 nuclear-encoded genes previously implicated with sulfide processes or Oxidative Phosphorylation. While many signals of selection were population-specific, two key sulfide detoxification genes showed evidence of selection across in all five populations. This work indicated that this shared variation under selection had a monophyletic origin, a hypothesis that was tested in Chapter Four.

In Chapter Two (Ryan et al., 2025), I generated new genome assemblies and applied comparative genomic approaches to investigate synteny and demographic history across *Poecilia* and *Gambusia,* two genera that have multiple populations that have independently adapted to sulfide springs. These analyses revealed that not all sulfidic populations have reduced population sizes compared to their nonsulfidic counterpart and no large-scale difference in synteny were detected between ecotypes. Additionally, these new genomic resources were used in Chapter Three (Ryan et al., *in review*) to test the hypothesis of the role of structural variants in adaptation.

Chapter Three (Ryan et al., *in review*) then took a population level genome-wide approach, integrating long- and short-read DNA sequencing as well as RNA sequencing to explore the role of structural variants in repeated adaptation. In addition to identifying structural variants under selection that sort by ecotype, this work found associations between structural variants in noncoding regions and

differential gene expression. For example, I identified a deletion within the intron of one copy of sulfide detoxification gene *ethe1* that is fixed in all sulfidic populations, occurs at low frequencies in nonsulfidic populations, and is associated with expression differences between gene copies and ecotypes.

Finally, Chapter Four (Ryan et al., *in prep*) I explored both genome-wide signatures of selection and the sources of adaptive variation in these regions in 172 genomes. Across multiple approaches, we found strong evidence that gene flow has shaped the evolutionary history of this species complex and facilitated the repeated adaptation to sulfide springs. Gene flow explains patterns of divergence and allele frequencies for both Single Nucleotide Polymorphism (SNPs) and structural variants across the genome. Furthermore, these results suggest that although many regions are under selection, and much of these regions are not shared among sulfide adapted populations, variation in sulfide detoxification genes is consistently under selection in all five populations.

Together, this dissertation demonstrates that while the same genes and pathways – particularly those involved in sulfide detoxification – repeatedly underpin adaptation supporting the predictability of evolution, these outcomes are contingent on the stochastic nature of gene flow and geographic isolation.

Chapter 1 SELECTION ON STANDING GENETIC VARIATION MEDIATES CONVERGENT EVOLUTION IN EXTREMOPHILE FISH

Kara Ryan[1], Ryan Greenway[2,3], Jake Landers[1], Lenin Arias-Rodriguez[4], Michael Tobler[3] Joanna L. Kelley[1,5]

[1] School of Biological Sciences, Washington State University, 100 Dairy Road, Pullman, WA 99164, USA
[2] Zoology and Evolutionary Biology, Department of Biology, University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Germany
[3] Division of Biology, Kansas State University, 116 Ackert Hall, Manhattan, KS 66506, USA
[4] División Académica de Ciencias Biológicas, Universidad Juárez Autónoma de Tabasco (UJAT), C.P. 86150, Villahermosa, Tabasco, México
[5] Department of Ecology and Evolutionary Biology, University of California Santa Cruz, 130 McAllister Way, Santa Cruz, CA 95060, USA

Abstract

Hydrogen sulfide is a toxic gas that disrupts numerous biological processes, including energy production in the mitochondria, yet fish in the *Poecilia mexicana* species complex have independently evolved sulfide tolerance several times. Despite clear evidence for convergence at the phenotypic level in these fishes, it is unclear if the repeated evolution of hydrogen sulfide tolerance is the result of similar genomic changes. To address this gap, we used a targeted capture approach to sequence genes associated with sulfide processes and toxicity from five sulfidic and five nonsulfidic populations in the species complex. By comparing sequence variation in candidate genes to a reference set, we identified similar population structure and differentiation, suggesting that patterns of variation in most genes associated with sulfide processes and toxicity are due to demographic history and not selection. But the presence of tree discordance for a subset of genes suggests that several loci are

evolving divergently between ecotypes. We identified two differentiation outlier genes that are associated with sulfide detoxification in the mitochondria that have signatures of selection in all five sulfidic populations. Further investigation into these regions identified long, shared haplotypes among sulfidic populations. Together, these results reveal that selection on standing genetic variation in putatively adaptive genes may be driving phenotypic convergence in this species complex.

Introduction

Convergent evolution, the independent evolution of similar traits in multiple lineages (Jonathan B. Losos, 2011), can occur at various levels of biological organization, from phenotype to nucleotide position. However, convergence at one level does not necessitate convergence in the underlying mechanism (Rosenblum et al., 2014). In some cases, convergent phenotypes are largely driven by the same amino acid change (e.g. echolocation (Y. Liu et al., 2010; Rossiter et al., 2011), but in others, there is little convergence at the amino acid or gene level (e.g. haemoglobin, (Natarajan et al., 2016). Moreover, the degree to which convergent phenotypes are the result of selection on standing variation or on de novo mutation remains unclear. In some cases, the source of adaptive variation is largely de novo mutation (e.g. pigmentation in deer mice (Linnen et al., 2009), but in others, the majority source is standing variation (Alves et al., 2019; Haenel et al., 2019; Jones et al., 2012; Lai et al., 2019; N. M. Reid et al., 2016). Determining how often convergent phenotypes are driven by similar genomic changes will provide a better understanding of the

repeatability of evolution (Gould, 1990; Rosenblum et al., 2014) and the relative contributions of selection on standing variation and de novo mutation in adaptation.

Extreme environments provide valuable systems to explore how selection may drive convergence across multiple biological levels. The strong selective regimes found in extreme environments often promote the evolution of convergent phenotypes (Riesch et al., 2015; Xu et al., 2020). These systems are particularly valuable when there are naturally replicated environments that have been independently colonized by multiple populations, allowing the investigation of how selection has independently shaped variation in the genome (Tobler et al., 2018). Freshwater springs in the Río Grijalva basin in Southern Mexico are an example of such a system. Springs in this area contain naturally occurring hydrogen sulfide (H_2S) at concentrations orders of magnitude higher than those considered lethal for most animals (Tobler et al., 2018, 2016). Despite the highly toxic conditions, sulfide springs in four river drainages have been independently colonized by fish populations from the *Poecilia mexicana* species complex, including *P. mexicana*, *P. thermalis*, and *P. sulphuraria* (Palacios et al., 2013; Tobler et al., 2018). Additionally, closely related populations inhabit nonsulfidic streams within the same drainages (Figure 1.1), providing a unique comparative framework to explore evolutionary processes that potentially give rise to convergent adaptations (Tobler et al., 2018). The sulfide spring populations from the Puyacatengo and Tacotalpa drainages show evidence for recent divergence between ecotypes, estimated at ~10,000 years ago in Tacotalpa and ~300 years ago in Puyacatengo (A. P. Brown et al., 2018). In contrast, sulfide spring populations in the

Pichucalco and Ixtapangajoya drainages are hypothesized to have diverged earlier, with estimates from the Pichucalco drainage suggesting a timing of divergence of ~19,000 years ago (Greenway et al., 2021). Sulfide spring populations are locally adapted and exhibit convergent phenotypes related to morphology (Greenway et al., 2019; Riesch et al., 2016; Tobler & Hastings, 2011), life history traits (Riesch et al., 2010, 2014) and physiology (Greenway et al., 2020; Pfenninger et al., 2014; Plath et al., 2013a; Tobler et al., 2011, 2016). These naturally replicated systems, therefore, provide a unique opportunity to explore the genomic basis of convergent adaptive phenotypes, from nucleotide position to pathway.

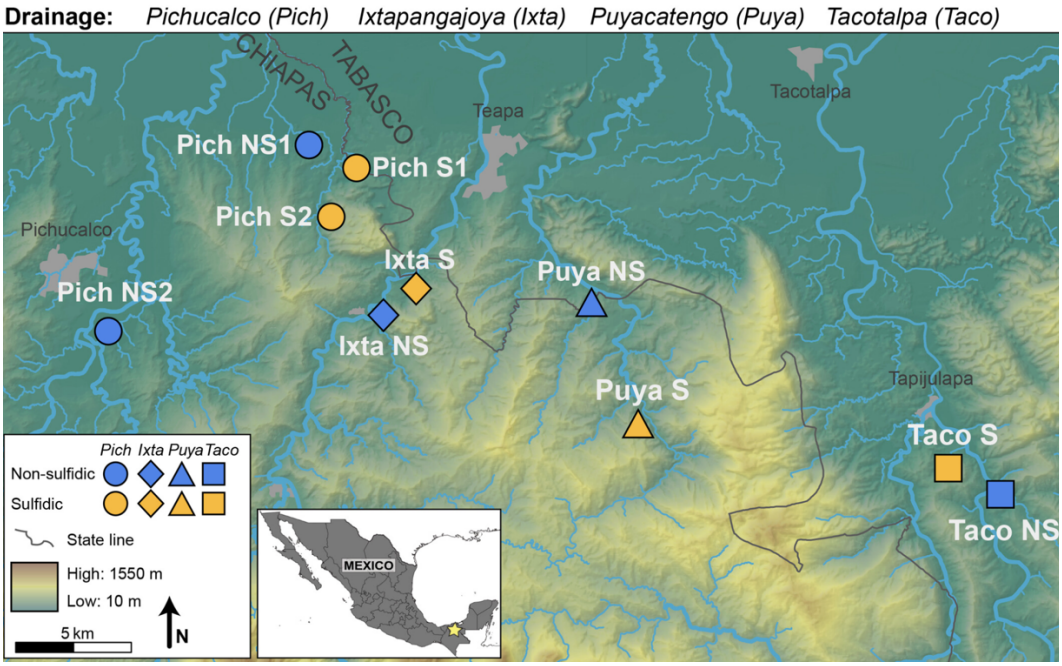


Figure 1.1: Map of Study region and sites. Samples were collected from 10 sites in the Río Grijalva basin. The study site location is indicated by a yellow star in the insert map of Mexico. Shape represents drainage and color represents ecotype. This figure was adapted from Hotaling et al. (2019).

The main mechanism of toxicity imposed by H_2S involves the inhibition of aerobic respiration. In the mitochondria, H_2S binds to Complex IV (cytochrome c oxidase, COX), inhibiting oxidative phosphorylation (OxPhos) and halting aerobic energy production even at micromolar concentrations (Cooper & Brown, 2008; Hill et al., 1984). In addition to disrupting energy production, elevated H_2S concentrations can produce harmful effects by modulating ion channels (García-Bereguiaín et al., 2008), modifying oxygen transport proteins (Pietri et al., 2011), interacting with transcription factors (Budde & Roth, 2010) and signaling molecules (Calvert et al., 2009), and disrupting posttranslational modification of proteins (Mustafa et al., 2009). Despite the presence of a highly conserved detoxification pathway across eukaryotes, the sulfide:quinone oxidoreductase (SQR) pathway in the mitochondria (Hildebrandt & Grieshaber, 2008; Libiad et al., 2014), environmental exposure to H_2S is still potentially lethal for most animals (Lagoutte et al., 2010). Although the biochemical action of H_2S is well understood, the ways in which these proteins may be modified for adaptation to H_2S-rich environments are largely unknown. At a molecular level, the strong selective pressure imposed by constant exposure to H_2S is predicted to drive adaptive modification of genes associated with OxPhos and H_2S detoxification. These genes are prime targets for natural selection due to their importance in cell survival, susceptibility to H_2S, and their highly conserved nature across taxa, which provides an opportunity to test for convergence at a molecular level. However, because of the genomic complexity and redundancy of these processes—more than 200 genes are associated with H_2S tolerance, H_2S detoxification, or OxPhos—it is

unclear to what extent the convergent evolution of sulfide tolerance is a result of convergence at the genomic level.

There is evidence for genomic convergence at the gene and nucleotide levels in subsets of populations of *P. mexicana*, that seems to arise from a combination of selection on standing variation and de novo mutation. For example, genes associated with the sulfide:quinone oxidoreductase pathway in the mitochondria, including *sulfide:quinone oxidoreductase* (*sqor*) and persulfide dioxygenase (*ethe1*) show evidence for selection on standing variation (A. P. Brown et al., 2018; Greenway et al., 2020; Pfenninger et al., 2015; Tobler et al., 2018). Additionally, these sulfur detoxification genes are differentially expressed between ecotypes (A. P. Brown et al., 2018; Kelley et al., 2016; Passow, Brown, et al., 2017; Passow, Henpita, et al., 2017; Tobler et al., 2014). Evidence for shared adaptive changes in subunits of OxPhos has been identified (A. P. Brown et al., 2018; Greenway et al., 2020; Kelley et al., 2016; Pfenninger et al., 2014), including evidence for selection acting on de novo mutations in mitochondrially encoded subunits of OxPhos complexes (Greenway et al., 2020; Pfenninger et al., 2014).

While previous studies support the importance of regions associated with H_2S detoxification and OxPhos, it remains untested whether selection has acted on shared regions of the genome across all sulfidic populations of this species complex. Furthermore, it is unclear if convergence is the result of selection on standing variation or de novo mutation. To address these questions, we used a targeted exon capture approach to sequence candidate genes associated with H_2S toxicity and

detoxification from five sulfidic and five nonsulfidic populations of *P. mexicana*. Using these data, we tested (1) whether the relationship among populations at H_2S candidate genes differs from the relationship at background genes, and (2) whether a subset of genes associated with H_2S detoxification has been targeted by selection in all drainages. This study highlights the importance of selection on standing genetic variation in the repeated evolution of complex traits in extreme environments.

Materials & Methods

Samples and sequencing

A total of 200 individuals were sampled from five sulfidic and five nonsulfidic populations (20 individuals per population) from the Pichucalco (Pich 1 and Pich 2), Ixtapangajoya (Ixta), Puyacatengo (Puya), and Tacotalpa (Taco) drainages in the Río Grijalva basin, Mexico (Figure 1.1; Table S1.1). Sampling included populations of *P. sulphuraria* (Pich sulfidic), *P. thermalis* (Ixta sulfidic), and *P. mexicana mexicana* (all other populations).

Probes were designed for capture sequencing by Rapid Genomics to target a total of 415 nuclear-encoded genes, comprised of 250 candidate genes and 165 background genes (Table S1.2). The 166 candidate genes associated with sulfide detoxification and sulfur processing were identified using Gene Ontology (GO) terms (Table S1.3). The candidate set also included 84 nuclear-encoded OxPhos genes, identified using a BLASTn search of the *P. mexicana* reference genome for genes encoding subunits of OxPhos from (F. Zhang & Broughton, 2013). The background

set contained 73 housekeeping genes from Zhang & Broughton (2013) that are highly expressed in all cell types and involved in critical functions, providing an appropriate comparator to OxPhos genes (Amsterdam et al., 2004; Warrington et al., 2000), and 92 additional nuclear-encoded genes involved in mitochondrial functions (excluding OxPhos and sulfide-related genes identified above) selected from the MitoCarta2.0 database (Calvo et al., 2016; Pagliarini et al., 2008) at random.

DNA was extracted from muscle tissue preserved in RNAlater (Ambion, Inc.) using the Gentra Puregene Tissue Kit following the manufacturer's protocol for purifying DNA from 5 to 10 mg of tissue with the following modifications: (i) tissues were homogenized using a micro pestle, (ii) centrifugation was carried out for 3.5 mins following the addition of protein precipitations solution, and (iii) centrifugation was performed for 2 mins following the addition of isopropanol. DNA was quantified using a Qubit fluorometer and was visualized on a 1% agarose gel.

Library preparation was performed by Rapid Genomics utilizing the Illumina high-throughput workflow and proprietary chemistry. Briefly, DNA was sheared to a mean fragment length of 400 base pairs (bp). The resulting fragments were end-repaired, followed by the incorporation of Illumina unique dual-indexed adapters and PCR enrichment. Probes from Rapid Genomics set RG_3101 were hybridized to the libraries and enriched for the targets of interest. Sequencing was performed on an Illumina HiSeq system with paired-end 150 bp reads. The resulting raw data were demultiplexed using Illumina's BCLtoFastq.

Quality control, variant calling, and filtering

Reads were quality checked using FastQC v0.11.9 (Andrews, 2010), and outputs were summarized using MultiQC v1.11 (Ewels et al., 2016). Reads were trimmed using the Cutadapt (M. Martin, 2011) wrapper TrimGalore v0.6.6 (https://github.com/FelixKrueger/TrimGalore) with parameters --stringency 5, --length 40, and default parameters trimming adaptors and reads with a quality score less than 20. Trimmed reads were aligned to the *P. mexicana* genome (NCBI accession: GCA_001443325.1 (Warren et al., 2018) using BWA-MEM v0.7.17 (H. Li, 2013). The resulting SAM files were converted to BAM files using Samtools v1.8 (H. Li et al., 2009). The BAM files were then sorted, and duplicates were marked using Picard Tools v2.21.4 (Broadinstitute/Picard, 2014/2022) *SortSam* and *MarkDuplicates*. Variants were called according to the Genome Analysis Tool Kit (GATK) v4.2.5.0 (Van der Auwera & O'Connor, 2020) best practices for data pre-processing for variant discovery and germline short variant discovery (single nucleotide polymorphisms [SNPs] and insertions/deletions [Indels]). GATK's *HaplotypeCaller* was used in gvcf mode to generate intermediate per-sample Genomic Variant Call Format (GVCF) files, which were then consolidated using *GenomicsDBImport*. Samples were then joint genotyped using GATK *GenotypeGVCFs*, retaining invariant sites using the -allSites parameter and combined using *CombineVariants*.

The resulting VCF files were filtered for sequencing depth and missingness (--minDP 20, --max_missing 0.9) using VCFtools v0.1.16 (P. Danecek et al., 2011).

Additionally, VCFtools was used to remove a nonsulfidic individual from the Pichucalco drainage that was identified as a first-generation hybrid. To generate an all-sites VCF with proper variant filtering, we separated the files into variant (--mac 1) and invariant (--maf 0) sites using VCFtools. We filtered the variant file for quality and minor allele frequency (--minQ 30 --maf 0.01) using VCFtools. Additionally, we removed loci that were significantly out of Hardy Weinberg Equilibrium within each population ($p < .001$) using dDocent Perl script *filter_hwe_by_pop.pl* (Puritz, Hollenbeck, et al., 2014; Puritz, Matz, et al., 2014). The resulting filtered variant sites VCF were then concatenated to the invariant sites using BCFtools v1.10.2 *concat* (Petr Danecek et al., 2021) and intersected with the original bed file containing the targeted regions for probe design to remove off-target variants and split the data to target and background using BEDTools v2.27.1 (Quinlan & Hall, 2010). The resulting filtered all-sites VCF file was converted to phylip format for phylogenetic analysis using python script vcf2phylip.py (https://github.com/edgardomortiz/vcf2phylip). In addition to an all-sites VCF, we filtered the candidate and background VCF files to retain only biallelic SNPs using VCFtools and remove SNPs found in high LD (+prune -l 0.8 -w 1000) using BCFtools. Unless otherwise stated, analyses were performed using the VCF containing biallelic, LD-pruned SNPs.

Analysis of evolutionary relationships among populations

To investigate population structure, a principal component analysis (PCA) was performed using Plink2 (Chang et al., 2015). Iqtree2 v2.1.3 (Minh et al., 2020) Ultra-Fast Bootstrap (Minh et al., 2013) (-B 1000 -bnni) approach was used to generate a maximum likelihood tree for the background and candidate gene sets using the unpruned, all-sites phylip file. ADMIXTURE v1.3.0 (Alexander et al., 2009) was used to investigate individual ancestry. Pong v1.5 (Behr et al., 2016) was used to visualize ADMIXTURE clustering.

Analysis of population genetic differentiation

Estimates of F_{ST}, heterozygosity, and nucleotide diversity (π) for each gene set were calculated using Stacks v2.59 *populations* (Catchen et al., 2013). Fixed differences between each comparison were identified by first filtering for private alleles and then filtering for a maximum minor allele frequency of 0 using VCFtools. Pixy v1.2.5 (Korunes & Samuk, 2021) was used to summarize nucleotide diversity and estimate genetic differentiation at the gene level between sulfidic and nonsulfidic populations from the same drainage using the all-sites VCF. First, Nei and Li′s nucleotide diversity (π) (Nei M & Li W H, 1979) was estimated for all populations, and the difference in nucleotide diversity between pairs ($\Delta\pi_{ecotype}$) was calculated by subtracting π of sulfidic from nonsulfidic populations ($\pi_{NS} - \pi_{S}$), such that loci with decreased nucleotide diversity in sulfidic populations would result in $+\Delta\pi_{ecotype}$. Next, we estimated relative population differentiation between pairs using Weir and Cockerham's estimate of F_{ST} (Weir & Cockerham, 1984). Because this estimator can

result in a negative value for populations that contain more variation within, we replaced all negative F_{ST} values with 0. Estimates of F_{ST} can be influenced by differences in nucleotide diversity within populations (Cruickshank & Hahn, 2014), therefore we also calculated d_{xy} (Nei M & Li W H, 1979) as an absolute measure of population differentiation.

We first identified outlier loci putatively under selection using a multivariate approach in Minotaur v0.0.1 (Verity et al., 2017). We used the distributions of F_{ST}, d_{xy}, and $\Delta\pi_{ecotype}$ per gene to calculate the Mahalanobis distance (Mahalanobis, 1936), and loci with greater-than-expected differentiation based on a 95% confidence intervals were considered putatively under selection. We then compared these drainage-specific gene lists to identify shared outlier genes. In addition to the per-gene approach, we identified outlier SNPs. We used VCFtools to calculate Weir and Cockerham's F_{ST} between all sulfidic and nonsulfidic populations on a per-site basis using a VCF containing biallelic SNPs that were not LD pruned. Outlier SNPs were identified using a 99.5% cut-off. SNPs were annotated using SNPeff v5.0e (Cingolani et al., 2012), and the potential effects of high F_{ST} nonsynonymous mutations were assessed using PolyPhen2 (Adzhubei et al., 2013).

Haplotype network

To generate haplotype networks, we split the non-LD pruned VCF containing biallelic SNPs by population using VCFtools and phased the sites using the Popgen Pipeline Platform script vcf_phase.py using the Beagle v5.1 algorithm (B. L.

Browning et al., 2018; S. R. Browning & Browning, 2007). The resulting VCFs were split into individual VCFs using BCFtools query and view. For each individual, a fasta file containing each haplotype was generated using BCFtools consensus (using the -H 1 and -H 2 parameter for haplotype 1 and haplotype 2, respectively). The header of the resulting fasta files was fixed to match the appropriate sample and haplotype before concatenating. Each gene-specific fasta file was then aligned using mafft v7.429 (Katoh & Standley, 2013) using default parameters. Haplotype networks were then generated for each gene using R v4.1.2 packages pegas v1.1 (Paradis, 2010) and ape v5.6.2 (Paradis & Schliep, 2019) using a statistical parsimony network (TCS) approach (Clement et al., 2000; Templeton et al., 1992) with singleton haplotypes filtered prior to network generation. The resulting network was visualized using Cytoscape (Shannon et al., 2003) v3.9.1.

Results

Targeted capture sequencing

Targeted capture sequencing resulted in 0.2–1.9 million 150-bp paired-end reads per sample, with an average of 0.7 million reads and an average Phred score greater than 35. GC content in reads ranged from 43% to 47%. Of the 839,328 bp targeted, 784,629 were retained after filtering, of which 9446 (1.2%) were SNPs. LD filtering resulted in a final set of 7277 SNPs. Genes associated with H_2S had more SNPs in high LD (26.1%) when compared to OxPhos (19.1%) and background (19.2%) genes (Table 1.1).

Table 1.1: The number of genes, exons, and sites captured by targeted sequencing and number of filtered sites, single nucleotide polymorphisms (SNPs), and linkage-disequilibrium pruned SNPs used in this analysis.

Set	Genes	Exons	Targeted sites	Filtered sites	SNPs	Pruned SNPs
Target: Sulfide	166	1330	421,765	398,664	5153	3808
Target: OxPhos	84	432	96,472	88,421	973	787
Background	165	1341	321,091	297,544	3320	2682
Total	415	3103	839,328	784,629	9446	7277

Similar population structure inferred from background and candidate genes

To investigate the relationship among populations within each gene set, we used admixture to cluster individuals based on ancestry. The background set supported a best *K* of 7, while both the OxPhos and sulfide sets supported a best *K* of 8. Nonetheless, the relationship among populations was similar (Figure S1.1; Figure 1.2a). Variation in the best *K* was the result of nonsulfidic individuals from Pich clustering as a single population in the background set and as two populations in the OxPhos and sulfide sets (Figure 1.2a). Generally, sulfide spring populations were recovered as distinct clusters, but Puya sulfidic and nonsulfidic individuals clustered as a single population in all gene sets (Figure 1.2a).

PCA of LD-filtered SNPs separated *P. sulphuraria* (Pich 1 and Pich 2, sulfidic) and *P. thermalis* (Ixta, sulfidic) individuals from *P. mexicana* individuals (all others) along PC axis 1, which explained 30.3%–34.9% of variation depending on the gene set (Figure 1.2b). PC axis 2 separated *P. mexicana* individuals by ecotype and explained 11.8%–13.2% of variation. Interestingly, the relationship among individuals from Puya and sulfidic individuals from Taco varied across gene sets

(Figure 1.2b). For example, Puya individuals clustered as a distinct group from Taco sulfidic individuals in both the background and the sulfide set (Figure 1.2b). In contrast, sulfidic individuals from Taco and Puya clustered as a single group in the OxPhos set (Figure 1.2b). PC axis 3 and 4 showed similar clustering patterns between the sulfide and Oxphos sets, but interestingly, the background set clustered sulfidic individuals from Taco with nonsulfidic individuals from Pich (Figure S1.2).

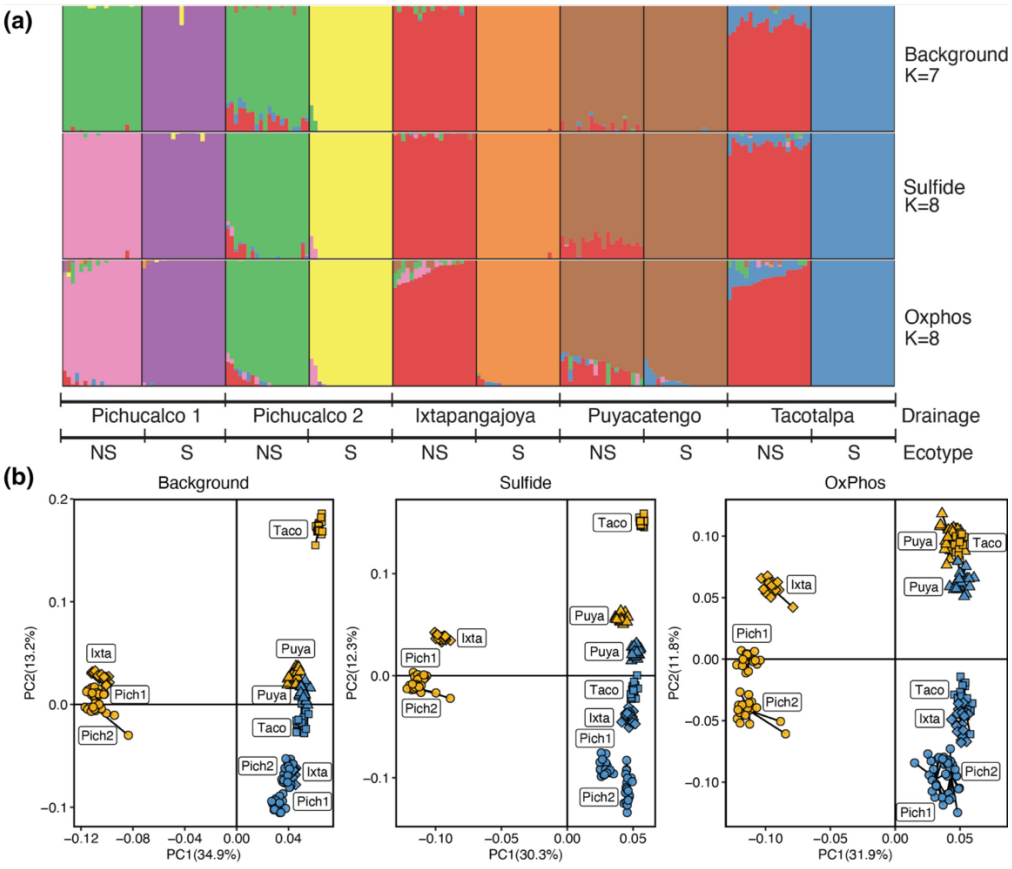


Figure 1.2:Analysis of population structure in background, sulfide processing, and OxPhos genes. (a) Best *K* from admixture analyses of each gene set ordered by drainage, from west to east. (b) Principal component analysis of unlinked single nucleotide polymorphisms for each gene set. Color represents ecotype (sulfidic in yellow, nonsulfidic in blue) and shape represents drainage of origin.

In addition to clustering analyses, we estimated relative differentiation using Weir and Cockerham F_{ST} for each pairwise comparison. In all gene sets, we found the highest F_{ST} between the Taco sulfidic population and the sulfidic populations from Pich and Ixta and the lowest F_{ST} between sulfidic and nonsulfidic populations from Puya (Figure S1.3). Additionally, F_{ST} tended to be lower between comparisons of nonsulfidic populations than between sulfidic populations or between populations of differing ecotypes across all gene sets (Figure S1.4). Observed and expected heterozygosity (H_O/H_E) was reduced in all sulfidic populations when compared to the adjacent nonsulfidic population, except for Puya, but this pattern was consistent across gene sets (Table S1.4). Similarly, nucleotide diversity (π) was lower in sulfidic populations compared to nonsulfidic populations (Table S1.5). We also compared distributions of per gene F_{ST}, d_{xy}, and $\Delta\pi$ between ecotypes from the same drainage. Within a drainage, the distributions of variation in both sulfur processing and OxPhos genes were similar to the background (Figure S1.5). But between drainages, these distributions varied greatly (Figure S1.5). Due to the overall similarity with the background, our results suggest that most of the variation in genes associated with OxPhos and sulfide processing is a result of demographic processes and not selection acting on many potentially adaptive loci.

Tree discordance between background and potential target genes

Despite limited evidence of selection acting on many loci across our targeted genes, phylogenetic analysis revealed tree discordance between background, sulfide

processing, and OxPhos genes. The topology of the highest supported maximum likelihood tree of the background set supported previous studies (A. P. Brown et al., 2019; Greenway et al., 2020; Palacios et al., 2013) that suggest three independent colonizations of sulfidic springs—a more ancient colonization by the *P. sulphuraria* clade as well as two more recent colonizations by *P. mexicana* in the Taco and Puya drainages (Figure 1.3). In contrast, the sulfide set clustered all populations by ecotype, contradicting the expected population tree (Figure 1.3). The OxPhos set was similar to the background set, except that the Taco sulfidic individuals cluster with both Puya populations instead of as sister taxa with the nonsulfidic population from the same drainage (Figure 1.3).

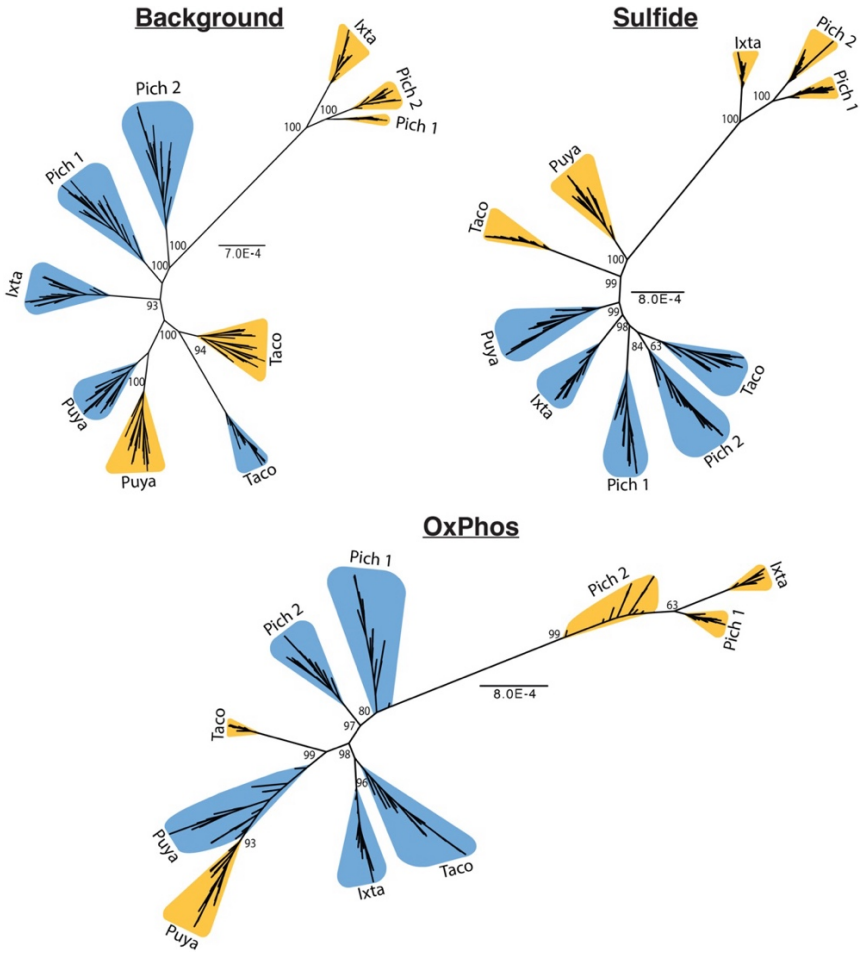


Figure 1.3: Tree discordance between maximum likelihood tree of background, sulfide processing, and OxPhos genes. Colour represents ecotype (sulfidic in yellow, nonsulfidic in blue). Bootstrap support for populations splits shown as a percent of 1000 bootstraps.

Evidence for selection in sulfide processing genes

A small subset of loci associated with sulfide processes were outliers and therefore putatively under selection in all sulfidic populations. The distribution of differentiation (F_{ST} and d_{xy}) and $\Delta\pi_{ecotype}$ were similar among gene sets but varied among drainages, suggesting variation in demographic history (Figure S1.5). We identified six Mahalanobis outlier gene regions putatively under selection that were

shared among all comparisons (Figure S1.6), of which five were associated with sulfide processes (Table S1.6). These regions included both copies of *ethe1*, galactose-specific lectin nattectin-like, ladderlectin-like, and C-type lectin domain family 10 member A-like. The final outlier locus was a background gene, succinate-CoA ligase ADP-forming subunit beta (*sucla2*). Other notable outlier genes associated with sulfide processes shared among some, but not all comparisons, included solute carrier family 26 member 1 (*slc26a1*) in the Pich 1, Ixta, Puya, and Taco comparisons, mercaptopyruvate sulfurtransferase (*mpst*) in Pich 2, Ixta, Puya and Taco, and *sqor* in the Ixta, Puya and Taco comparisons. Additionally, the OxPhos-associated gene COX subunit 8A (*cox8a*) was considered an outlier in the Pich 1, Pich 2, Puya, and Taco comparisons.

In addition to identifying shared outlier genes between ecotypes within the same drainage, we identified 45 highly differentiated SNPs between all sulfidic and nonsulfidic individuals based on a 99.5% empirical cut-off (Figure 1.4; Table S1.7). Of these 45 outlier SNPs, 35 were associated with sulfide processes, seven in OxPhos genes, and three in background genes (Table S1.7). Notably, the top 10 most differentiated SNPs (F_{ST} 0.87–0.95) were located in two sulfide detoxification genes, *sqor* and *ethe1* (Figure 1.4a). Four were in the 3' UTR of *ethe1.a*, and two were nonsynonymous mutations in *sqor*. The nonsynonymous mutations identified in *sqor* included a change from alanine to valine and from arginine to lysine, but both amino acid changes had a low predicted impact on the structure and function of the protein according to the Polyphen2 score (0.003–0.036 for Ala to Val, 0 for Arg to

Lys). Of the remaining top outlier SNPs, three were synonymous mutations in *sqor* and one was a synonymous mutation in *ethe1* (Table S1.7). In addition to *sqor* and *ethe1*, we found four solute carrier family genes (*slc13a1, slc25a35, slc26a1,* and *slc26a5*) that contained highly differentiated sites (F_{ST} 0.73–0.77, Table S1.7). Of the seven outlier SNPs found in OxPhos genes, six were in COX assembly homologue COX15. We identified three nonsynonymous positions in *cox15*, including an amino acid change from serine to cysteine that was predicted to be possibly damaging (PolyPhen2 = 0.952, Table S1.7). We identified three background genes that each contained a single outlier SNP, hypoxia inducible domain family member 1A (*higd1a*), 2-oxoglutarate dehydrogenase (*ogdh*), and ribosomal protein S15 (*rps15*) (Table S1.7).

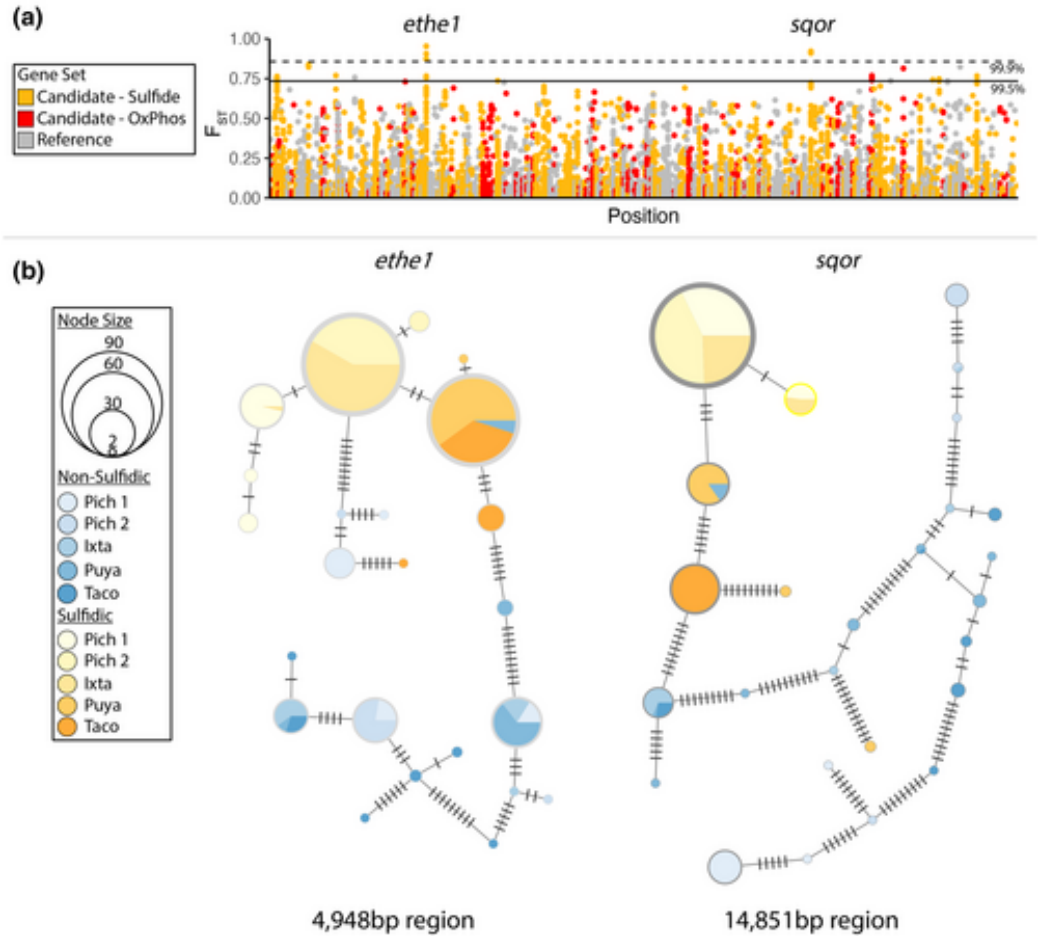


Figure 1.4: Differentiation and evidence for a monophyletic origin of regions putatively under selection in sulfidic populations. (a) Per base pair F_{ST} between all sulfidic and all nonsulfidic individuals coloured by gene set: H_2S detoxification (gold), OxPhos (red) and background (grey). Outlier cut-offs are indicated by horizontal lines (solid line 99.5%, dashed line 99.9%). (b) Haplotype network for outlier genes putatively under selection *ethe1.a* (left) and *sqor* (right). Yellow shades represent sulfidic individuals' haplotype and blues represents nonsulfidic individuals' haplotype. Shade represents population, with the lightest shade representing western populations and darker shades represents eastern populations (see Figure 1.1). Node size represents number of haplotypes found in the dataset. Number of mutations between haplotypes is labelled on the branch as tick marks. Singletons have been removed.

To better understand sequence variation in regions putatively under selection, we generated haplotype networks for outlier genes of interest.

Both *ethe1.a* and *sqor* showed a reduced number of sulfidic haplotypes compared to nonsulfidic haplotypes, evidence for a monophyletic origin (Figure 1.4b), and the topology of these networks varied greatly from background genes (Figure S1.7). There is evidence of haplotype sharing between sulfidic and nonsulfidic individuals from Puya in *sqor* (~15,000 bp region) and among Puya sulfidic, Taco sulfidic, and Puya nonsulfidic individuals in *ethe1.a* (~5000 bp region) (Figure 1.4b). However, this low level of sharing is not surprising given the results of Admixture, PCA, and tree discordance, which suggest ongoing gene flow among these populations (Figures 1.2 and 1.3). Similar to *ethe1.a* and *sqor*, we see a partitioning of haplotypes by ecotype in the *cox15* haplotype network and a low level of sharing between ecotypes (Figure S1.8).

Discussion

Although convergent phenotypic evolution is common, it remains unclear whether convergent phenotypes typically arise through similar genomic changes (Kitano et al., 2022; A. Martin & Orgogozo, 2013). Furthermore, it is unknown how often these similar genomic changes are the result of de novo mutation, standing ancestral variation, or introgressed loci (Rosenblum et al., 2014). Convergent adaptive traits that result from selective pressures imposed by extreme environments provide an opportunity to explore how strong selection may shape convergence at various levels of biological organization. In this study, we utilized a naturally replicated extreme environment—sulfide-rich springs that harbour populations of

fishes that have independently adapted to highly toxic conditions—to test hypotheses about the role of genomic convergence in the independent evolution of sulfide tolerance. In addition to identifying drainage-specific genes under selection, we identified two candidate genes associated with H_2S detoxification in the mitochondria, *sqor* and *ethe1*, that show evidence for selection on shared variation. This study suggests that the convergent evolution of H_2S tolerance in the *P. mexicana* species complex is the result of both shared and unique genomic changes.

Many regions putatively under selection are associated with H_2S detoxification

H_2S is detoxified in the mitochondria via a series of enzymatic reactions associated with the SQR pathway (Libiad et al., 2014; Olson, 2018). This pathway begins with SQOR binding H_2S followed by a series of reactions involving a group of enzymes including ETHE1 that oxidizes H_2S to a variety of excretable compounds that allow elimination of oxidized sulfur molecules from the system (Hildebrandt & Grieshaber, 2008; Libiad et al., 2014; Olson, 2018). Our findings highlight the importance of ETHE1 and SQOR in adaptation to H_2S-rich environments. Previous studies have shown that *ethe1* is differentially expressed between ecotypes (A. P. Brown et al., 2018; Kelley et al., 2016; Passow, Brown, et al., 2017; Passow, Henpita, et al., 2017) Consistent with this, we found highly differentiated SNPs in the 3' UTR region of *ethe1*, which may play a role underlying these patterns of differential gene regulation and expression (Mayr, 2019). Additionally, previous studies have shown that sulfidic populations have higher SQOR activity and lower endogenous levels of

H_2S than nonsulfidic populations at increasing levels of H_2S exposure, suggesting an increased detoxification ability (Greenway et al., 2020). Our analyses identified multiple, highly differentiated positions between all sulfidic and nonsulfidic individuals in coding regions of *sqor*, with two variants leading to differences in the encoded amino acids. Our findings raise the question of whether these amino acid changes are involved in, or potentially directly responsible for, observed increases in SQOR activity in sulfidic populations. Future work to categorize the activity of this enzyme as well as understanding the individual and combinatoric effects of these amino acid changes will be necessary to understand how these changes impact SQOR function and activity.

In addition to strong patterns of differentiation in two key detoxification enzymes, we see repeated patterns of increased differentiation in several other candidate genes, including solute carrier genes. Epistasis could be an important driver of these patterns. For example, mutations that increase the capacity to detoxify H_2S may only be beneficial if they occur when the genomic background contains alleles that allow for an increased capacity to remove the byproducts of detoxification out of the cell. Theoretically, there could be an epistatic relationship between *sqor*, *ethe1*, and solute carrier families, such as slc13 (M. J. Bergeron et al., 2013) and slc26 (Alper & Sharma, 2013) that can remove detoxification byproducts out of the cell, limiting the number of mutational paths as seen in other adaptations (Weinreich et al., 2006).

Although we see limited evidence for selection on genes associated with OxPhos compared to H_2S detoxification genes, previous studies that included mitochondrially encoded subunits of OxPhos have identified parallel amino acid changes associated with increased cox resistance (Greenway et al., 2020; Pfenninger et al., 2014). It is likely that both H_2S regulation and resistance are crucial in adapting to this extreme environment, despite our results showing limited evidence for selection in nuclear-encoded OxPhos genes, beyond *cox8a* and *cox15*. This result could be explained by only needing adaptive modifications to the reactive core of COX, which is mitochondrially encoded, to gain H_2S resistance. Future work to determine if mitonuclear coevolution is necessary for H_2S tolerance or if modification of only mitochondrial genes is sufficient.

Sources of shared variation among drainages

Adaptive variation underlying convergent phenotypes may arise independently among populations as the result of selection on de novo mutations (in the same or different genes), or may be shared via selection on introgressed or ancestral variation (K. Reid et al., 2021; Rosenblum et al., 2014; Stern, 2013). Consistent with previous studies, phylogenetic and admixture analyses provide evidence for the independent evolution of H_2S tolerance among sulfide spring populations (Pfenninger et al., 2014; Tobler et al., 2018). In addition, the sulfidic populations in Puya and Taco are more recently diverged from their nonsulfidic ancestor than those in other drainages (Pfenninger et al., 2014; Tobler et al., 2018).

However, our analyses support selection on shared variation among sulfidic populations in a small subset of genes. This result raises questions as to the source of this putatively adaptive variation. Despite geographical barriers between drainages, seasonal flooding could provide a mechanism for the movement of fish between drainages and the potential for the export of adaptive alleles from a source population, known as the transporter hypothesis (Schluter & Conte, 2009). Recent simulation work provided quantitative proof-of-concept for the transporter hypothesis, showing that only a few individuals are needed for the export of adaptive alleles between populations in similar habitats through gene flow with connected populations in other habitat types (Galloway et al., 2020). We identified shared putatively adaptive haplotypes, greater than 15 kb in length, that could be indicative of recent introgression.

In a recent study, Todesco et al. (2020) identified recently introgressed adaptive haplotype blocks greater than 1 Mb in sunflowers. Given our study used exon capture data, which limits our ability to detect long haplotypes, we were unable to determine if the source of adaptive variation was recent introgression or standing ancestral variation. Distinguishing between the two potential sources of shared adaptive variation remains an important unanswered question in this system.

Evolutionary biologists have long been fascinated by examples of repetitive evolutionary trajectories. But even with the expansion of low-cost sequencing, it remains unclear how often convergent phenotypes are driven by similar genomic changes. Our study adds to the growing body of literature that suggests selection on

shared variation may be an important driver of phenotypic convergence within closely related populations and species (A. P. Brown et al., 2019; Waters & McCulloch, 2021). Future work using whole genome data in this species complex is necessary to elucidate the relative role of selection on introgressed and standing variants. Furthermore, understanding the source of adaptive alleles across the family Poeciliidae, which have independently evolved sulfide tolerance almost 20 times (Tobler et al., 2018), will provide important insight into fundamental questions of the predictability and repeatability of evolution.

Supplemental Material

Supplemental Table 1.1: Populations included in this study. Coordinates are from Greenway et al., 2020.

Site	Site Name	Drainage	H_2S	Species	Lat	Long
Pich 1 S	Baños del Azufre	Pichucalco	+	P.sulphuraria	17.552	-92.998
Pich 1 NS	Vet Station	Pichucalco	-	P.mexicana	17.556	-93.008
Pich 2 S	La Gloria	Pichucalco	+	P.sulphuraria	17.532	-93.015
Pich 2 NS	Rosita	Pichucalco	-	P.mexicana	17.485	-93.103
Ixta S	La Esperanza	Ixtapangajoya	+	P.thermalis	17.511	-92.980
Ixta NS	Rio Ixtapangajoya	Ixtapangajoya	-	P.mexicana	17.495	-92.998
Puya S	La Lluvia, small spring	Puyacatengo	+	P.mexicana	17.463	-92.895
Puya NS	Vicente Guerrero	Puyacatengo	-	P.mexicana	17.51	-92.914
Taco S	El Azufre I	Tacotalpa	+	P.mexicana	17.442	-92.774
Taco NS	Bonita	Tacotalpa	-	P.mexicana	17.426	-92.752

Supplemental Table 1.2: Excel file Table_S2.xlsx. Capture gene list including location, gene name, Gene ID, Transcript ID & Protein ID. In the gene set columns, Broughton refers to reference genes from Broughton & Zhang 2013, NuclearRef refers to nuclear encoded genes involved in other mitochondrial functions, NuOx are nuclear encoded oxidative phosphorylation genes and Sulfide Detox/Response refers to genes with a known sulfide related function. See methods for details about these gene sets.

Supplemental Table 1.3: Gene ontology terms used to generate targeted gene list.

GO Term	Biological Process
GO:0070221	sulfide oxidation, using sulfide:quinone oxidoreductase
GO:0006790	sulfur compound metabolic process
GO:0044273	sulfur compound catabolic process
GO:0070813	hydrogen sulfide metabolic process
GO:0000096	sulfur amino acid metabolic process
GO:0006090	pyruvate metabolic process
GO:0006749	glutathione metabolic process
GO:1901687	glutathione derivative biosynthetic process
GO:0000098	sulfur amino acid catabolic process
GO:0008272	sulfate transport
GO:0006534	cysteine metabolic process
GO:0044272	sulfur compound biosynthetic process

Supplemental Table 1.4: Observed (Ho) and expected (HE) heterozygosity including standard error for each gene set: background, oxphos, and sulfide.

Site	Background H_o	Background H_E	Oxphos H_o	Oxphos H_E	Sulfide H_o	Sulfide H_E
Pich1S	0.098 ± 0.003	0.092 ± 0.003	0.09 ± 0.006	0.082 ± 0.005	0.101 ± 0.003	0.094 ± 0.002
Pich1NS	0.138 ± 0.004	0.134 ± 0.003	0.144 ± 0.007	0.134 ± 0.006	0.15 ± 0.003	0.139 ± 0.003
Pich2S	0.078 ± 0.003	0.076 ± 0.003	0.102 ± 0.007	0.089 ± 0.006	0.082 ± 0.003	0.078 ± 0.002
Pich2NS	0.15 ± 0.004	0.145 ± 0.003	0.154 ± 0.007	0.141 ± 0.006	0.158 ± 0.003	0.149 ± 0.003

Ixta S	0.078 ± 0.003	0.073 ± 0.003	0.084 ± 0.006	0.073 ± 0.005	0.068 ± 0.002	0.064 ± 0.002
Ixta NS	0.133 ± 0.003	0.129 ± 0.003	0.138 ± 0.007	0.127 ± 0.006	0.139 ± 0.003	0.127 ± 0.003
Puya S	0.132 ± 0.004	0.127 ± 0.003	0.145 ± 0.008	0.132 ± 0.006	0.141 ± 0.003	0.128 ± 0.003
Puya NS	0.138 ± 0.004	0.13 ± 0.003	0.159 ± 0.008	0.148 ± 0.007	0.149 ± 0.003	0.136 ± 0.003
Taco S	0.071 ± 0.003	0.067 ± 0.003	0.067 ± 0.007	0.053 ± 0.005	0.081 ± 0.003	0.071 ± 0.002
Taco NS	0.162 ± 0.004	0.152 ± 0.003	0.161 ± 0.007	0.149 ± 0.006	0.156 ± 0.003	0.145 ± 0.003

Supplemental Table 1.5: Estimated nucleotide diversity (π) including standard error for each gene set: background, oxphos, and sulfide.

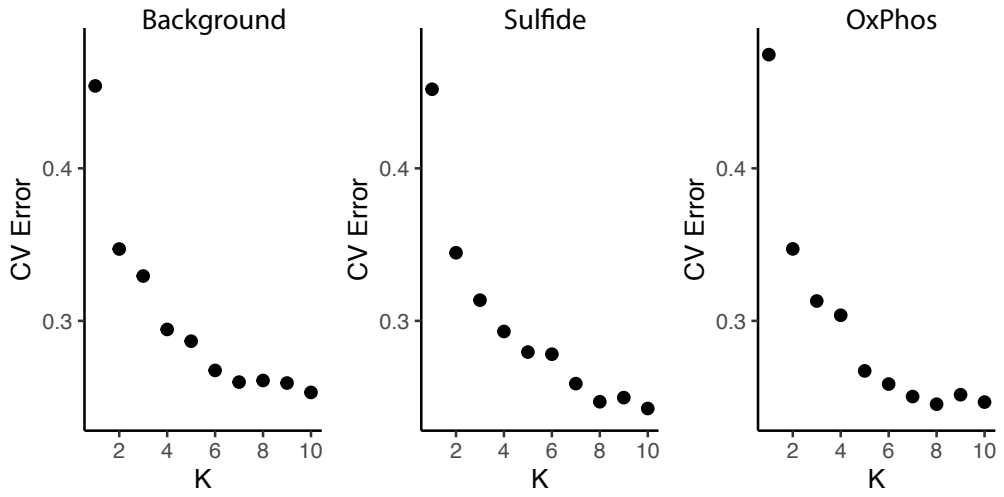
Site	Background π	Oxphos π	Sulfide π
Pich1	0.095 ± 0.003	0.084 ± 0.005	0.097 ± 0.003
Pich1 NS	0.138 ± 0.003	0.138 ± 0.006	0.143 ± 0.003
Pich2 S	0.078 ± 0.003	0.091 ± 0.006	0.08 ± 0.002
Pich2 NS	0.149 ± 0.003	0.145 ± 0.006	0.153 ± 0.003
Ixta S	0.075 ± 0.003	0.075 ± 0.005	0.066 ± 0.002
Ixta NS	0.133 ± 0.003	0.13 ± 0.006	0.131 ± 0.003
Puya S	0.13 ± 0.003	0.135 ± 0.007	0.131 ± 0.003
Puya NS	0.133 ± 0.003	0.153 ± 0.007	0.139 ± 0.003
Taco S	0.069 ± 0.003	0.055 ± 0.005	0.073 ± 0.002
Taco NS	0.156 ± 0.004	0.153 ± 0.006	0.149 ± 0.003

Supplemental Table 1.6: Shared outlier genes in three or more comparisons across ecotype within drainages.

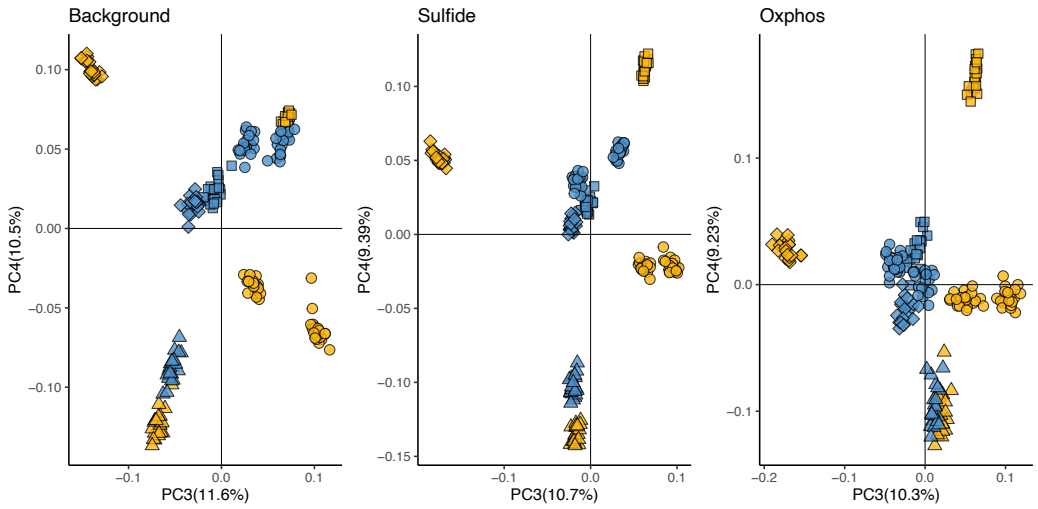
Gene Accession	**Drainage comparison outlier**	**Gene Set**	**Gene name**
LOC106905994	Pich 1/Pich 2/Ixta/Puya/Taco	Sulfide	galactose-specific lectin nattectin-like
LOC106907337	Pich 1/Pich 2/Ixta/Puya/Taco	Sulfide	ladderlectin-like
LOC106917689	Pich 1/Pich 2/Ixta/Puya/Taco	Sulfide	persulfide dioxygenase ETHE1, mitochondrial-like

LOC106917690	Pich 1/Pich 2/Ixta/Puya/Taco	Sulfide	ETHE1 persulfide dioxygenase
LOC106925607	Pich 1/Pich 2/Ixta/Puya/Taco	Sulfide	C-type lectin domain family 10 member A-like
sucla2	Pich 1/Pich 2/Ixta/Puya/Taco	Background	succinate-CoA ligase ADP-forming subunit beta
higd1a	Pich 1/Pich 2/Ixta/Puya	Background	HIG1 hypoxia inducible domain family, member 1A
LOC106908428	Pich 2/Ixta/Puya/Taco	Sulfide	mercaptopyruvate sulfurtransferase
LOC106914136	Pich 1/Ixta/Puya/Taco	Sulfide	solute carrier family 26 member 1
LOC106926066	Pich 1/Pich 2/Puya/Taco	Oxphos	cytochrome c oxidase subunit 8A, mitochondrial-like
LOC106907017	Pich 1/Ixta/Taco	Sulfide	solute carrier family 13 member 2-like
LOC106910378	Pich 1/Pich 2/Taco	Sulfide	ladderlectin-like
LOC106914353	Pich 1/Pich 2/Ixta	Oxphos	NADH:ubiquinone oxidoreductase subunit C2
LOC106919459	Pich 1/Pich 2/Ixta	Oxphos	cytochrome c oxidase subunit 6B1-like
LOC106926427	Pich 1/Ixta/Puya	Background	caspase 10, apoptosis-related cysteine peptidase
mgst1	Pich 1/Pich 2/Taco	Sulfide	microsomal glutathione S-transferase 1
ndufb3	Pich 2/Ixta/Taco	Oxphos	NADH:ubiquinone oxidoreductase subunit B3
ndufv3	Pich 1/Pich 2/Ixta	Oxphos	NADH:ubiquinone oxidoreductase subunit V3
sqor	Ixta/Puya/Taco	Sulfide	sulfide quinone oxidoreductase

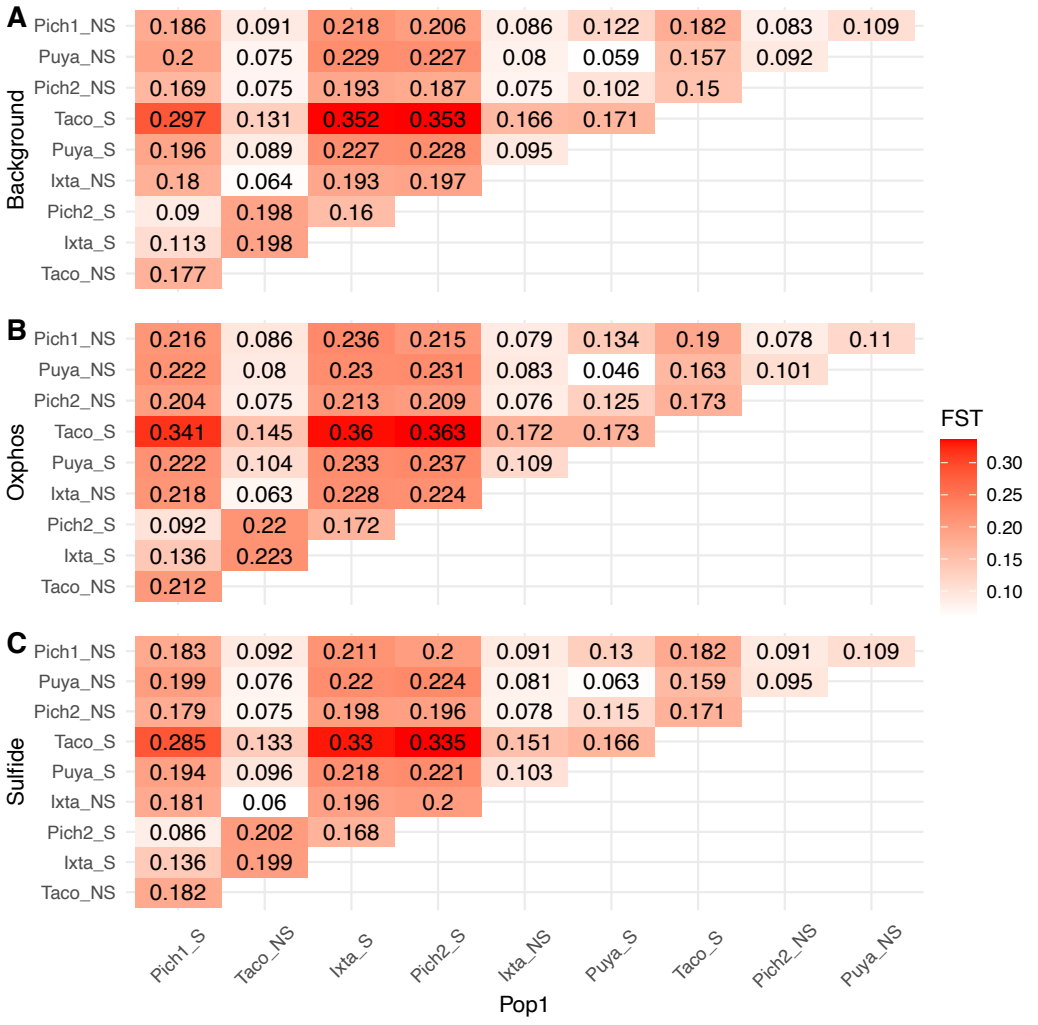
Supplemental Table 1.7: Excel file Capture_Table_S7.xlsx. SNP positions, Weir and Cockerham F_{ST} between all sulfidic and nonsulfidic individuals, gene name, and SNPeff annotation including the type of variant, SNPeff class and amino acid change of 99.5% outliers (denoted by line in excel file). SNPeff class is an estimation based on the putative impact of the variant.



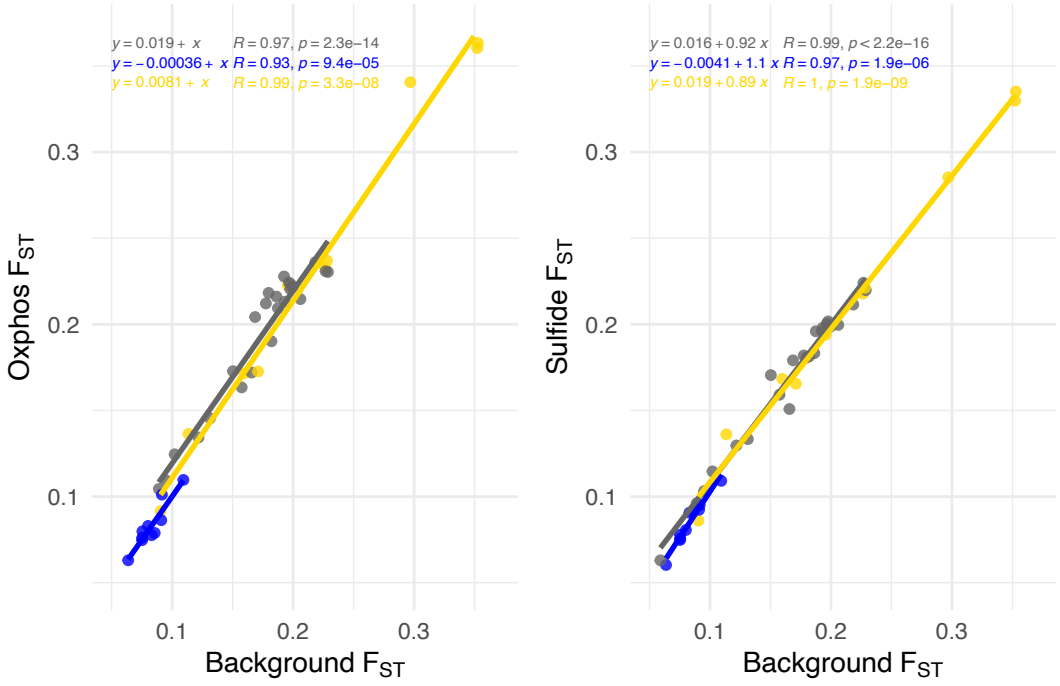
Supplemental Figure 1.1: Admixture CV error for background, sulfide, and OxPhos gene sets for K of 1-10.



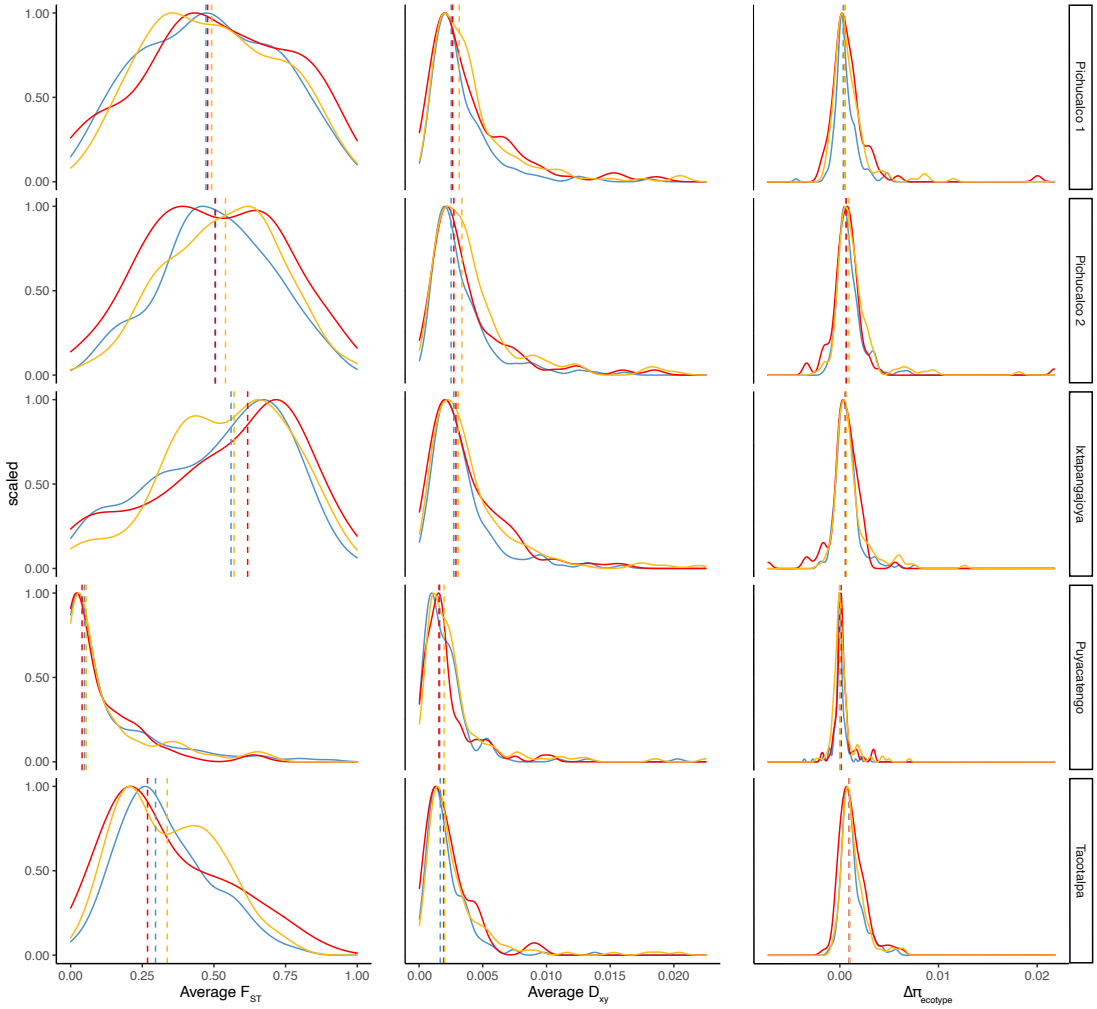
Supplemental Figure 1.2: Principal components three and four of linkage disequilibrium (LD) pruned single nucleotide polymorphisms (SNPs) that account for 22.1% 22.6% of total variation in background, sulfide and oxphos genes. Yellow represents the sulfidic ecotype and blue represents the nonsulfidic ecotype. Shape represents the drainage of origin (■: Tac; ▲: Puy; •: Pich; ◆: Ixta).



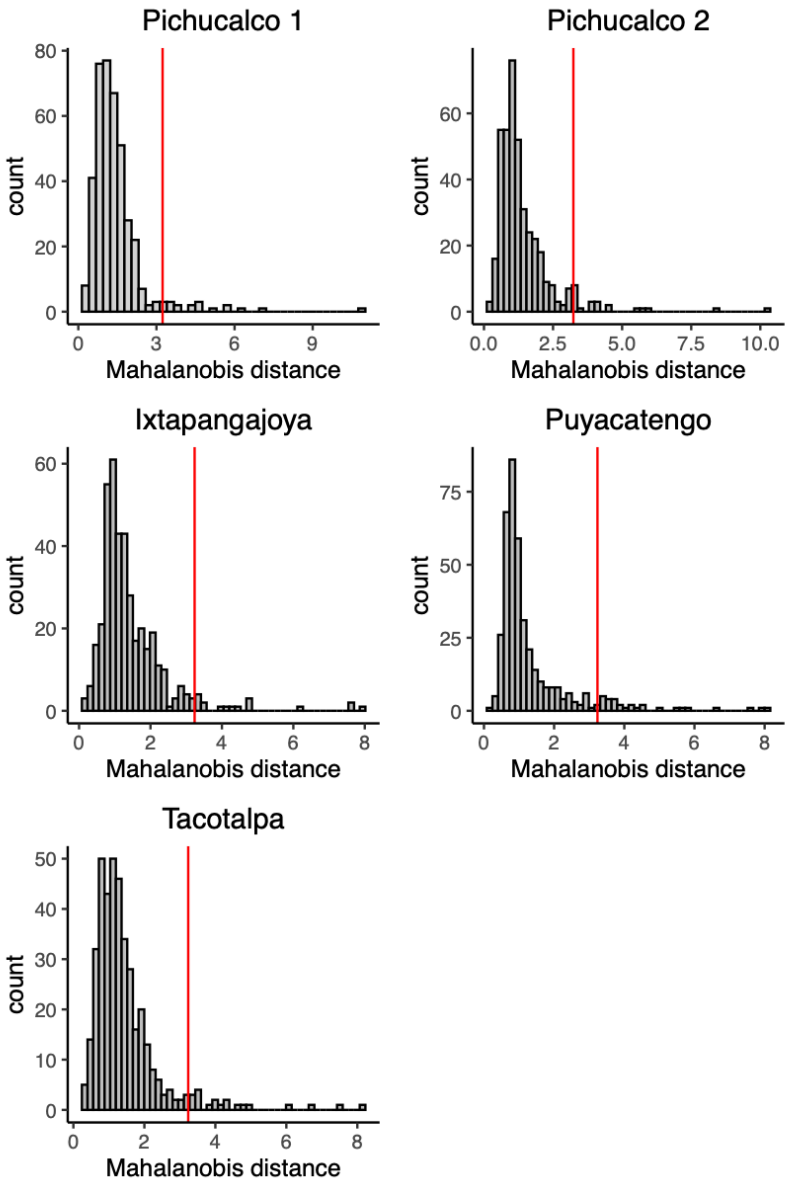
Supplemental Figure 1.3: F_{ST} (Weir and Cockerham) for all pairwise comparisons of populations for (a) background, (b) OxPhos, and (c) sulfide genes.



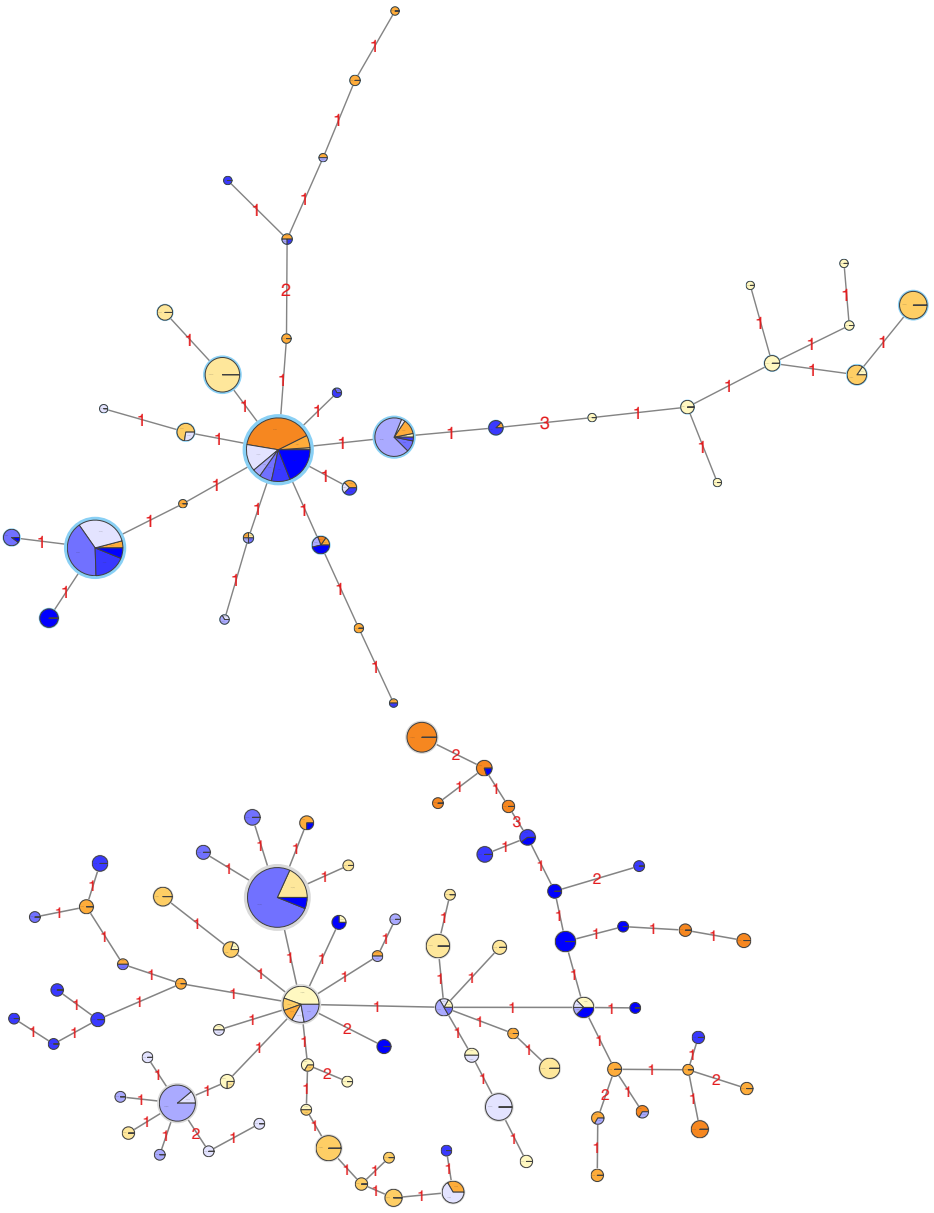
Supplemental Figure 1.4: F_{ST} (Weir and Cockerham) for all pairwise comparisons of populations of Oxphos-related genes (left) and sulfide-related genes (right) compared to background genes. Yellow represents comparisons between sulfidic populations; blue represents comparisons between nonsulfidic populations and grey represents comparisons of differing ecotypes.



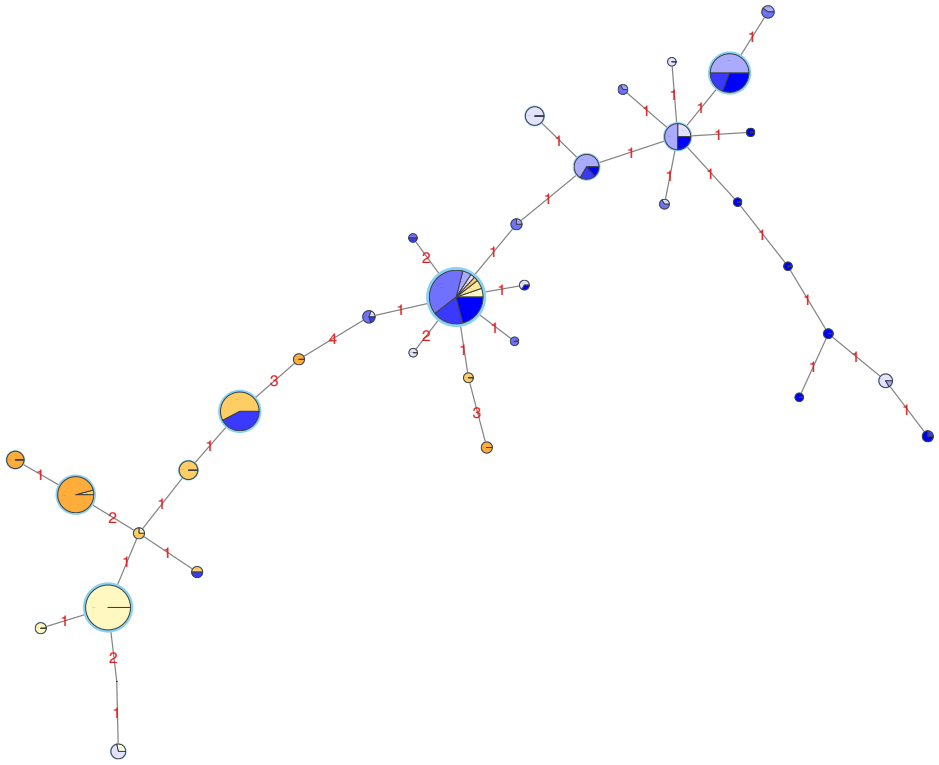
Supplemental Figure 1.5: The distribution of differentiation (F_{ST} and Dxy) and difference in diversity between ecotypes ($\Delta\pi$ ecotype) for reference (blue), oxphos (red) and sulfide (gold) genes.



Supplemental Figure 1.6: Distributions of Mahalanobis distances for each gene plotted by drainage. Populations of different ecotypes of the same drainage were compared. Vertical line (red) denotes 95% outlier.



Supplemental Figure 1.7: Haplotype network of two randomly selected reference genes: gapdh (top) and pdss2 (bottom). Yellow shades represent sulfidic individuals' haplotype and blues represents nonsulfidic individuals' haplotype. Shade represents population, with the lightest shade representing western populations and darker shades represents eastern populations (see Figure 1). Node size represents number of haplotypes found in the dataset. Number of mutations between haplotypes is labeled on the branch. Note: singletons have been removed.



Supplemental Figure 1.8: Haplotype network of *cox15.* The legend is the same as figure S1.7.

Chapter 2 NEW GENOME ASSEMBLIES FOR POECILIIDAE: A FOUNDATION FOR ADAPTATION STUDIES

Kara Ryan [1], Rishi De-Kayne [1,2], Jeremy Davis [1], Lenin Arias-Rodriguez [3], Michael Tobler [4,5,6], Joanna L. Kelley [1]

1. Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95060, USA
2. Current address: Department of Integrative Biology, University of California Berkeley, Berkeley, CA 94720, USA
3. División Académica de Ciencias Biológicas, Universidad Juárez Autónoma de Tabasco (UJAT), Villahermosa, México
4. Department of Biology, University of Missouri - St. Louis, St. Louis, MO, 63121, USA
5. Whitney R. Harris World Ecology Center, University of Missouri - St. Louis, St. Louis, MO, 63121, USA
6. WildCare Institute, Saint Louis Zoo, St. Louis, MO 63110, USA

Abstract

Multiple lineages in the family Poeciliidae have independently adapted to hydrogen-sulfide-rich springs. The independent colonizations of such springs mean that there are naturally replicated lineages that provide a strong model for studying adaptation and convergent evolution. However, there are limited genomic resources for many genera and species across Poeciliidae. Here, we present six high-quality, chromosome-level, annotated genome assemblies for *Poecilia* and *Gambusia* populations, five of which are the first for the species or ecotype, and the remaining assembly improved the current reference genome contiguity by more than 100-fold. Using these new assemblies, we compared repeat content and modeled historical changes in effective population size.

Introduction

Vertebrates have adapted to survive and even thrive in extreme environments (Bell, 2012). Numerous fish species, for example, have evolved a range of adaptive phenotypes to persist in a variety of stressful conditions, including extreme temperatures, oxygen concentrations, and toxicant levels, to name a few (Riesch et al., 2015). Determining the mechanisms resulting in, and the genetic basis underlying, convergent adaptive phenotypes in extremophile lineages is particularly valuable, given that it remains unclear if genomic convergence is a common feature of repeated adaptation (Xu et al., 2020). This knowledge informs understanding of constraints on potential adaptive solutions to similar stressors and predictability at various levels of biological organization.

Fishes within the family Poeciliidae lend themselves to investigations of convergent evolution because multiple independent lineages have adapted to springs rich in hydrogen sulfide (H_2S) (Greenway et al., 2014; Tobler et al., 2018). Across taxa, endogenously produced H_2S is an important signaling molecule in proper physiological function (L. Li et al., 2011). However, environmental H_2S is toxic as it halts aerobic ATP production via inhibition of oxidative phosphorylation (Cooper & Brown, 2008). H_2S also creates hypoxic conditions in aquatic environments (Tobler et al., 2016). Despite these highly toxic conditions, 19 evolutionary independent lineages of the family Poeciliidae have adapted to H_2S-rich springs (Tobler et al., 2018).

The presence of multiple populations of Poeciliid fishes that have independently adapted to H_2S-rich springs and closely related populations in adjacent, nonsulfidic habitats provide a strong comparative framework for exploring convergent evolution. Sulfide-adapted populations have convergently evolved a range of phenotypic shifts, including having enlarged heads, mouths, and gills (Camarillo et al., 2020) and changes to life history traits (Riesch et al., 2014), metabolism (Passow, Arias-Rodriguez, et al., 2017), (mitochondrial function (Greenway et al., 2020), and even behavior (Bierbach et al., 2018; Riesch et al., 2009). At the genomic level, there is evidence of convergence in gene regulation, both in terms of wild-type gene expression and expression changes upon exposure to H_2S in the laboratory (A. P. Brown et al., 2018; Greenway et al., 2020; Kelley et al., 2016; Passow, Brown, et al., 2017; Passow, Henpita, et al., 2017; Perry et al., 2024). However, there is varied support for genomic convergence of amino acids, gene sequence, and gene pathways involved in H_2S adaptation across sulfidic populations of *P. mexicana* (A. P. Brown et al., 2018; De-Kayne et al., 2024; Greenway et al., 2020, 2024; Pfenninger et al., 2014; Ryan et al., 2023). Despite some evidence for convergence at the amino acid level from independent *de novo* mutation (Greenway et al., 2020, Pfeninger et al., 2014), most convergence at the nucleotide level is the result of shared variation rather than independent *de novo* mutation in the mitochondrial genome (Greenway et al., 2020, Ryan et al., 2023, Brown et al., 2019). While chromosome-level assemblies are available for Poeciliidae species, these taxa have diverged from sulfide-adapted lineages of interest approximately 15-20MYA (TimeTree), limiting our ability to

accurately map short-read sequencing data to evolutionary distance. Furthermore, existing reference genomes typically represent a non-extremophile ecotype that differ greatly in life history, morphology, physiology and other important biological traits. The lack of high-quality genomic resources for extremophile lineages has limited the ability to use comparative genomic and population genetic approaches to understand sulfide adaptation at the genomic level.

To address this resource gap, we generated six new genome assemblies, including four extremophile populations and two closely related, non-extremophile populations across two genera, *Poecilia* and *Gambusia*, using PacBio HiFi sequencing. Three of these assemblies are the first assemblies of the species (*Gambusia sexradiata* nonsulfidic ecotype, *Gambusia eurystoma,* and *Poecilia sulphuraria*). Of the remaining assemblies, one (*Poecilia mexicana* nonsulfidic ecotype) improves the current reference genome contiguity, measured as N50, from 275 kilobases (kb) to 33 Megabases (Mb), and the remaining two (*Poecilia mexicana* sulfidic ecotype and *Gambusia sexradiata* sulfidic ecotype) are the first genome assemblies for the sulfidic ecotypes of this species. We used these new assemblies to estimate repetitive content and explore historical changes in effective population size. We found that H_2S-adapted fishes tend to have higher repetitive content than nonsulfidic populations within a genus. Furthermore, we found that most, but not all, H_2S-adapted populations have reduced effective population sizes compared to closely related, nonsulfidic populations. These six new chromosome-scale reference genomes

provide new resources for population genetic and comparative studies across Poecliidae.

Methods

Sampling

Organisms sequenced in this study include species in the genera *Poecilia* (sulfidic: *P. mexicana* and *P. sulphuraria;* nonsulfidic: *P. mexicana*) and *Gambusia* (sulfidic: *G. eurystoma* and *G. sexradiata*; nonsulfidic: *G. sexradiata*). Samples were collected with a seine (2 × 4 meters) from hydrogen-sulfide-rich springs in the Rio Grijalva basin in Southern Mexico, and from adjacent nonsulfidic streams and lagunas in the same region (Table S2.1). For all individuals except *P. sulphuraria*, within 12 hours of capture, fish were sacrificed, dissected using previously sterilized scissors and forceps, and preserved in 2 ml of RNAlater (Ambion, Inc.), stored at room temperature for one week then at –80°C until DNA extraction. For the *P. sulphuraria* individual, the lab-reared sample was dissected, flash-frozen using liquid nitrogen, and stored at –80°C until DNA extraction. Both *P. mexicana* individuals and all *Gambusia* individuals were males. The *P. sulphuraria* individual was female due to sample availability.

Library preparation and sequencing

Due to sample availability and individual size, tissue choice varied among samples but was chosen to optimize the high molecular weight DNA quantity. DNA

was extracted from the heart for *P. sulphuraria* and *P. mexicana* nonsulfidic ecotype, a fin clip for *P. mexicana* sulfidic ecotype, a whole fish except for the skin, gills, and gut for *G. eurystoma*, and from skeletal muscle for both *G. sexradiata* ecotypes. DNA was extracted using the Circulomics Nanobind High Molecular Weight DNA extraction kit. Tissues were homogenized using an Omni Bead Ruptor 96 using a 2.8-mm ceramic bead per sample or by cryofracture using liquid nitrogen and a Covarias tube. A 2-hour proteinase K digestion was added to increase yield due to small amounts of initial tissue input. DNA concentration was assessed using the Qubit™ High Sensitivity DNA kit in triplicate, Quantus™ Fluorometer, and Nanodrop. DNA quality and fragment size were evaluated using an Agilent™ Femto Pulse system and TapeStation. SMRTbell library preparation and sequencing was done at the UC Davis Genome Center using two SMRT cells on a PacBio Revio sequencer. Only HiFi reads, generated using the PacBio CCS algorithm with the SMRT Link software, were used for the remainder of the analysis.

Quality control and k-mer complexity analysis

Read quality was assessed using FastQC v0.11.9 (Andrews, 2010), and read summary statistics were generated using SeqKit v2.6.1. We used Jellyfish v2.2.10 (Marçais & Kingsford, 2011) *count* and *histo* to perform a k-mer-based complexity analysis of each individual (k-mer size=21). GenomeScope2.0 (Ranallo-Benavidez et al., 2020) was then used to estimate genome size, heterozygosity, and the percentage of duplicates.

Genome Assembly

Reads were assembled using Hifiasm v0.16.1-r375 (Cheng et al., 2021, 2022). Contaminants were identified using Kraken2 using a custom database that included archaea, bacteria, protozoa, humans, fungi, plants, and *P. mexicana* RefSeq genomes/proteins. The remaining contigs were scaffolded into pseudochromosomes using ragtag v2.1.0 (Alonge et al., 2022) *scaffold* with a 23-chromosome *Poecilia reticulata* reference (GCF_000633615.1) for *Poecilia* and a 24-chromosome *Gambusia affinis* reference (GCF_019740435.1) for *Gambusia* using default parameters, where gaps are set to 100bp. Due to the presence of an obvious misassembly, which joined three unique chromosomes into one >90 Mb sequence, we subset the *P. reticulata* reference to include only the three highly-syntenic chromosomes using seqtk subseq (https://github.com/lh3/seqtk) and used ragtag *correct* with the –inter parameter. We then re-scaffolded the assembly against *P. reticulata*. Finally, we identified and plotted synteny for *Poecilia* and *Gambusia* using ntSynt v.1.0.2 (Coombe et al., 2024) and ntSynt-Viz v.1.0.0 (Coombe et al., 2025), visualizing synteny blocks greater than 0.5 Mb. A neighbor-joining cladogram was generated using the pairwise distance between synteny blocks using ntSynt-Viz v.1.0.0 and Quicktree v2.5 (https://github.com/khowe/quicktree). Read coverage was assessed with MosDepth v0.3.4 (Pedersen & Quinlan, 2018). Assemblathon_stats.pl (Bradnam et al., 2013) was used to calculate N50, L50, and percent GC content for the contig and chromosome level assemblies. Completeness was assessed using

BUSCO v5.4.3 (Manni et al., 2021) with the cyprinodontiformes_odb10 lineage dataset.

To annotate the mitochondria, we used MitoHiFi v3.2.2 (Uliano-Silva et al., 2023) to identify contigs that correspond to the mitochondria. To do so, we used the *P. mexicana* mitochondrial reference genome (NC_035306.1) for *Poecilia* individuals and *G. affinis* (NC_004388.1) for *Gambusia* individuals as references to search the assembled contigs. For assemblies where multiple mitochondria were present, contigs with lower read support were removed from the assembly.

Genome structural annotation

To support the pseudo-chromosome status of the assemblies, we used A Telomere Identification toolkit (tidk) v.0.2.65 (M. R. Brown et al., 2025) to find known repeats for Poeciliidae. First, we build the existing database of repeats using tidk build. Then, we modified the table to add Poeciliidae-specific repeats from Telobase (Lyčka et al., 2024). Then, we used tidk find and tidk plot to find and plot telomeric repeats for the Cypriniformes clade. Extension *De-novo* TE Annotator v2.2.2 (Ou et al., 2019) was used to generate a transposable element (TE) library using coding sequences from previous genome assemblies of *P. mexicana* (NCBI: GCF_001443325.1) for *Poecilia* and *G. affinis* (NCBI: GCF_019740435.1) for *Gambusia* species. The resulting TE library was used to finalize the repeat annotation with RepeatMasker v4.1.5 (Smit, Hubley & Green RepeatMasker

at http://repeatmasker.org). Structural gene annotation was performed using the BRAKER pipeline v3.0.3 (Lars Gabriel et al., 2023) using the vertebrate OrthoDB v11 (Kuznetsov et al., 2023) protein database modified to include *G. affinis* and *P. mexicana* proteins as our protein reference. We used previously collected RNA sequencing (RNAseq) data (McGowan et al., 2019; Passow, Brown, et al., 2017; Passow, Henpita, et al., 2017), which included brain, eye, gill, and liver from four populations of *P. mexicana,* including two sulfidic and nonsulfidic populations and gill tissue with and without exposure to hydrogen sulfide in common-garden-reared fish as RNA evidence for *Poecilia*. For *Gambusia* species, we used *G. affinis*, *G. sexradiata,* and *G. eurystoma* RNAseq data which included liver, gill, and brain tissue (Coffin et al., 2022; Greenway et al., 2020). RNAseq reads were trimmed using TrimGalore v0.6.6 (github.com/FelixKrueger/TrimGalore) with default parameters, then mapped to each genome using Hisat2 v2.2.1 (D. Kim et al., 2019). The resulting SAM files were converted to Binary Alignment Map (BAM) files and sorted using Samtools v1.11 (H. Li et al., 2009). Alignment statistics were generated using Picard v2.21.4 (broadinstitute.github.io/picard/).

To assess the completeness of the BRAKER annotations, we first filtered the GTF file to retain the largest isoform per gene using AGAT v0.7.0 (Dainat, n.d.) script agat_sp_keep_longest_isoform.pl. We then used Gff Utilities GffRead (Pertea & Pertea, 2020) to get the protein sequence from the GFF. We ran BUSCO v5.4.3 in protein mode with the cyprinodontiformes_odb10 lineage dataset to assess completeness.

Estimation of demographic history

To investigate the demographic history of these species, we modeled changes in effective population size (N_e) over time using PSMC (H. Li & Durbin, 2011). We first aligned long reads from each species to the corresponding nonsulfidic genome within that genus using minimap2 v2.26 (H. Li, 2018) and used Samtools v1.11 to sort and convert the SAM files to BAM files. Then, we used BCFtools v.1.10.2 (Petr Danecek et al., 2021) *mpileup* and *call* to call and filter variants (-Q 30 -q 30 -d 5 -D 50). Consensus sequences were generated using BCFtools vcfutils.pl. The PSMC companion scripts fq2psmcfa and splitfa were used to convert the consensus sequence into a psmc format file and split long chromosomes into shorter segments. We then ran psmc v0.6.5-v67 with 100 bootstraps (-N25 -t15 -r5 -b -p "1 + 1 + 1 + 1 + 25*2 + 4 + 6" following suggestions in (Hilgers et al., 2024). For visualization, the results were imported into R using the psmcr package (github.com/emmanuelparadis/psmcr) and plotted using ggplot2 using a mutation rate of 2.9e-8 (Y. Lin et al., 2023) and a generation time of 6 months (Jourdan et al., 2014).

Results

Genome sequencing and Assembly

PacBio HiFi sequencing resulted in 1,316,776 to 2,484,081 HiFi reads per sample with a read N50 of 6,888 to 14,370 bp, a median read length of 5,499 to 11,499, an average Phred quality score of 39, and a guanine-cytosine (GC) content of

38.89% to 39.50% (Table S2.1). The k-mer complexity analysis estimated that *P. mexicana* from the sulfidic habitat had the lowest heterozygosity (0.103%) and *P. mexicana* from the nonsulfidic habitat had the highest heterozygosity (0.461%) (Figure S2.1).

Assembled genome sizes ranged from 722.0 Mb to 728.1 Mb for *Poecilia,* which is shorter than the fragmented *P. mexicana* reference genome (GCF_001443325.1, 801 Mb) but is similar to the chromosome-level assembly of *Poecilia reticulata* (GCF_000633615.1, 731 Mb) (Fraser et al., 2020) and the *P. mexicana* haplotype of *Poecilia Formosa* (GCA_013036085.2, 745 Mb). For the *Gambusia* species, assembled genome sizes ranged from 686.8 Mb to 670.0 Mb, which is similar to the *G. affinis* assembly (GCF_019740435.1 680 Mb) (Table S2.3). For all assemblies, over 99% of bases were scaffolded into chromosomes ($N = 23$ for *Poecilia* and $N = 24$ for *Gambusia*) (Table S2.4) with a genome N50 of 29.6 to 32.3 Mb. The average GC% was slightly lower in *Gambusia* (39.03 ± 0.01) than in *Poecilia* (39.50 ± 0.017) (Table S2.5).

Genome Synteny and Completeness

Assembled chromosomes showed a high level of synteny with previously identified linkage groups for *P. reticulata* and *G. affinis,* with some notable exceptions, including several translocations and inversions (Figure 2.1 A&B). This includes two translocations into LG6 from LG16 and LG19 that were assembled in all *Poecilia* assemblies but not present in *P. reticulata* (Figure 2.1A). However, the

lack of Hi-C data for these samples limits further investigation into these large-scale chromosomal rearrangements. Similar to previous work on *Gambusia* (Jourdan et al., 2016), the phylogeny generated from genome-wide synteny blocks placed *G. eurystoma* within *G. sexradiata* (Figure 2.1B). Benchmarking Universal Single-Copy Ortholog (BUSCO) analyses revealed that all genome assemblies included 94.6% to 95.5% of expected complete genes, based on 15,213 genes (Table 2.1) consistent with other assemblies in Poeciliidae (Fraser et al., 2020; van Kruistum et al., 2020; Y. Zhang et al., 2023).

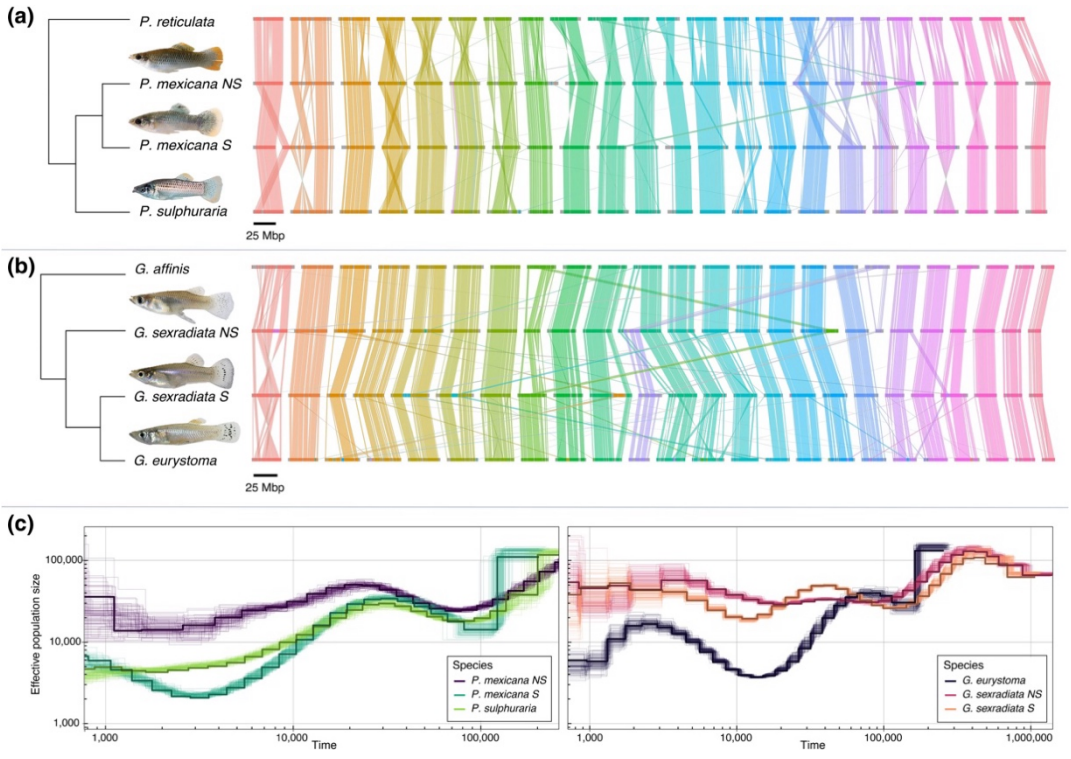


Figure 2.1: Genome synteny, completeness, and repeat content. a) Syntenic regions greater than 0.5 Mb between linkage groups 1 to 23 of *P. reticulata* and *Poecilia* genome assemblies. b) Syntenic regions greater than 0.5 Mb between linkage groups 1 to 24 of *G. affinis* and *Gambusia* genome assemblies. The

neighbor-joining cladograms were generated using pairwise distances. Photos of fish assembled in this analysis, shown on the cladogram, are not to scale and are credited to Michael Tobler. c) Historical changes in effective population size inferred from the Pairwise Sequentially Markovian Coalescent (PSMC) model for Poecilia (left) and Gambusia (right).

Table 2.1: Genome assembly and annotation statistics. Contig and scaffold-level genome assembly completeness, including BUSCO scores (shown in % of total BUSCOs) of genomes and annotations (cyprinodontiformes_odb10, *n* = 15,213), including percent of complete single copy (CS), complete duplicated (CD), fragmented (F) and missing (M) genes.

Species and ecotype	**# of contigs**	**Contig assembly size**	**Contig N50**	**# of scaffolds**	**scaffold N50**	***N* gaps (bp)**
P. sulphuraria (S)	110	723,450,955	28,388,437	64	32,287,255	4,600
P. mexicana (S)	109	721,996,367	27,957,565	80	31,502,199	15,700
P. mexicana (NS)	186	728,088,823	20,367,937	75	33,770,996	11,000
G. eurystoma (S)	714	686,808,650	2,735,428	159	30,179,505	54,600
G. sexradiata (NS)	225	669,107,982	16,588,268	75	30,375,227	14,800
G. sexradiata (S)	265	669,986,110	14,260,791	61	29,623,307	20,300

	Genome BUSCO				**Annotation BUSCO**			
Species	**CS**	**CD**	**F**	**M**	**CS**	**CD**	**F**	**M**
P. sulphuraria (S)	94.9	0.6	0.3	4.2	94.5	0.8	0.3	4.4
P. mexicana (S)	94.8	0.6	0.4	4.2	94.6	0.7	0.4	4.3
P. mexicana (NS)	94.0	1.3	0.4	4.3	91.9	1.3	0.4	6.4
G. eurystoma (S)	91.2	3.4	0.4	5.0	90.7	2.5	0.4	6.4
G. sexradiata (NS)	93.3	1.3	0.3	5.1	95.1	1.4	0.3	3.2
G. sexradiata (S)	93.2	1.8	0.4	4.6	94.6	1.7	0.2	3.5

Genome annotation

The proportion of the genome spanned by repetitive elements was similar among species of the same genus. Nonsulfidic *P. mexicana* had the lowest proportion of repeat content (20.56% of the genome) and *G. eurystoma* had the highest (24.01% of the genome) (Figure S2.3). *Gambusia* species had significantly more repetitive

content than *Poecilia* species (mean *Gambusia* 23.82 ± 0.16% compared to *Poecilia* 20.86 ± 0.31%, $t = 14.7$, df = 3, p-value < 0.001), with the largest differences due to DNA transposons (Figure S2.3). A range of 23,217 to 25,604 genes were annotated (Table S2.6) with other *Poecilia* and *Gambusia* annotations of roughly 25,000 to 28,000 genes, of which ~22,000 to 25,000 are protein-coding (GCF_001443325.1, GCF_019740435.1).

Demographic history

Previous genomic studies have suggested that H_2S-adapted populations have a smaller effective population size (N_e) than closely related nonsulfidic populations (A. P. Brown et al., 2018; Greenway et al., 2024). To explore this, we used Pairwise Sequential Markovian Coalescent (PSMC) to estimate the historical change in N_e (Figure 2.1C). In *Poecilia*, we found that all populations had a similar N_e (~12.5k individuals) until approximately 30 kya, when both H_2S-adapted populations showed a reduction in N_e relative to the nonsulfidic population (Figure 2.1C). At approximately 12 kya (N_e ~10.5k), the demographic histories of H_2S-adapted *Poecilia* split, where the *P. mexciana* sulfidic ecotype (*P. mexicana* S) experienced a larger reduction in population size followed by expansion (N_e = 2k individuals at 3 kya to 5k individuals at 1kya) compared to *P. sulphuraria,* which experience a slowly stabilizing decline to ~5k individuals at 1kya. This pattern was surprising given that previous work has supported that H_2S-adapted and nonadapted populations of *P. mexicana* have only diverged approximately 10 kya (Brown et al.,

2018) and that *P. mexicana* has diverged from *P. sulphuraria* approximately 20 kya (Greenway et al., 2024). However, the germline mutation rate is highly variable in some Poeciliid species (Y. Lin et al., 2023); therefore, this could result from poor estimates of the mutation rate for these populations or the result of complex evolutionary histories, such as ongoing gene flow.

Interestingly, the *Gambusia* species did not follow the same trend (Figure 1C). Generally, *Gambusia* had a higher N_e than *Poecilia*. Similar to H_2S-adapted populations in *Poecilia*, *G. eurystoma* consistently had a lower N_e than the nonsulfidic population and showed a large population decline (N_e of 40k reduced to 4k) from ~70 kya to 10.5 kya, followed by a population expansion until ~2.5 kya (N_e = 15k). However, H_2S-adapted *G. sexradiata* maintained a similar N_e as the nonadapted populations, although the timing and/or magnitude of population declines and expansions differ until 2 kya, where bootstrap support weakens. The challenges associated with estimating effective population size in populations with complex evolutionary histories make these results well-suited for further investigation with other complementary methods, such as site frequency spectrum-based approaches.

Discussion

Lineages that have independently adapted to extreme environments provide an opportunity to explore how often phenotypic convergence is underpinned by convergence at the genome level. The generation of high-quality genome assemblies and annotations provides an opportunity to utilize population genetic and comparative

genomic approaches to explore the genetic basis of convergent traits from single nucleotide variation through large-scale genomic rearrangements. These six new reference genomes, including genomes for lineages displaying a variety of interesting phenotypes, will add to the expanding number of genomic resources for understanding the biology of Poeciliidae.

Supplementary material

Supplemental Table 2.1 Study sites, species, ecotype, and sampling coordinates for individuals sequenced.

Site	**Species**	**Ecotype**	**Sampling coordinate**
La Gloria	*P. sulphuraria*	Sulfidic	17.532, -93.015
El Azufre I	*P. mexicana*	Sulfidic	17.442, -92.774
Arroyo Bonita	*P. mexicana*	Nonsulfidic	17.426, -92.752
Baños el Azufre	*G. eurystoma*	Sulfidic	17.552, -92.998
Mogote de Puyacatengo	*G. sexradiata*	Sulfidic	17.582, -92.899
Caparroso	*G. sexradiata*	Nonsulfidic	18.341, -92.792

Supplemental Table 2.2: Results of HiFi sequencing, including the number of HiFi sequences per sample, the total number of Gigabases (Gb) generated for each individual, median read length, read N50, and GC content.

Species	**HiFi seqs**	**Total bp (Gb)**	**Median Read Length (bp)**	**Read N50**	**GC %**
P. sulphuraria (S)	2,156,321	23.37	10,999	14,370	39.50
P. mexicana (S)	2,273,495	25.00	10,499	13,640	39.49
P. mexicana (NS)	1,487,498	16.75	11,499	13,377	39.36
G. eurystoma (S)	1,316,776	11.53	8,499	10,498	39.03
G. sexradiata (NS)	2,484,081	15.98	5,499	6,888	38.99
G. sexradiata (S)	1,529,192	13.80	8,499	10,752	38.89

Supplemental Table 2.3: HiFiasm assembly results prior to scaffolding, including the number of contigs assembled, the size of the assembly, the number of contigs greater than 10Mb, the assembly N50 and the L50 count.

Species	**# of contigs**	**Assembly size**	**Contigs >10M**	**N50**	**L50**
P. sulphuraria (S)	110	723,450,955	23	28,388,437	11
P. mexicana (S)	109	721,996,367	23	27,957,565	10
P. mexicana (NS)	186	728,088,823	28	20,367,937	14
G. eurystoma (S)	714	686,808,650	2	2,735,428	71
G. sexradiata (NS)	225	669,107,982	24	16,588,268	14
G. sexradiata (S)	265	669,986,110	23	14,260,791	17

Supplemental Table 2.4: Scaffolding results against *Poecilia reticulata* (N=23 linkage groups) and *Gambusia affinis* (N=24 linkage groups) as references. Scaffolding results include the number of contigs placed into linkage groups, placed base pairs (bp), the number of unplaced contigs, unplaced bp, the number and size of gaps, and the percent of bases assembled into linkage groups.

Species and ecotype	**Placed contigs**	**Placed bp**	**Unplaced contigs**	**Unplaced bp**	**Gap bp**	**Gap seqs**	**% in linkage groups**
P. sulphuraria (S)	70	721,960,724	40	1,490,231	10,000	100	99.79
P. mexicana (S)	124	720,140,949	56	1,861,118	5,500	55	99.79
P. mexicana (NS)	134	725,596,201	52	2,492,622	11,000	110	99.66
G. eurystoma (S)	571	680,079,900	143	6,728,750	54,600	546	99.02
G. sexradiata (NS)	173	665,561,320	52	3,5466,62	14,800	148	99.47
G. sexradiata (S)	228	667,602,937	37	2,383,173	20,300	203	99.64

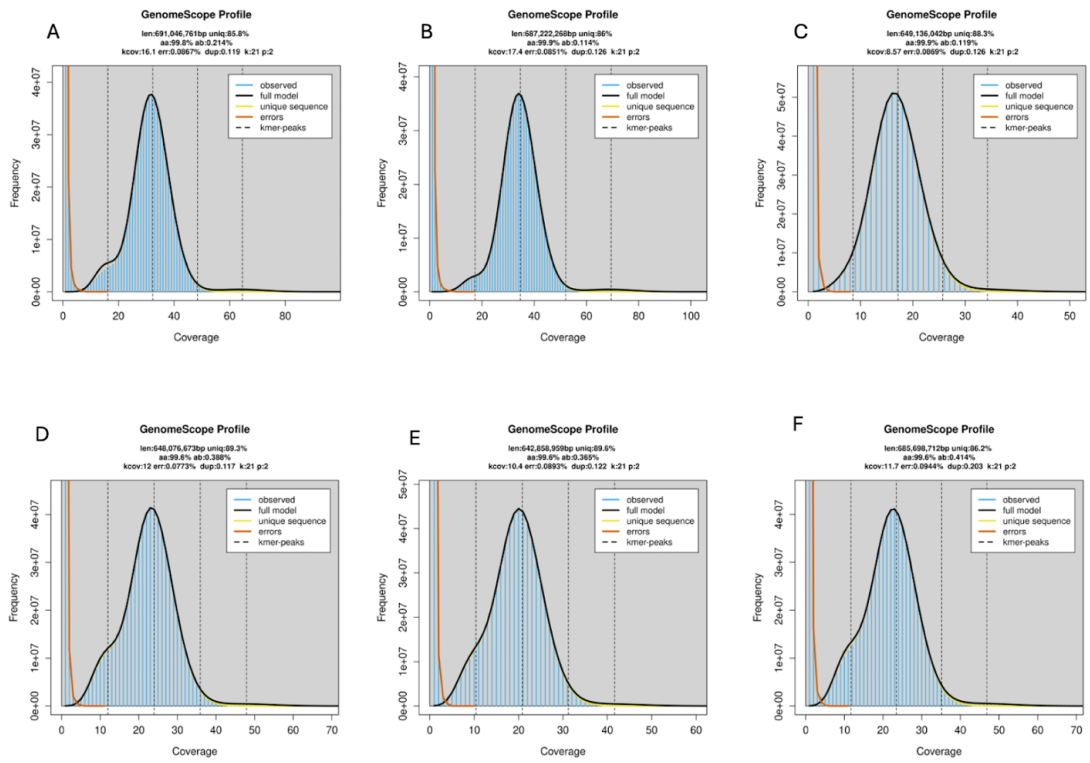
Supplemental Table 2.5: Final genome assembly results, including the number of sequences, the total genome length, the shortest unplaced contig, the average contig length, the N50, the GC%, and the total number of gaps.

Species and ecotype	**# seqs**	**Length**	**Min len**	**Mean len**	**N50**	**GC %**	**N gaps**
P. sulphuraria (S)	64	723,455,555	9,156	11,303,993	32,287,255	39.51	46
P. mexicana (S)	80	722,012,067	77	9,025,150	31,502,199	39.51	157
P. mexicana (NS)	76	728,099,823	14,743	9,580,260	33,770,996	39.48	110

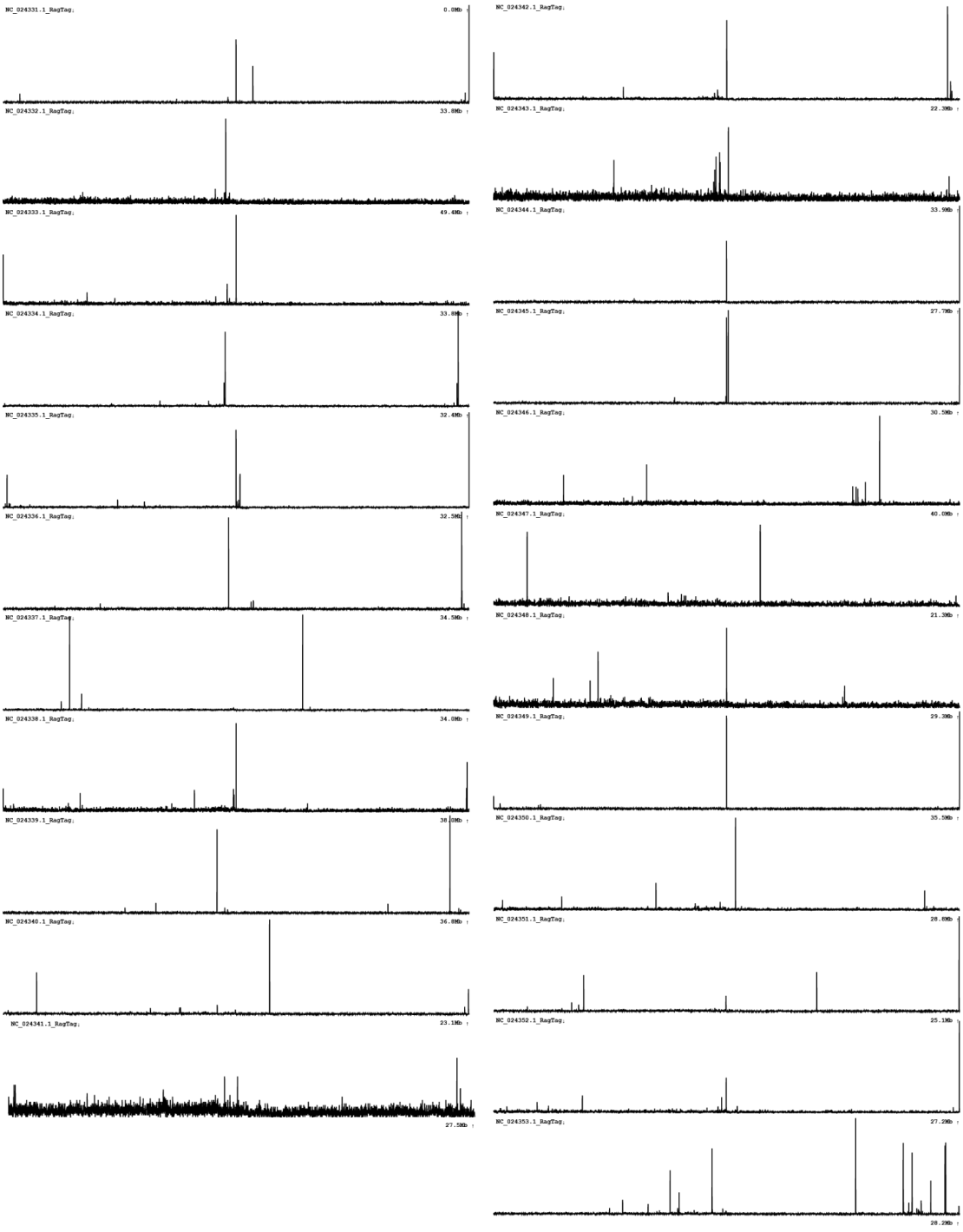
G. eurystoma (S)	168	686,863,250	5,110	4,088,471	30,179,505	39.04	546
G. sexradiata (NS)	77	669,122,782	5,104	8,689,906	30,375,227	39.03	148
G. sexradiata (S)	62	670,006,410	4,927	10,806,555	29,623,307	39.02	203

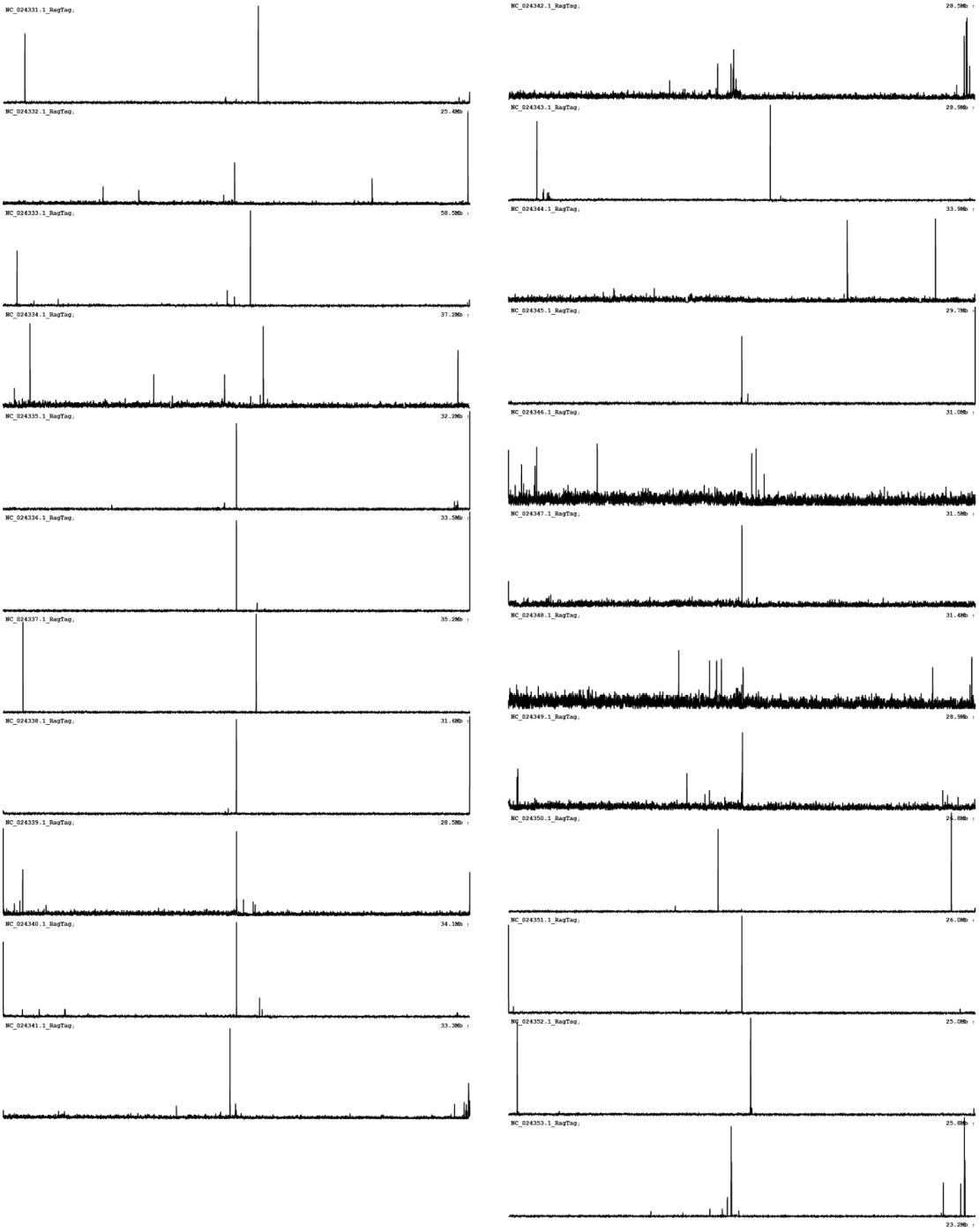
Supplemental Table 2.6: Gene annotation using the Braker3 pipeline. Total number of annotated genes, mean exon length (bp), and mean intron length (bp).

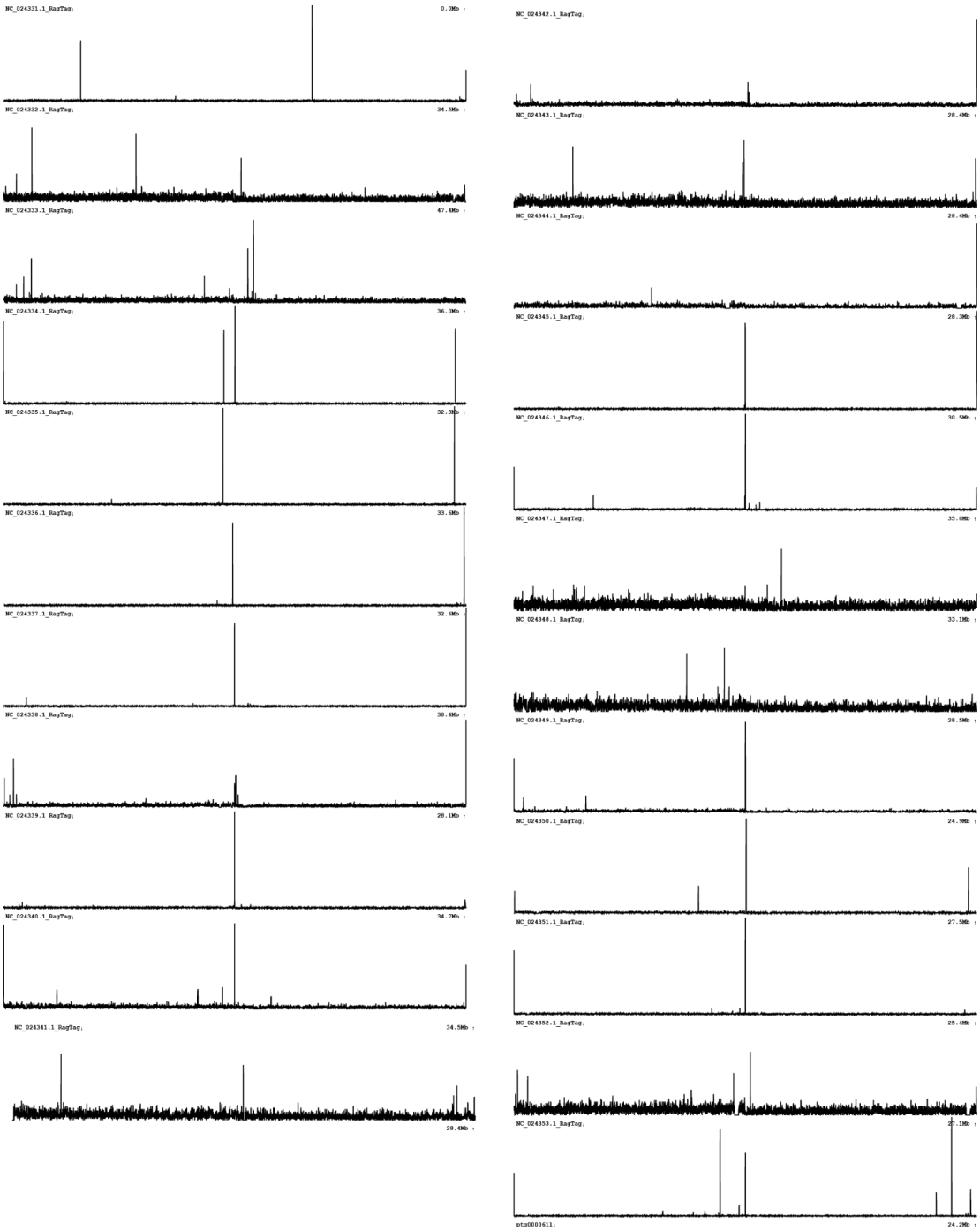
Species	**# of genes**	**Mean exon length (bp)**	**Mean intron length (bp)**
P. sulphuraria (S)	25,604	201	723
P. mexicana (S)	25,499	207	729
P. mexicana (NS)	23,700	173	762
G. eurystoma (S)	23,622	173	659
G. sexradiata (S)	23,579	178	677
G. sexradiata (NS)	23,217	174	628

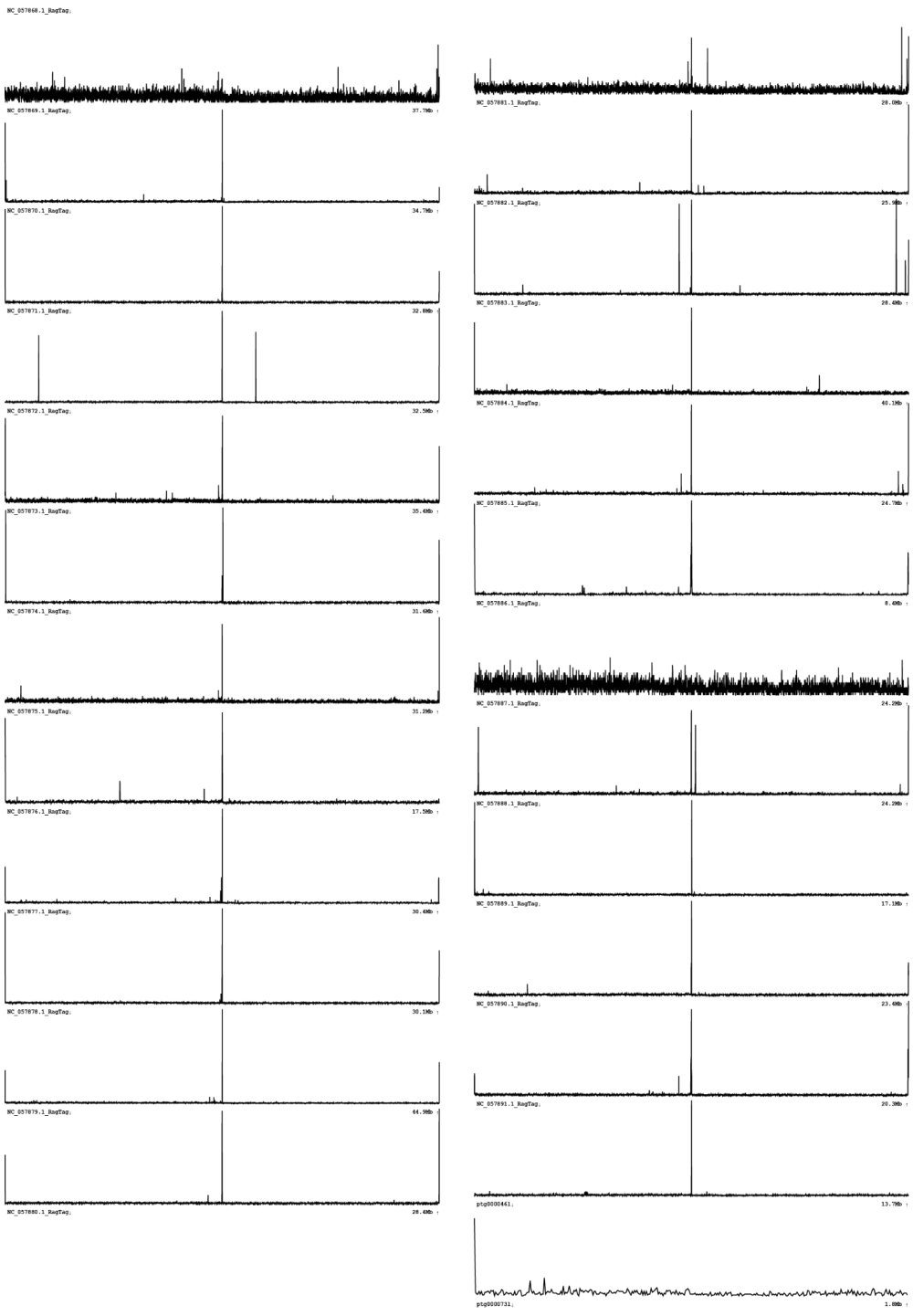


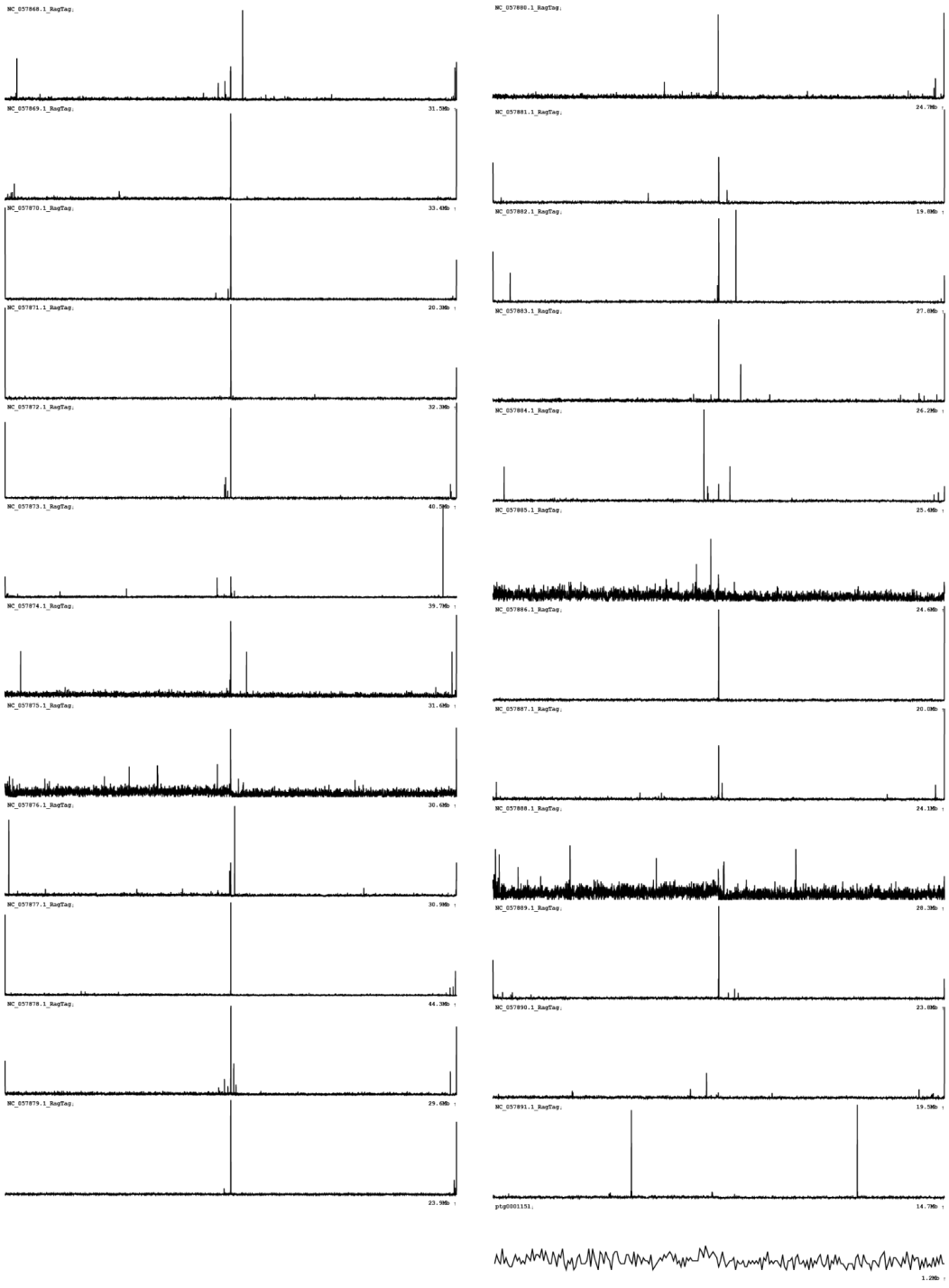
Supplemental Figure 2.1: K-mer spectrum and fitted models for (A) *P. sulphuraria* (S), (B) *P. mexicana* (S), (C) *G. eurystoma* (S) 2028, (D) *G. sexradiata* (NS), (E) *G. sexradiata* (S), and (F) *P. mexicana* (NS) Bonita. K-mer plots were generated using genomescope 2.0 with parameters set to K-mer length = 21, ploidy = 2, maximum K-mer coverage = -1, and average k-mer coverage for polyploid genome = -1.

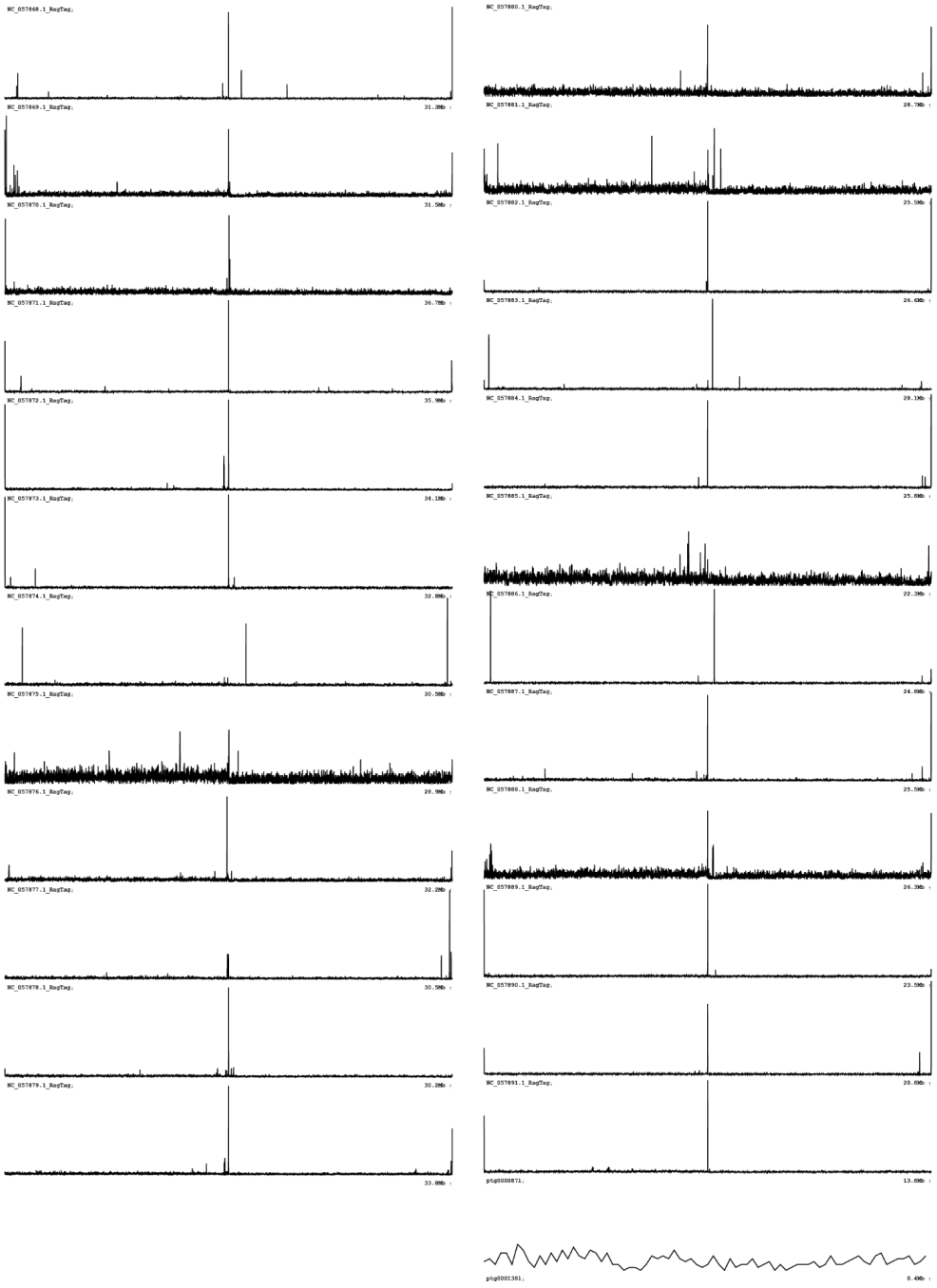

NC_024331.1_RagTag; 0.0Mb
NC_024332.1_RagTag; 33.0Mb
NC_024333.1_RagTag; 49.4Mb
NC_024334.1_RagTag; 33.8Mb
NC_024335.1_RagTag; 32.4Mb
NC_024336.1_RagTag; 32.5Mb
NC_024337.1_RagTag; 34.5Mb
NC_024338.1_RagTag; 34.0Mb
NC_024339.1_RagTag; 38.0Mb
NC_024340.1_RagTag; 36.8Mb
NC_024341.1_RagTag; 23.1Mb
27.5Mb
NC_024342.1_RagTag;
NC_024343.1_RagTag; 22.3Mb
NC_024344.1_RagTag; 33.9Mb
NC_024345.1_RagTag; 27.7Mb
NC_024346.1_RagTag; 30.5Mb
NC_024347.1_RagTag; 40.0Mb
NC_024348.1_RagTag; 21.3Mb
NC_024349.1_RagTag; 29.3Mb
NC_024350.1_RagTag; 35.5Mb
NC_024351.1_RagTag; 28.6Mb
NC_024352.1_RagTag; 25.1Mb
NC_024353.1_RagTag; 27.2Mb
28.2Mb


NC_024331.1_RagTag;
NC_024332.1_RagTag;
25.4Mb ↑
NC_024333.1_RagTag;
58.5Mb ↑
NC_024334.1_RagTag;
37.2Mb ↑
NC_024335.1_RagTag;
32.2Mb ↑
NC_024336.1_RagTag;
33.5Mb ↑
NC_024337.1_RagTag;
35.2Mb ↑
NC_024338.1_RagTag;
31.6Mb ↑
NC_024339.1_RagTag;
28.5Mb ↑
NC_024340.1_RagTag;
34.1Mb ↑
NC_024341.1_RagTag;
33.3Mb ↑
NC_024342.1_RagTag;
28.5Mb ↑
NC_024343.1_RagTag;
28.5Mb ↑
NC_024344.1_RagTag;
33.9Mb ↑
NC_024345.1_RagTag;
29.7Mb ↑
NC_024346.1_RagTag;
31.0Mb ↑
NC_024347.1_RagTag;
31.5Mb ↑
NC_024348.1_RagTag;
31.4Mb ↑
NC_024349.1_RagTag;
28.9Mb ↑
NC_024350.1_RagTag;
26.0Mb ↑
NC_024351.1_RagTag;
26.0Mb ↑
NC_024352.1_RagTag;
25.0Mb ↑
NC_024353.1_RagTag;
25.0Mb ↑
23.2Mb ↑

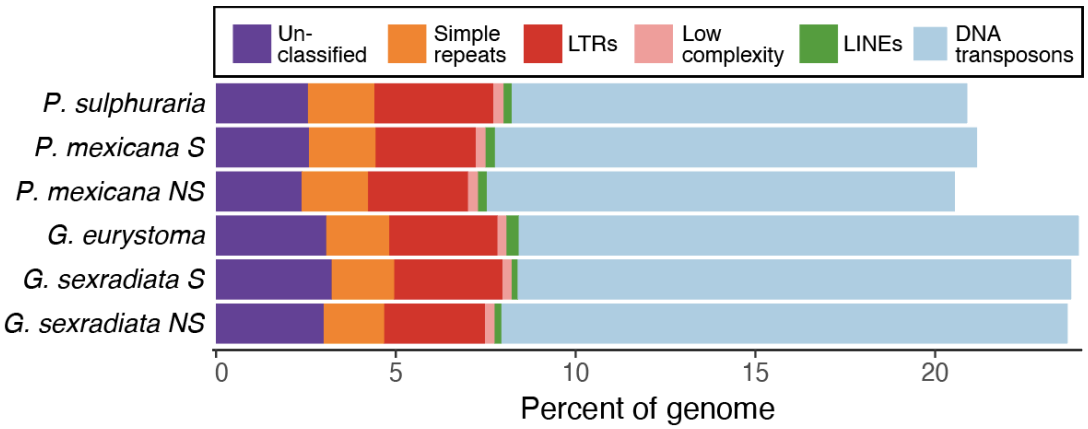

NC_024331.1_RagTag:
0.0Mb
NC_024332.1_RagTag:
34.5Mb
NC_024333.1_RagTag:
47.4Mb
NC_024334.1_RagTag:
36.0Mb
NC_024335.1_RagTag:
32.3Mb
NC_024336.1_RagTag:
33.6Mb
NC_024337.1_RagTag:
32.6Mb
NC_024338.1_RagTag:
30.4Mb
NC_024339.1_RagTag:
28.1Mb
NC_024340.1_RagTag:
34.7Mb
NC_024341.1_RagTag:
34.5Mb
28.4Mb
NC_024342.1_RagTag:
NC_024343.1_RagTag:
28.4Mb
NC_024344.1_RagTag:
28.4Mb
NC_024345.1_RagTag:
28.3Mb
NC_024346.1_RagTag:
30.5Mb
NC_024347.1_RagTag:
35.0Mb
NC_024348.1_RagTag:
33.1Mb
NC_024349.1_RagTag:
28.5Mb
NC_024350.1_RagTag:
24.9Mb
NC_024351.1_RagTag:
27.5Mb
NC_024352.1_RagTag:
25.4Mb
NC_024353.1_RagTag:
27.1Mb
ptg000011:
24.2Mb


NC_057868.1_RagTag;
NC_057869.1_RagTag;
37.7Mb
NC_057870.1_RagTag;
34.7Mb
NC_057871.1_RagTag;
32.8Mb
NC_057872.1_RagTag;
32.5Mb
NC_057873.1_RagTag;
35.4Mb
NC_057874.1_RagTag;
31.6Mb
NC_057875.1_RagTag;
31.2Mb
NC_057876.1_RagTag;
17.5Mb
NC_057877.1_RagTag;
30.6Mb
NC_057878.1_RagTag;
30.1Mb
NC_057879.1_RagTag;
64.9Mb
NC_057880.1_RagTag;
28.4Mb
NC_057881.1_RagTag;
28.0Mb
NC_057882.1_RagTag;
25.9Mb
NC_057883.1_RagTag;
28.4Mb
NC_057884.1_RagTag;
40.1Mb
NC_057885.1_RagTag;
24.7Mb
NC_057886.1_RagTag;
8.4Mb
NC_057887.1_RagTag;
24.2Mb
NC_057888.1_RagTag;
24.2Mb
NC_057889.1_RagTag;
17.1Mb
NC_057890.1_RagTag;
23.6Mb
NC_057891.1_RagTag;
20.3Mb
ptg000461;
13.7Mb
ptg000731;
1.6Mb


NC_057868.1_RagTag:
NC_057869.1_RagTag:
31.5Mb
NC_057870.1_RagTag:
33.4Mb
NC_057871.1_RagTag:
20.3Mb
NC_057872.1_RagTag:
32.3Mb
NC_057873.1_RagTag:
40.9Mb
NC_057874.1_RagTag:
39.7Mb
NC_057875.1_RagTag:
31.4Mb
NC_057876.1_RagTag:
30.6Mb
NC_057877.1_RagTag:
30.9Mb
NC_057878.1_RagTag:
44.3Mb
NC_057879.1_RagTag:
29.6Mb
23.9Mb
NC_057880.1_RagTag:
24.7Mb
NC_057881.1_RagTag:
19.6Mb
NC_057882.1_RagTag:
27.6Mb
NC_057883.1_RagTag:
26.2Mb
NC_057884.1_RagTag:
25.6Mb
NC_057885.1_RagTag:
24.6Mb
NC_057886.1_RagTag:
20.6Mb
NC_057887.1_RagTag:
24.1Mb
NC_057888.1_RagTag:
28.3Mb
NC_057889.1_RagTag:
23.6Mb
NC_057890.1_RagTag:
19.5Mb
NC_057891.1_RagTag:
ptg000115l:
14.7Mb
1.2Mb



Supplemental Figure 2.2: Annotation of telomeres in (1) P. mexicana NS, (2) P. mexicana S, (3) P. sulphuraria, (4) G. sexradiata NS, (5) G. sexradiata S, and (6) G. eurystoma.



Supplemental Figure 2.3: Repeat content by percent of genome.

Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears Genome Biology and Evolution. Ryan, K., De-Kayne, R., Davis, J., Arias-Rodriguez, L., Tobler, M., & Kelley, J. L. 2025. The dissertation author was the primary investigator and author of this paper.

Chapter 3 Pangenome-based analysis reveals ecotype-associated structural variants across multiple locally adapted Poeciliid fishes

Kara Ryan [1], Rishi De-Kayne [1,2], Lenin Arias-Rodriguez [3], Michael Tobler [4,5,6], Joanna L. Kelley [1]

1. Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95060, USA
2. Current address: Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA
3. División Académica de Ciencias Biológicas, Universidad Juárez Autónoma de Tabasco (UJAT), Villahermosa, México
4. Department of Biology, University of Missouri - St. Louis, St. Louis, MO, 63121, USA
5. Whitney R. Harris World Ecology Center, University of Missouri - St. Louis, St. Louis, MO, 63121, USA
6. WildCare Institute, Saint Louis Zoo, St. Louis, MO 63110, USA

Abstract

Extreme environments provide powerful natural laboratories for studying evolutionary convergence across taxa. While convergent phenotypes frequently evolve in response to similar selective pressures, consistent molecular signatures at the nucleotide level are often elusive, potentially due to overlooked genetic variants such as structural variants. Here, we leverage recent advances in long-read sequencing and pangenome approaches to investigate the role of structural variants in repeated adaptation to toxic springs rich in hydrogen sulfide (H_2S) across multiple independently evolved populations within the *Poecilia mexicana* species complex. We identified more than 99,000 structural variants over 50 base pairs in length, including insertions, deletions, duplications and inversions, using complementary graph- and read-based methods. We then integrated these with population genomic data and RNA-sequencing data to detect variants in highly differentiated regions

between ecotypes and assess their functional impacts. Structural variants in both coding and regulatory regions were enriched in genes involved in sulfide metabolism and ion transport, including *ethe1*, *slc13a1*, and *sqor*. Notably, we found ecotype-specific structural variants in promoter and intronic regions of key detoxification genes that exhibited elevated expression in H_2S-adapted populations. Our results highlight the critical and underappreciated role of structural variation in shaping adaptive phenotypes and emphasize the need to incorporate structural variants into models of genomic convergence.

Introduction

Extreme environments provide striking examples of convergent evolution across taxa, suggesting there are a limited number of solutions for evolution. These environments serve as natural experiments, offering powerful frameworks to investigate convergent evolution across biological scales (Rappaport & Oliverio, 2024). Strong selective pressures frequently drive the independent evolution of similar phenotypes—from molecular function to behavior—across lineages. For example, in subterranean habitats, the combination of darkness and nutrient scarcity results in a suite of convergent traits, such as loss of pigment and eyes, across distantly related taxa (Balart-García et al., 2023; Christiansen, 1961; Culver & Pipan, 2009; Derkarabetian et al., 2010). While some cases of phenotypic convergence are underpinned by repeated *de novo* mutations in the same genes or amino acid positions (e.g. *prestin* in echolocating bats and dolphins (Z. Liu et al., 2014)), growing

evidence suggests that many convergent traits lack consistent sequence-level signals (Corbett-Detig et al., 2020). Instead, convergence often occurs at higher levels, such as within gene regulatory networks or biological pathways or functions (Morales et al., 2024), even in extreme environments (F. Li et al., 2021; Xu et al., 2020).

Most genomic studies of convergent evolution have focused on single nucleotide polymorphisms (SNPs), potentially overlooking other forms of genetic variation. Structural variants—including insertions, deletions, duplications, and inversions—represent additional forms of genomic variation that are abundant genome-wide but have been historically difficult to confidently identify and genotype, despite their ability to affect gene regulation and protein function. For example, structural variants have been identified as a major driver of gene expression differences among human populations (Scott et al., 2021) and can shape functional differences in proteins, often through disordered conformations (Jilani et al., 2022; M. Lin et al., 2017). The absence of consistent sequence-level signals across taxa could reflect either the existence of multiple genomic solutions to similar selective pressures or a bias in investigating SNPs over structural variants. Therefore, incorporating structural variants into studies of genomic convergence is critical for a comprehensive understanding of the genomic architecture underpinning adaptive traits.

Structural variants have been recognized as important in evolutionary processes for over a century (McClintock, 1931; Sturtevant, 1913) and can directly influence adaptation and speciation (Berdan et al., 2024; Mérot, Oomen, et al., 2020;

L. Zhang et al., 2021). They underpin a broad range of phenotypes, including variation in coloration and color patterning (Joron et al., 2011; K.-W. Kim et al., 2022), morphology (Mérot, Llaurens, et al., 2020), and behavior (Küpper et al., 2016; Yan et al., 2020). In fishes, for example, copy number variation in the *fads2* gene is associated with dietary differences in stickleback (Ishikawa et al., 2021), and chromosomal inversions help maintain migratory and resident ecotypes in Atlantic cod (Kirubakaran et al., 2016). However, structural variants have historically been underrepresented in evolutionary genomic studies due to technical limitations in detecting them with short-read sequencing and alignment-based methods, especially in repetitive or structurally complex regions. The increasing accessibility of long-read sequencing has renewed interest in structural variants across a wider range of taxa (Bertolotti et al., 2020; Mérot, Oomen, et al., 2020; Mérot et al., 2023). Notably, genomic patterns of diversity inferred by structural variants can differ from those estimated from SNPs (Stuart et al., 2023), and structural variants' roles in evolutionary processes, such as adaptation and speciation, should be studied directly (Berdan et al., 2024; Feulner & De-Kayne, 2017). Despite increasing recognition of their evolutionary significance, few studies have comprehensively examined the role of structural variants in repeated In the livebearing fish family Poeciliidae, multiple lineages have independently adapted to toxic springs rich in hydrogen sulfide evolving convergent phenotypes in life history, morphology, physiology, and gene expression (Greenway et al., 2020; Kelley et al., 2016; Plath et al., 2013a; Riesch et al., 2014; Tobler et al., 2018, 2016). Despite convergence across levels of

organization, evidence for convergence at the gene and nucleotide level is limited (Greenway et al., 2020, 2024; Pfenninger et al., 2014). Furthermore, where signals of selection on similar regions have been detected, they appear to be primarily driven by selection on shared, standing variation and not independent *de novo* mutations (Ryan et al., 2023). Importantly, these previous studies have focused mainly on SNP variation, leaving the contribution of structural variants largely unexplored, although one study identified a few shared structural variants among sulfide-adapted populations of *Poecilia mexicana* (Brown et al. 2019). These were identified despite using short reads (100 base pair, bp), a highly fragmented reference genome, and a small sample size (N=1 per population), which limited the ability to detect variants and downstream analyses. Recent advances, including high-quality reference assemblies (Ryan et al., 2025) and improved long- and short-read sequencing technologies (Begum et al., 2021; Mahmoud et al., 2019), enable a more comprehensive analysis of structural variants, their potential role in repeated adaptation to sulfide springs in *P. mexicana*, and the extent to which adaptive structural variation represents a common feature of repeated adaptation in closely related lineages.

In this study, we investigate the role of structural variants in repeated adaptation to sulfide springs in *Poecilia mexicana*. Using Pacific Biosciences (PacBio) HiFi long read data from three individuals, two from sulfidic environments and one from nonsulfidic, we assembled pseudo-haplotype-resolved, contig-level genomes and constructed a *P. mexicana* pangenome graph. We then applied both graph- and long-

read based structural variant detection approaches to comprehensively identify variants 50 bp or longer. We then generated population-level, whole-genome resequencing data from five sulfidic and five closely related, nonsulfidic populations to assess variation in the frequency of structural variants among populations and identify highly differentiated SNPs.=. Finally, we integrated previously published RNA-sequencing (RNA-seq) data from (Kelley et al., 2016) and (Passow, Henpita, et al., 2017) to assess overlap between differentially expressed genes and structural variants in noncoding regions. By integrating SNPs, structural variants, and gene expression data from multiple sulfidic and nonsulfidic populations of *P. mexicana*, this work offers new insights into the genomic architecture of convergence in extreme environments and addresses broader questions about the predictability of evolution at the molecular level.

Materials & Methods

Focal populations

We analyzed ten populations of the *Poecilia mexicana* species complex distributed across four river drainages in the Río Grijalva basin of Chiapas and Tabasco, Mexico: Pichulcalo (Pich), Ixtapangajoya (Ixta), Puyacatengo (Puya), and Tacotalpa (Taco). These included five sulfide-spring (S) populations and five nonsulfidic (NS) populations. Sulfidic populations consisted of two *P. mexicana* populations (Taco S, Puya S), two *Poecilia sulphuraria* populations (Pich1 S, Pich2 S), and one *Poecilia thermalis* population (Ixta S). The closely related NS populations

included five *P. mexicana* populations (Taco NS, Puya NS, Ixta NS, Pich1 NS, Pich2 NS) (Table S3.1).

Identification of structural variants using long reads

We identified structural variants using previously generated PacBio HiFi reads (Ryan et al., 2025) from three focal populations: Taco S, Pich2 S, and Taco NS (Bioproject PRJNA984711). Briefly, DNA was extracted using the Circulomics Nanobind High Molecular Weight DNA extraction kit. Quantaty was assessed using the Qubit™ High Sensitivity DNA kit in triplicate, Quantus™ Fluorometer, and Nanodrop and quality was assessed using an Agilent™ Femto Pulse system and TapeStation. SMRTbell library preparation and sequencing was performed at UC Davis Genome Center on a single SMRT cell on a PacBio Revio sequencing platform.

Reads were mapped to the *P. mexicana* genome assembly (GCA_051168715.1 (Ryan et al., 2025)) using pbmm2 (https://github.com/PacificBiosciences/pbmm2), a wrapper for minimap2 (H. Li, 2018). Variants were called per sample using Sawfish v2.0.1 (Saunders et al., 2025) *discovery* followed by joint genotyping with *joint-call*. Variants were filtered to remove low-quality calls (QUAL < 10), excessive depth (12 time the gc-corrected haploid-depth estimate), redundant breakpoints already represented in the VCF, or conflicting multi-breakpoint genotypes flagged by sawfish. Additionally, we used Sniffles2 v2.6.1 (Smolka et al., 2024) to generate an intermediate sniffles (SNF) file

for each individual and performed multi-sample variant calling to generate a Variant Call Format (VCF) file. Variants were filtered using BCFtools v1.21 (Petr Danecek et al., 2021) to retain high quality variants (QUAL ≥ 30) and to exclude imprecise variants, homozygous reference calls, and break-end variants.

Pangenome generation, variant calling

To generate a pangenome graph representative of assembled haplotypes, we used HiFiasm 0.25.0-r726 (Cheng et al., 2021) to generate contig-level, pseudo-haplotype-resolved assemblies for each individual. FASTA headers were renamed to conform to the PanSN-spec: Pangenome Sequence Naming standard (https://github.com/pangenome/PanSN-spec). A reference-guided pangenome was constructed using the Minigraph-Cactus pipeline (Hickey et al., 2024). First, a structural variant graph (rGFA) was generated using Minigraph v0.20-r599 (H. Li et al., 2020), with the chromosome level *P. mexicana* NS assembly as the backbone. Haplotype assemblies were mapped back to graph using Minigraph, and a base-pair-level multiple genome alignment was generated using Cactus. The final pangenome graph was then prepared for downstream analysis using the VG toolkit (Garrison et al., 2018) *view* and *index* functions. Variants were extracted using VG *deconstruct* and the resulting VCF was filtered using VCFbub (*vcfbub: use variant nesting information to flter overlapping sites from vg deconstruct output*, n.d.) to remove nested sites unless nested under variants greater than 10 kilobases (kb). We then annotated variant types using Truvari v5.3.0 (English et al., 2022) *anno svinfo.*

Variant merging across calling methods

Each structural variant callset (Sniffles2, Sawfish and graph-derived) was normalized using BCFtools *norm* to left-align indels, correct reference alleles, and decompose multiallelic records. Redundant variants within each VCF were then collapsed using Truvari v5.3.0 *collapse* with parameters -r 500 -p 0.95 -P 0.95 -S 100000 -s 50 -k maxqual –passonly. To generate a nonredundant merged set of variants from long-read methods, we used Jasmine v1.1.5 (Kirsche et al., 2023). Duplications were temporarily converted to insertions (--preprocess_only –dups_to_ins) to reduce merging errors. Variants were merged and redundant variants were removed (max_dist=500 –nonlinear-distance). After merging, duplicates were restored to their original variant type and variants with unknown variant type were filtered out. To merge across read and graph-based approaches, we first normalized and removed redundance calls from the graph-based method using BCFtools and Truvari as described above. Then, variants were merged between graph and read based methods using SURVIVOR (Jeffares et al., 2017) v1.0.7.

Short-read library preparation and sequencing

Whole-genome sequencing Illumina libraries were prepared from 80 individuals (8 per population) previously sampled for capture sequencing (Ryan et al., 2023). Genomic DNA, stored at –80 °C for approximately six years, was thawed and quantified using Qubit the high-sensitivity DNA kit. Samples, chosen to optimize

DNA quality and quantity, were shipped overnight on ice to the Genomics and Cell Characterization Core Facility (GC3F) at the University of Oregon. At GC3F, 10 ng of DNA per sample was dried in a SpeedVac in a 384-well plate and 6 µl of a diluted Bead-Linked Transposase (Illumina) (1:60 in buffer: 10 mM Tris-HCL, 5 mM $MgCl_2$, 8% PEG) was added. The samples were vortexed to resuspend the DNA and incubated at 53 °C for 30 minutes. After the plate was spun down and placed on a magnet to gather beads, 5 µl of supernatant was removed. Then, 1 µl of barcoded Illumina Nextera adapter primers (IDT), 5 µl 2x Equinox Master-mix (Watchmaker Genomics), and 3 µl Nuclease Free Water (NEB) were added to each well. PCR was conducted as follows: 68 °C for 3 min, 98 °C for 3 min, 10 cycles of 98 °C for 45 s, 62 °C for 30 s, 68 °C for 2 min, followed by final extension at 68 °C for 1 min. After amplification, 20 µl of nuclease free water was added to each well and mixed. A 2 µl aliquot per sample was pooled and purified twice with 0.8x volume cleaning with Total Pure NGS beads (Omega Biosciences). The pooled library was run on Illumina MiSeq Nano PE150. Based on read count normalization, a rebalanced pool was sequences on two lanes of an Illumina NovaSeq S4 (PE150).

Identification of highly differentiated regions

Read quality was assessed using FastQC v0.12.1 (Andrews, 2010) and trimmed using fastp v0.32.2 (Chen et al., 2018). Trimmed reads were mapped to the *P. mexicana* reference genome (GCA_051168715.1) using bwa-mem2 (Vasimuddin et al., 2019). The resulting SAM files were converted to BAM and sorted and indexed

using Sambamba (Tarasov et al., 2015) v0.8.2. Single nucleotide polymorphisms (SNPs) were called using BCFtools *mpileup* (--skip-indels). BCFtools *view* was used to filter bialleleic SNPS (-m2 -M2 -v snps). The resulting SNPs were further filtered using VCFtools (P. Danecek et al., 2011) for quality (--minQ 30), missingness (--max-missing 0.9), depth (--min-meanDP 3 --max-meanDP 250) and minor allele count (--mac 8). Retained SNPs were then thinned to retain 1 SNP per 1 kb to reduce the effects of linkage disequilibrium. We then used pixy (Korunes & Samuk, 2021) to calculate the average Weir and Cockerham F_{ST} (Weir & Cockerham, 1984) for 10-kb windows across the genome. We converted windowed F_{ST} values to Z-score and defined outlier regions based on a 0.99 F_{ST} empirical threshold ($F_{ST} > 0.431$) and Z-score (> 2.996).

Population level genotyping of structural variants

To estimate allele frequencies of structural variants, we used GraphTyper2 v2.7.7 (Eggertsson et al., 2019) with the genotype_sv module to genotype the previously identified variants using the short-read BAM files generated above. To do so, we modified the merged VCF file to anchor the reference sequence to the first position of the insertion sequence for graph generation. For genotyping, the default minimum coverage filter (10X) was disabled to account for lower genome-wide sequencing depth across individuals. Aggregate genotype call per individual were determined by the highest genotype quality across different genotyping models, including both coverage and breakpoint-based models. The subsequent genotyped

VCFs were merged using BCFtools. Allele counts and frequencies were than calculated with BCFtools separately for sulfidic and nonsulfidic groups. We retained variants that were genotyped in at least 20 individuals per ecotype. Finally, allele frequency difference between sulfidic and nonsulfidic populations were calculated in R by subtracting the allele frequency in the non-sulfidic population from the frequency in the sulfidic population.

Analysis of structural variants and gene expression

We imported structural variants into R using VariantAnnotation v1.54.1 and GenomicRanges v.1.60.0. The GFF annotation file was imported, and features were extracted using rtracklayer v.1.68.0. Promoters were defined at the 1-kb window upstream of the transcription start site (strand-aware) and introns were defined as gene regions excluding exons using *setdiff* and *reduce*. Then, overlaps between highly differentiated regions (F_{ST} outlier windows), genomic features, and structural variants were identified using *findoverlaps()*.

Integrative Genomics Viewer (IGV) was used to visualize variants of interest in long and short read alignments. Gene Ontology (GO) enrichment was performed with ClusterProfiler v4.16.0, with GO terms considered significantly enriched if Benjamini-Hochberg adjusted p-value was less than 0.05.

We reanalyzed RNA-seq count data from gill tissue focusing on both field-collected samples (Kelley et al., 2016) as well as samples from a nonsulfidic common-garden H_2S exposure experiment, that included both laboratory-control and

H_2S exposure groups from two population pairs of sulfidic and nonsulfidic populations (Tacotalpa and Puyacatengo, n = 4–6 per group) (Passow, Henpita, et al., 2017). Using EdgeR v4.4.2, we filtered low-count genes, calculated normalization factors and estimated tagwise dispersion. We modeled differential expression using glmFit and tested contrasts between ecotypes using glmLRT. Differentially expressed genes (Benjamini-Hochberg adjusted p-value < 0.05) were then intersected with small variants (50–1,00 bp) present in promoters and intron regions.

Generation of gene trees

In order to explore variation among populations, we generated genes trees for both copies of *ethe1*. First, we downloaded the refseq proteins of both copies of *ethe1* from NCBI for *Gambusia affinis*, *Xiphophorus helleri, Poecilia reticulata*, *Poecilia mexicana*, *Fundulus heteroclitus*, and for single copy *ethe1* for *Danio rerio*. We then used mafft v.7.526 to align nucleotide gene sequences. From there, we used Iqtree2 v.2.2.6 to generate a maximum likelihood tree with 1,000 ultrafast bootstrap replicates. We then generated consensus sequences for both *ethe1* gene regions from the VCF file generated above for each *P. mexicana* individual. The headers in the resulting consensus sequences were renamed to include the sample name and then concatenated. These resulting fasta files were then aligned using mafft and a maximum likelihood tree with 1,000 ultrafast bootstrap replicates were generated for the *P. mexicana* individuals. Trees were visualized in R using ggtree v3.16.3, treeio v1.32.0, and ape v.5.8.1.

Results

Structural variant discovery with long reads

To characterize structural variation, we analyzed long-read sequencing data from three individuals representing two sulfidic populations (*P. sulphuraria* and *P. mexicana* S Taco) and one nonsulfidic population (*P. mexicana* NS Taco), with an estimated coverage of 22-32X and a median read length of 10,263 - 11,005 base pairs (bp) (Table S3.1; Ryan et al., 2025). These long-read datasets were used for complementary reference-based structural variant discovery (Sawfish and Sniffles2) and for pangenome-based variant detection. Then, these structural variants were merged and filtered to result in a set of variants for downstream population-level genotyping.

After filtering and collapsing redundant variants within each method, Sawfish detected 114,017 variants (53,417 deletions, 60,213 insertions, 185 inversions, and 202 duplications), while Sniffles2 identified 86,212 variants (40,524 deletions, 45,371 insertions, 260 inversions, 57 duplications) across three individuals with long read sequencing data (Ryan et al., 2025). Merging these variants and reducing redundancy in calls resulted in 93,759 unique variants (43,586 deletions, 49,599 insertions, 340 inversions and 234 duplications). Pangenome generation and structural variant calling was performed using haplotype-resolved assemblies (113–405 haplotypes per individual; Table S3.3) identified 69,409 variants (33,198 deletions, 35,911 insertions, and 300 variants of unknown type, likely inversions and break-end

variants). Merging graph-based variants calls with the long-read based call set resulted in 99,016 variants (45,581 deletions, 52,893 insertions, 308 inversions, and 234 duplications) (Figure S3.1A). Variant length ranged from 50 bp to over 10 megabases (Mb), with distinct distributions of sizes observed across SV types (Figure S3.1B). Median length was 244 bp for deletions, 383 bp for insertions, 2,699 bp for inversions, and 18,788 bp for duplications.

Genotyping of structural variants and variation in allele frequencies between ecotypes

To quantify allele frequency variation between habitats, we generated short read sequencing data from eight individuals per population from 10 populations of the *Poecilia mexicana* species complex (five sulfidic and five nonsulfidic). Whole-genome short-read sequencing of 80 individuals across ten populations produced an average of 26.8 million reads per individual across two sequencing lanes (range: 16.9–56.6 million reads) and a mean estimated coverage of 11.71X (9.44 – 21.19X; Table S3.2). Using these data, we genotyped the structural variants discovered in three individuals with long reads to explore the difference in allele frequencies between ecotypes. After filtering, we were able to successfully genotype and analyze 39,496 structural variants, ranging from 50 bp to greater than 5 Mb in size, comprising of 35,068 deletions, 2423 insertions, 2995 duplications, and ten inversions. We explored the most differentiated structural variants between ecotypes (top 1% difference in allele frequency) (Figure S3.2, Table S3.4). These consisted of

34 deletions, two duplications, and three insertions, consistent with our reduced ability to genotype insertions (Mahmoud et al., 2019; Sedlazeck et al., 2018). The mean variant size of these highly differentiated variants was 781 bp, and the median was 264 bp. Thirty-one of these variants overlapped with gene regions (Table S3.4). The highest difference in allele frequencies between ecotypes was a 419-bp deletion in an intron of *ethe1* (AF_S=1.0, AF_{NS} = 0.15).

Single nucleotide polymorphism detection and identification of highly differentiated regions

To identify highly differentiated regions between ecotypes independent of structural variants, we identified genome-wide single nucleotide polymorphisms (SNPs). After filtering, we retained 578,183 SNPs. Genome-wide average F_{ST} between sulfide and nonsulfidic populations was 0.136. We identified highly differentiated genomic windows (10 kb, F_{ST} > 0.431, Z-score > 2.996) as outliers that are putatively under selection (Figure S3.3). Outlier windows overlapped with 823 genes (668 unique genes, Table S3.5).

Analysis of large and mid-sized structural variants

To identify candidate structural variants potentially underlying adaptation to sulfide springs, we first examined large variants (>50 kb) due to their ability to influence large genomic regions and potentially many genes at once. To avoid potentially spurious SV calls, we excluded variants exceeding 5 Mb. Large variants consisted of

54 deletions, 70 insertions, 32 inversions, and 8 duplications (Figure S3.4). Two noteworthy large variants overlapped highly differentiated region: a deletion on scaffold 12 and an inversion on scaffold 6 that span extensive regions of high genomic differentiation.

The 3.6-Mb deletion on scaffold 12 affected 140 unique genes, including N-acetylglucosamine-6-sulfatase (*gns*) and 2-oxoglutarate dehydrogenase (*ogdh*), but did not have significant enrichment of any biological processes or functions. This variant was at low frequency in both sulfidic (0.0375) and nonsulfidic (0.0625) populations, suggesting that this is likely a rare variant and not involved in adaptation, despite being in a highly differentiated region (Table S3.6). The inversion on scaffold 6 encompassed 33 unique genes, 27 of which had functional annotation information (Figure 3.1A). These genes included two copies of *slc13a1*, a Na^+-sulfate cotransporter gene involved in sulfate homeostasis, and *hyal4* (hyaluronidase-4), which modifies the structure of chondroitin sulphate. Gene Ontology (GO) analysis of the genes in the inversion on scaffold 6 identified significant enrichment for biological processes and molecular functions related to hyaluronan catabolism (GO:0030214, GO:0030212) and glycosaminoglycan metabolism and catabolism (GO:0006027, GO:0030203, GO:0006026, GO:0006022) (Figure 3.1B). However, we were unable to genotype this variant across individuals to assess its distribution among ecotypes, given variation in depth across short reads.

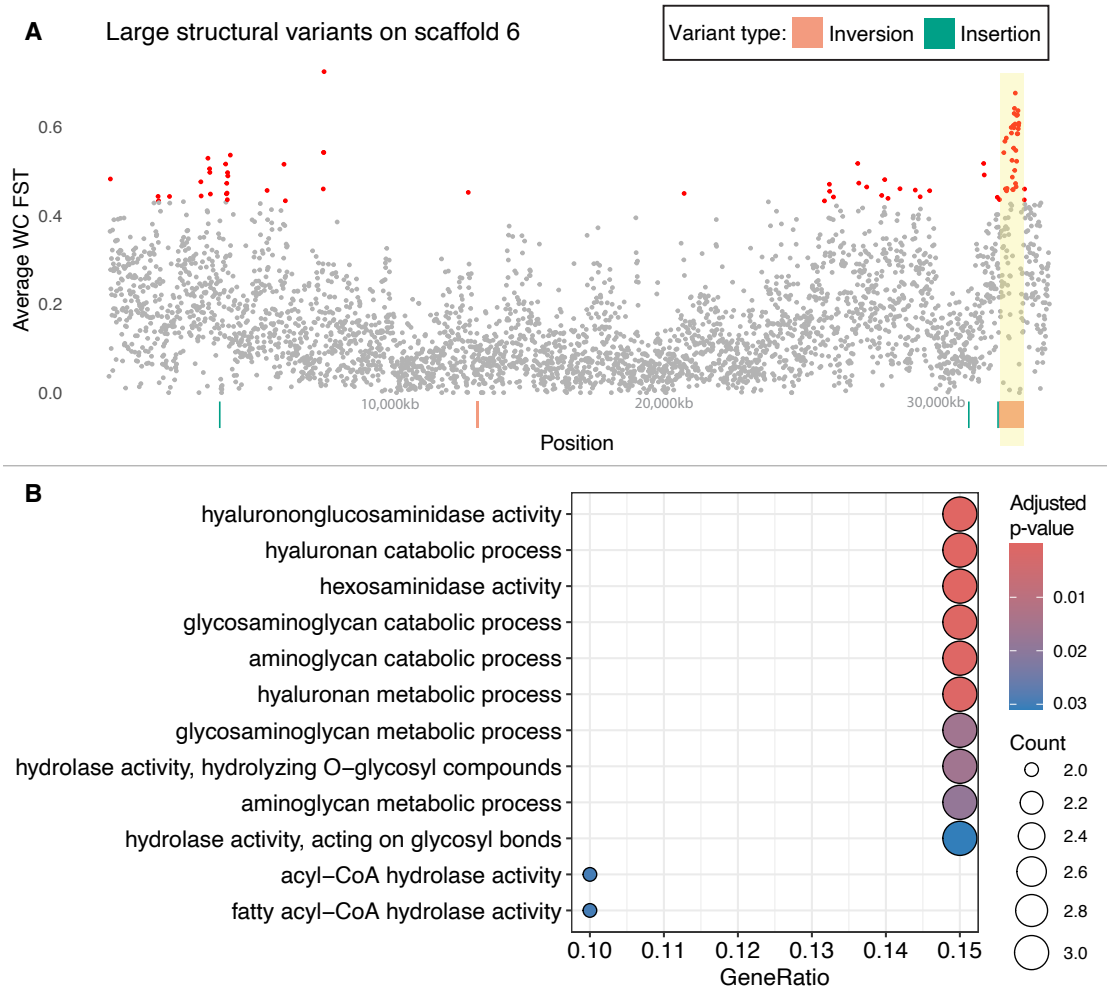


Figure 3.1: Large inversion highly differentiated between ecotypes. (A) Average Weir and Cockerham FST (WC FST) between all sulfidic and nonsulfidic individuals in 10-kb windows. Outliers with Z score > 2.996 and a FST > 0.431 (0.99 empirical cutoff) are colored red. Structural variants greater than 50 kb are plotted below the axis. Inversions are plotted in orange and insertions are plotted in green. (B) Gene Ontology (GO) enrichment analysis of genes in the inversion highlighted in yellow in (A). All significantly enriched (Benjamini Hochberg corrected p-value < 0.05) GO terms are plotted.

We identified 1,117 unique, medium sized variants (10–50kb) consisting of 492 deletions, 560 insertions, 60 duplications, and 5 inversions. Scaffold 18 was a hotspot for variants of this size (118 variants), whereas all other scaffolds contained

fewer than 76 variants each (Figure S3.5). Of these medium sized variants, 25 overlapped with highly differentiated region regions, harboring 17 annotated genes. Notably, genes involved in oxidative phosphorylation (OxPhos) complex assembly were detected: NADH dehydrogenase [ubiquinone] 1 alpha subcomplex assembly factor 5 (*ndufaf5*) (involved in Complex I assembly) and protein PET117 cytochrome c oxidasecChaperone (*pet117*). However, GO enrichment analysis of these genes did not yield significant functional enrichment.

Two variants were at high frequency in sulfidic populations (AF > 0.94) and low to mid frequency in nonsulfidic populations (AF < 0.34): a 13.8-kb deletion of LOC106932973 (CUB and zona pellucida-like domain-containing protein 1) and a 13.1-kb deletion that did not affect a gene (Table S3.4). Additionally, there were three variants at low frequency in sulfidic populations (< 0.15) and mid to high frequency in nonsulfidic populations (> 0.56): a 11.7-kb deletion that affects *slc22a13*, a 19.6-kb deletion that affects dihydropyrimidinase-related protein 2, rac GTPase-activating protein 1, and protein phosphatase 1 regulatory subunit 3C-B, and a 13.9-kb deletion that did not influence any genes.

Next, we analyzed variants 1–10 kb in size. These variants consisted of 8,344 deletions, 9,912 insertions, and 102 duplications. Although most variants occurred in intergenic regions, 124 genes overlapped with variants of this size, of which 94 genes had functional annotations. Some of these genes occurred in highly differentiated regions, including candidate genes directly involved in sulfide metabolism. These genes included persulfide dioxygenase (*ethe1a* and *ethe1b*) and sulfide: quinone

oxidoreductase (*sqor*). These genes were significantly enriched in a biological process (sulfur compound metabolic process; GO:0006790) and molecular functions secondary active transmembrane transporter activity (GO:0015291), sodium ion transmembrane transporter activity (GO:0015081), solute: monoatomic cation symporter activity (GO:0015294), hexosaminidase activity (GO:0015929) and solute: sodium symporter activity (GO:0015370) (Figure 3.2A).

The 2.3-kb deletion in *sqor* shows significant difference in AF between sulfidic (AF=0.025) and nonsulfidic populations (AF=0.590). An additional 21 variants had significantly higher AF in sulfidic populations than nonsulfidic populations, of which 7 variants affected 8 genes. Finally, 40 variants had significantly lower AF in sulfidic populations, which interacted with 17 genes (Table S3.4).

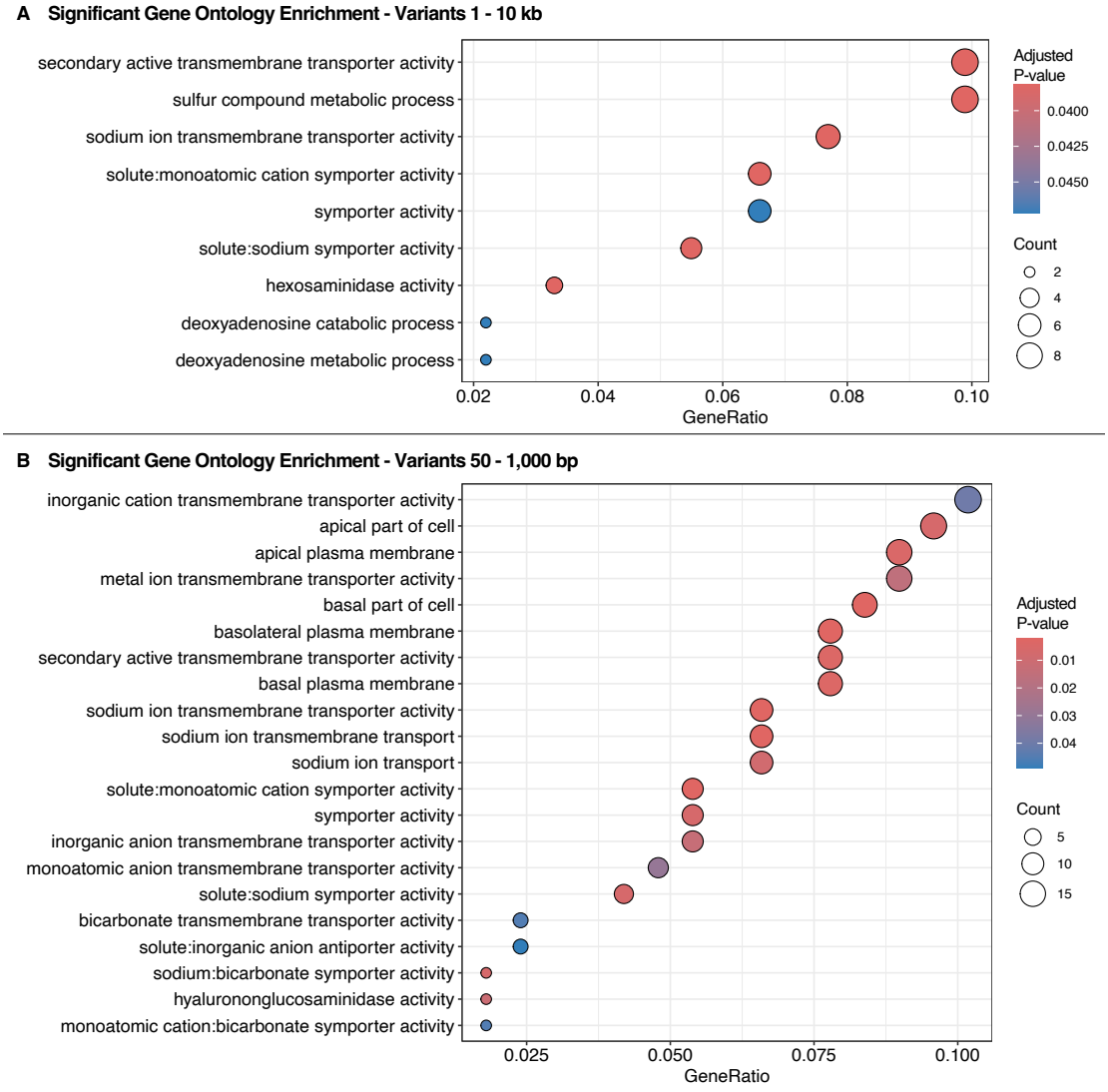


Figure 3.2: Gene ontology (GO) terms that are significantly enriched in variants (A) 1–10 kb or (B) 50–1,000 bp and in regions highly differentiated between ecotypes. All significant terms (Benjamini Hochberg adjusted p-value < 0.05) were plotted.

Analysis of variants shorter than 1kb

Most structural variants were 50 bp to 1 kb in size, with 33,995 deletions, 33,437 insertions and 99 inversions called in this size range. Of these ~68,000 small variants, 29,284 overlapped with annotated gene regions. These variants overlapped

with 181 genes that were highly differentiated between ecotypes. These included multiple genes that localize to the mitochondria, including 2-oxoglutarate dehydrogenase (*oghd*), persulfide dioxygenase (*ethe1*), sulfide:quinone oxidoreductase (*sqor*), pyruvate carboxylase (*pc*) and alpha-aminoadipic semialdehyde synthase (*aass*). These variants were significantly enriched in genes involved in a broad range of transmembrane transport activities, including sodium ion transport (GO:0015081, GO:0035725, GO:0006814), metal ion transport (GO:0046873), and bicarbonate or inorganic anion exchange (GO:0015106, GO:0005452) (Figure 3.2B). These transporters spanned diverse molecular functions such as symporters, antiporters, and secondary active transporters (e.g., GO:0015291, GO:0015294, GO:0015103), and were enriched in localization to specialized regions of the plasma membrane (GO:0016323, GO:0009925, GO:0016324).

Given the potential specific functional and regulatory impacts of these variants, we investigated variants in this size class in more detail. Specifically, we focused on those in coding, putative promoter, and intronic regions. Among variants that overlapped with genes, 2,558 variants overlapped with exons, with 1,924 variants affecting protein-coding sequences. These coding-region variants included 996 deletions, 927 insertions, and one inversion. The mean variant length was 224 bp and the median was 129 bp. Interestingly, these exon-overlapping variants were more frequently in-frame (44.4% of variants) than expected based on the genome-wide distribution for this size class (35.9%) (Figure S3.6).

Associations of noncoding structural variants and differentially expressed genes

Regulatory changes have been shown to underlie adaptive divergence between sulfidic and nonsulfidic populations (Brown et al., 2018; Greenway et al., 2020; Kelley et al., 2016; Passow, Brown, et al., 2017; Passow, Henpita, et al., 2017). To investigate how structural variants in non-coding regions, specifically introns and promoters, relate to differential expression, we reanalyzed RNA-seq data generated by Kelley (2016) and Passow (2017).

Most small structural variants overlapping genes were in introns (27,191 variants across 10,196 genes), suggesting intronic variants are common. Among these, 361 genes also fell within highly differentiated regions. Additionally, we identified 4,650 structural variants in putative promoter regions. Of variants in putative promoters, 63 overlapped with highly differentiated regions. In the Puyacatengo drainage, 275 genes were significantly upregulated and 491 downregulated ($logFC > 1$ or < -1, BH-adjusted p-value < 0.05), while the Tacotalpa drainage showed 774 upregulated and 519 downregulated genes. Of these, 115 upregulated and 152 downregulated genes were shared between drainages. Forty-eight of the shared upregulated and 70 of the shared downregulated genes contain at least one intronic structural variant, with many containing multiple variants and 7 shared up- and 19 downregulated genes contain promoter variants (Table S3.6).

Structural Variation and Expression of Candidate Genes Encoding sulfide detoxification enzyme ETHE1

We specifically explored differential expression in both gene copies of *ethe1* due to their candidate role in sulfide detoxification, their observed structural variants in promoter and intronic regions, and the highly differentiated variation between ecotypes (Figure 3.3A). Gene *ethe1a* harbored two promoter variants (5,624-bp insertion and 64-bp deletion) and an intronic deletion (419 bp) between exons 6 and 7 that are shared between Taco S and *P. sulphuraria* and not *P. mexicana* NS, and a 1.2-kb insertion between exons 1 and 2 unique to *P. sulphuraria* (Figure 3.3A). Of these, the 419-bp deletion is fixed in sulfidic individuals and at low frequency in nonsulfidic individuals (Figure 3.3B), with genotypes highly concordant between both breakpoint and coverage based genotyping models. In contrast, structural variation in *ethe1b* does not sort by ecotype across the genes, as the 1.2-kb insertion is unique to *P. sulphuraria*. Expression data from wild-caught gill tissues indicated substantial upregulation of both copies of *ethe1* in sulfide-adapted populations relative to nonsulfidic populations. Specifically, in Tacotalpa, expression was 16-fold higher for *ethe1a* and 14-fold higher for *ethe1b;* similarly, in Puyacatengo, *ethe1a* expression was 13-fold and *ethe1b* was 16-fold higher (Figure 3.3C). Moreover, differential expression between the gene copies is apparent: *ethe1a* exhibited consistently higher expression than *ethe1b* in sulfide-adapted populations (3.1-fold higher in Tacotalpa and 3.6-fold higher in Puyacatengo; Figure 3.3C). These expression differences between ecotypes and gene copies, although at a different magnitude, persisted under both under common garden nonsulfidic conditions and after experimental H_2S exposure (Figure 3.3D).

To characterize how variation in these genes is distributed among populations, we generated gene trees using data from the 80 individuals analyzed above along with publicly available sequences from Poeciliid fishes (*Gambusia affinis*, *Xiphophorus helleri,* and *Poecilia reticulata*), an outgroup that also shares the ancestral *ethe1* duplication (*Fundulus heteroclitus*), and *Danio rerio*, which contains a single copy *ethe1* (Supplemental Figure S3.7). Notably, both gene trees deviate from the established species phylogeny (Ryan et al 2023, Greenway et al 2020, Brown et al 2019). In both copies of *ethe1* sequences from sulfide adapted *P. mexicana* individuals cluster more closely with *P. sulphuraria* and *P. thermalis* (Pich and Ixta) than with conspecific or sister taxa (Supplemental Figure S3.7). But this clustering differs between paralogs: *ethe1a* clusters all sulfidic populations together, with very little variation among individuals, whereas in *ethe1b* only the sulfidic Puya population clusters with *P. sulphuraria* and *P. thermalis* (Pich and Ixta), while the Taco sulfidic population does not.

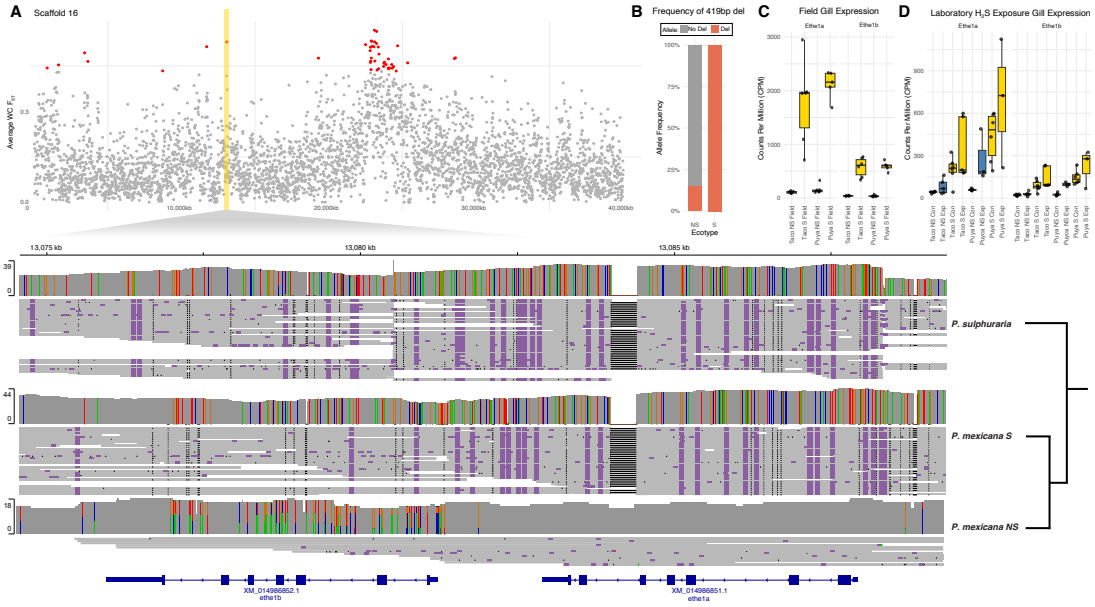


Figure 3.3: Sequence and gene expression variation in two copies of sulfide processing gene persulfide dioxygenase (*ethe1)*. (A) F_{ST} in 10-kb windows of scaffold 16. Genome wide 99% outlier windows are marked in red. Integrative Genomics Viewer (IGV) screenshot of long reads mapped to the *P. mexicana* NS reference genome and cladogram to show the relationship among populations. The top panel of each row is depth with overlayed colored bars representing SNP locations. In the bottom panel of each row, small insertions are labeled with purple and black lines represent deletions. Structure of *ethe1* copies from genome annotation represented below. Additional 1.2-kb insertion unique to *P. sulphuraria* at 13,080,525 also visible. (B) Frequency of 419-bp deletion in sulfidic and nonsulfidic individuals. (C) Gene expression in the gill of both copies of *ethe1* in field caught individuals from two population pairs of sulfidic and nonsulfidic populations. Data generated in (Kelley et al., 2016). (D) Gene expression in the gill of both copies of *ethe1* from individuals exposed to hydrogen sulfide in common garden conditions from two population pairs of sulfidic and nonsulfidic populations. Data originally generated in (Passow, Henpita, et al., 2017)

Discussion

We explored the role of structural variants in repeated adaptation to hydrogen-sulfide-rich springs in independent lineages of *Poecilia mexicana*. Through an integrative approach combining long-read sequencing, pangenome assembly, and analysis of differentiation with short-reads, we identified 99,016 structural variants

across multiple individuals in the *P. mexicana* species complex. Numerous variants were present in regions that were highly differentiated between sulfidic and nonsulfidic ecotypes with some sorting by ecotype and not taxonomic relationship. These variants, distributed across coding, intronic, promoter and intergenic regions, highlight how structural variants could contribute to adaptation, extending beyond the traditionally emphasized single nucleotide polymorphisms (SNPs). But, given these discoveries were made using three genomes across the species complex, they likely underrepresent the full structural variant diversity in the species. For example, recent work on House Finches (*Haemorhous mexicanus*) that utilized a pangenome approach to identify structural variants found over 800,000 structural variants across 16 individuals and an outgroup, spanning ~13 million years of evolution across large geographic space (Fang and Edwards 2024). Future work to expand this pangenome to include more individuals as long-read data become available could lead to new insights into independent and shared variation underpinning sulfide tolerance.

Noncoding structural variants associated with regulatory differences

Structural variants can substantially influence phenotypic evolution without influencing protein function by altering regulatory elements and thereby affecting gene expression (Chaisson et al., 2015; Harewood et al., 2012; Scott et al., 2021; Shaul, 2017). These regulatory differences may facilitate adaptive divergence under strong environmental selective pressures, such as those found in H_2S-rich springs. Notably, we discovered multiple promoters and intronic variants in previously

identified differentially expressed genes that have been implicated in adaptation (Kelley et al., 2016; Passow, Brown, et al., 2017; Passow, Henpita, et al., 2017). We discovered ecotype-specific variants in regulatory regions of genes central to H_2Sdetoxification. Specifically, a 64-bp deletion in the promoter and a 419-bp deletion in an intron of *ethe1a* correlated with significantly elevated and ecotype-specific gene expression patterns in gill tissues of sulfidic populations compared to nonsulfidic counterparts (Figure 3.3A). In addition to coding neofunctionalization, regulatory neofunctionalization has been proposed as a major source of evolutionary innovation (Ohno 1970; Moore and Purugganan 2005) due to the potential for protein expression in a novel context (Hughes et al. 2014). Such consistent differenced across independently adapted population pairs suggests regulatory neofunctionalization, potentially facilitated by the observed structural variants. In particular, the 419-bp deletion in *ethe1a* near the intron-exon boundary may influence mRNA processing or stability, contributing to its consistently elevated expression. The presence of identical structural variants in both *P. sulphuraria* and H_2S -adapted *P. mexicana* (Taco S), but not in closely related nonsulfidic *P. mexicana* populations (Figure 3.3A), further suggests a shared origin—likely via gene flow or incomplete lineage sorting—rather than independent parallel mutations in each lineage. Future investigation into exon usage between ecotypes, potentially using long-read RNA-seq, could shed light into the evolution of these paralogs, and how these variants influence both expression and splicing of these two copies of this H_2S-detoxification gene.

Structural variants affecting transporter and sulfide metabolism genes

We identified variants in transporter and sulfur-processing genes (Figure 3.1 & 3.2), including a highly differentiated large genomic inversion involving the sodium-sulfate cotransporter gene *slc13a1.* Given growing evidence that inversions can maintain adaptive variation and drive ecological divergence (Thorstensen et al., 2022; Zong et al., 2020), this inversion could assist in maintaining linked adaptive variation in the face of low levels of gene flow occurring between ecotypes in at least some populations (Brown et al., 2018; Greenway et al., 2024). Additional variants exist in these and related transport genes, including a 195-bp insertion within exon 1 of sulfate transporter *slc26a*2, a 60-bp deletion in exon 7 of *slc13a1*, and a 143-bp insertion in carbohydrate sulfotransferase 6 (*chst*). Future studies using rapidly advancing protein prediction models (Jumper et al., 2021) could shed light on the functional impacts of these, and other, coding variants in highly differentiated regions. Moreover, GO enrichment analysis revealed significant overrepresentation of many transport-related biological processes among genes harboring structural variants within highly differentiated genomic regions (Figure 3.2B). These results collectively underscore a complex genomic response encompassing both regulatory and protein-coding modifications to environmental H_2S exposure.

The presence of variants in both sulfur processing genes and transporter genes leads to new hypotheses about the mechanisms underlying repeated adaptation to sulfide springs. H_2S is produced endogenously in the body at low levels and is

important in a variety of biological processes, ranging from cell signaling to proper neurological and cardiac function (Cao et al., 2019; García-Bereguiaín et al., 2008; Kabil et al., 2014; L. Li et al., 2011; Peers et al., 2012). There are a variety of conserved mechanisms for processing surplus H_2S and related compounds to maintain tight regulation, given its inhibition of cytochrome c oxidase and mitochondrial respiration even at low concentrations (Cooper & Brown, 2008; Jiang et al., 2016; Szabo, 2018). These H_2S regulatory mechanisms, via sulfide-quinone oxidoreductase (*sqor*) (Landry et al., 2017) or superoxide dismutase [Cu-Zn] (*sod1*) (Switzer et al., 2023) in the mitochondria, produce a variety of downstream byproducts, such as sulfate and thiosulfate. Given evidence for differential SQR activity between ecotypes (Greenway et al., 2020) as well as differential expression of multiple sulfur processing genes between ecotypes (A. P. Brown et al., 2018; Greenway et al., 2020; Kelley et al., 2016; Passow, Brown, et al., 2017; Passow, Henpita, et al., 2017), it is likely that in addition to oxidizing H_2S to prevent the disruption of the electrochemical gradient, cells need to excrete the byproducts of this process to maintain homeostasis. Given that these byproducts need transporters to move across membranes, the enrichment in variants associated with sulfate transport indicates that modification of excretory pathways complements H_2S oxidization in mediating elevated tolerance to environmental H_2S. The presence of ecotype-specific structural variants occurring in transport genes parallels previous findings, where *slc26a5* was duplicated and upregulated in sulfidic ecotypes (A. P. Brown et al., 2019). Interestingly, Brown (2019) also found a duplication in *slc13a1*, but they did not find

the smaller deletion within the gene identified in this study. Future physiological work assessing not only the function of detoxification related proteins, but also of transport of byproducts across the membrane could shed light into the complexity of adaptation to sulfide springs.

Broader evolutionary significance of structural variation

Identifying genetic variants underpinning adaptive phenotypes remains challenging due to complex genotype-to-phenotype relationships. Nonetheless, our findings clearly illustrate that structural variants are present, and often common, in genomic regions putatively under selection (Ryan et al., 2023). These variants could play an integral role in repeated adaptation to H_2S-rich springs. Our results complement prior studies highlighting adaptive divergence in gene expression and physiological H_2S tolerance between ecotypes (Brown et al., 2018; Greenway et al., 2020; Kelley et al., 2016; Passow, Brown, et al., 2017; Perry et al., 2024; Tobler et al., 2016), suggesting structural genomic variants could act alongside SNP-driven mechanisms by modulating gene expression, splicing, and/or protein function.

Importantly, the repeated identification of identical structural variants across independently evolved sulfidic populations raises intriguing questions regarding their evolutionary origin. Given the improbability of identical mutations independently arising multiple times, our results favor the persistence of ancestral standing genetic variation or adaptive introgression mediated through gene flow among closely related lineages, suggested but not rigorously tested across populations in previous studies (

Brown et al., 2018; Greenway et al., 2020; Ryan et al., 2023). Future population genomic studies will be essential for distinguishing these scenarios, further elucidating evolutionary processes underlying adaptive genomic variation. This study provides evidence that structural variants could contribute to the repeated ecological adaptation, emphasizing the importance of integrating structural variants into future studies of local adaptation. These new insights pave the way to novel investigations into how structural variants influence both protein structure and gene regulation and drive adaptive divergence across natural populations.

Supplemental material

Supplemental Table 3.1: Focal populations included in this study including the species name, site name, site ID used in the manuscript, the drainage of origin, the habitat type, either sulfidic (S) or nonsulfidic springs (NS), latitude and longitude of the population, the estimated HiFi coverage and median read length for species with long reads, and the sample size of Illumina sequencing per population used in this study.

Species	**Site name**	**Site ID**	**Drainage**	**Habitat**	**Lat & Lon**	**Est. HiFi cov**	**Median HiFi read length**	**Illumina sample size**
P. sulphuraria	Baños del Azufre	Pich 1 S	Pichucalco	S	17.55, -92.99			8
P. mexicana	Vet Station	Pich 1 NS	Pichucalco	NS	17.55, -93.00			8
P. sulphuraria	La Gloria	Pich 2 S	Pichucalco	S	17.53, -93.01	32X	10,433	8
P. mexicana	Rosita	Pich 2 NS	Pichucalco	NS	17.48 -93.10			8
P. thermalis	La Esperanza	Ixta S	Ixtapangajoya	S	17.51, -92.98			8
P. mexicana	Rio Ixtapangajoya	Ixta NS	Ixtapangajoya	NS	17.49, -92.99			8

P. mexicana	La Lluvia, small spring	Puya S	Puyacatengo	S	17.46, -92.89			8
P. mexicana	Vicente Guerrero	Puya NS	Puyacatengo	NS	17.51, -92.91			8
P. mexicana	El Azufre I	Taco S	Tacotalpa	S	17.44, -92.77	34X	10,263	8
P. mexicana	Bonita	Taco NS	Tacotalpa	NS	17.42, -92.75	22X	11,005	8

Supplemental Table 3.2: Sample IDs from BioProject PRJNA1128810 used in this study, including estimated sequence coverage. Provided as separate Excel file "Pmex_SV_table_S2.xlsx"

Supplemental Table 3.1: Haplotype assemblies generated using HiFiasm including the species, site ID, habitat type, number of contigs assembled and the N50 for each haplotype assembly.

Species	**Site ID**	**Habitat**	**Hap 1 contigs**	**Hap 1 N50**	**Hap 2 contigs**	**Hap 2 N50**
P. mexicana NS	Taco NS	Nonsulfidic	405	5,429,777	315	7,850,947
P. mexicana S	Taco S	Sulfidic	164	27,106,189	169	26,192,946
P. sulphuraria	Pich2 S	Sulfidic	161	23,574,294	113	25,996,582

Supplemental Table 3.2: Alle frequencies of structural variants and gene annotation if it occurs in a gene. Provided as separate Excel file "Pmex_SV_table_S4.xlsx".

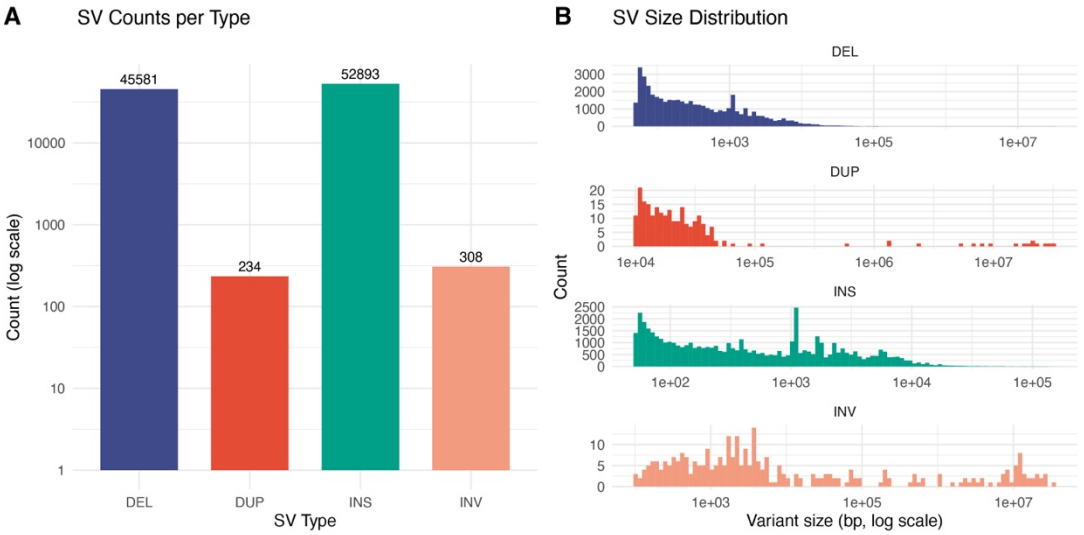
Supplemental Table 3.3: Table of genes present in F_{ST} outlier windows. Provided as separate Excel file "Pmex_SV_table_S5"

Supplemental Table 3.4: Allele frequency difference between all genotyped variants. Provided as separate Excel file "Pmex_SV_table_S6"

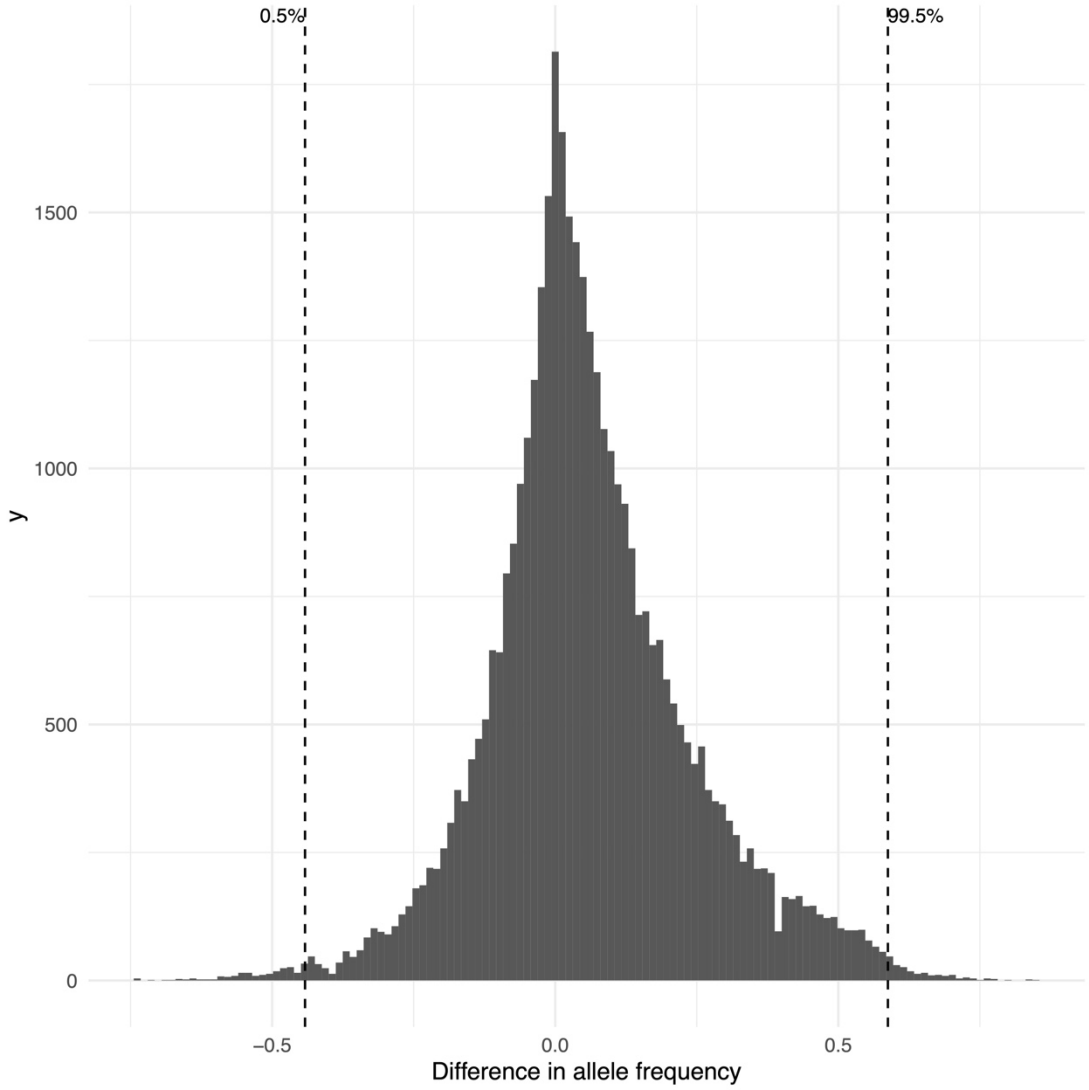
Supplemental Table 3.5 Shared, upregulated and downregulated genes containing structural variants in intronic or promoter regions.

Gene ID	Gene name	Differential expression
ethe1a	Persulfide dioxygenase	Upregulated
ppargc1b	peroxisome proliferator-activated receptor gamma coactivator 1-beta	Upregulated
gsr	glutathione reductase	Upregulated
pxdn	peroxidasin homolog	Upregulated

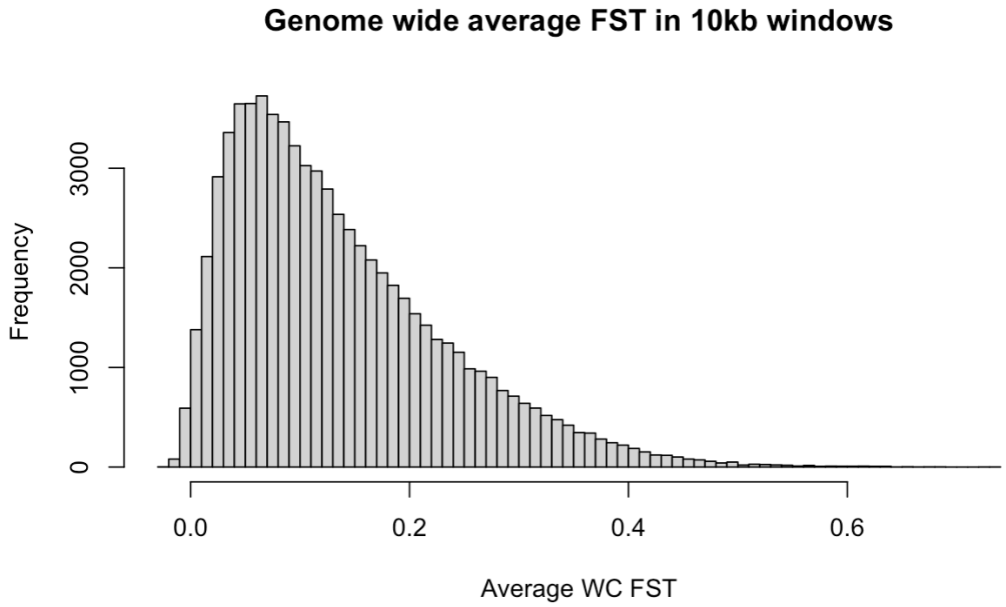
cdip1	cell death-inducing p53-target protein 1 homolog	Upregulated
apol6	apolipoprotein L6-like	Upregulated
LOC106913869	long noncoding RNA	Upregulated
slc4a4	Electrogenic sodium bicarbonate cotransporter 1	Downregulated
atp5j	ATP synthase-coupling factor 6	Downregulated
abo	histo-blood group ABO system transferase	Downregulated
map3k5	Mitogen-activated protein kinase 5	Downregulated
cish	Cytokine-inducible SH2-containing protein	Downregulated
krt16	Keratin, type I cytoskeletal 16	Downregulated
cyp3a5	Cytochrome P450 3A5	Downregulated
scamc2	Calcium-binding mitochondrial carrier protein	Downregulated
slc9a7	sodium/hydrogen exchanger 7	Downregulated



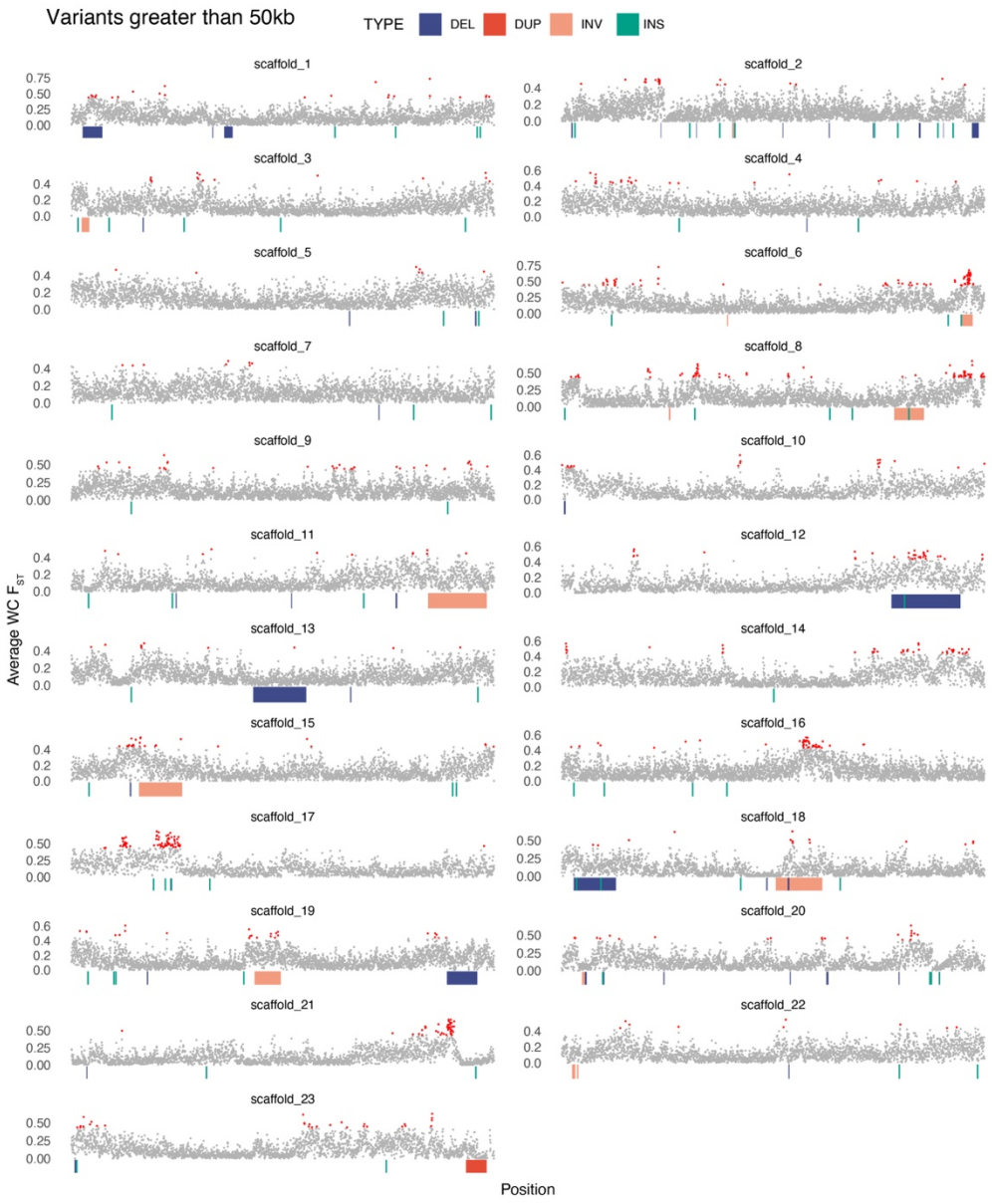
Supplemental Figure 3.1: (A) Count of variants found after merging via three methods of variant discovery for each variant type, including deletions (DEL), duplications (DUP), insertions (INS), and inversions (INV). (B) The size distribution of variants found for each variant type.



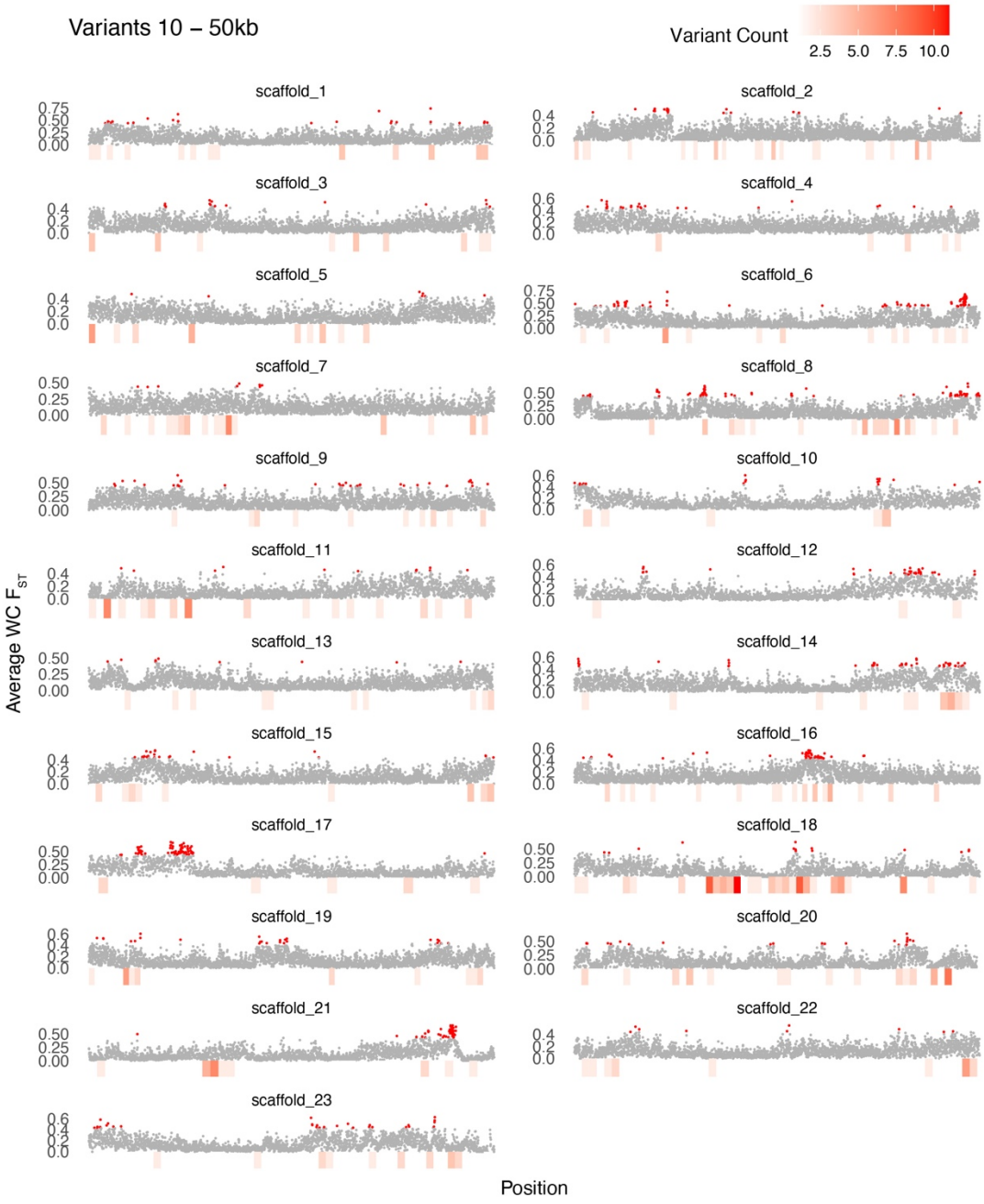
Supplemental Figure 3.2: Distribution in difference in allele frequency between sulfidic and nonsulfidic populations, where positive values are higher allele frequency in the sulfidic individuals and negative values are higher allele frequency in the nonsulfidic individuals. Dotted lines represent 99.5 and 0.05 percentile outlier threshold.



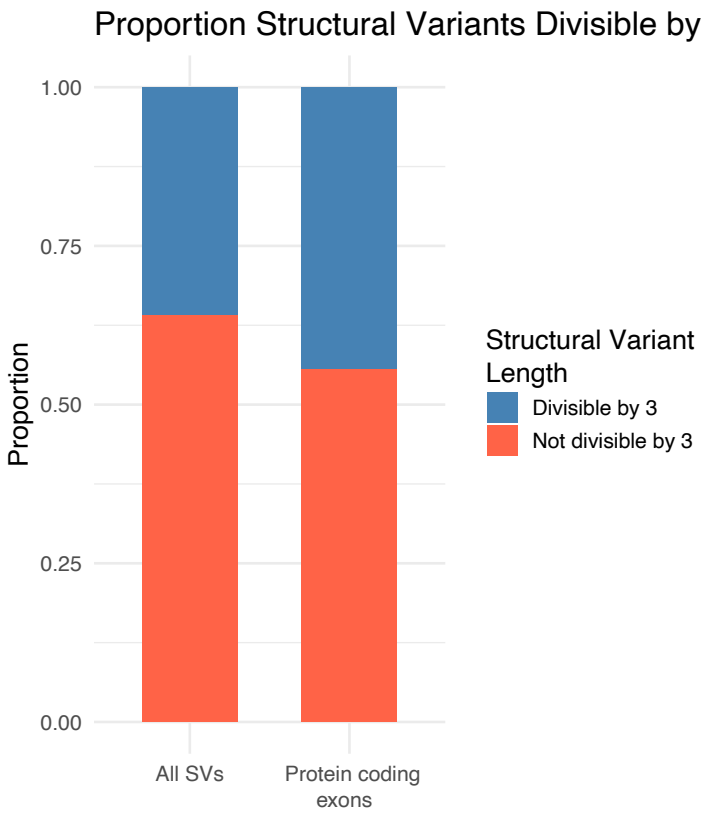
Supplemental Figure 3.3: Distribution of weighted average Weir and Cockerham F_{ST} (WC FST) calculated in 10kb windows across the genome.



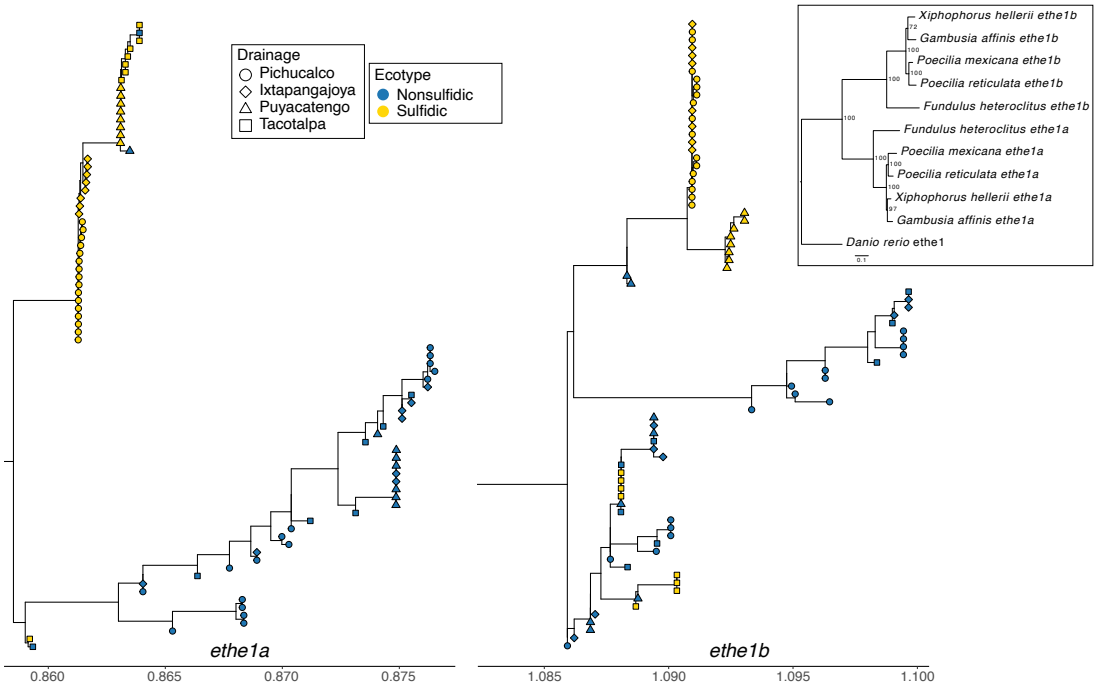
Supplemental Figure 3.4: Identification of highly differentiated (F_{ST}) outlier regions between sulfide adapted and nonsulfidic individuals and overlap with large variants greater than 50kb. Points indicated in red had a Z score > 2.996 and a F_{ST} > 0.431 (0.99 empirical cutoff). Location of variants between 50kb-5Mb plotted below FST outlier window.



Supplemental Figure 3.5: **:** Identification of highly differentiated (F_{ST}) outlier regions between sulfide adapted and nonsulfidic individuals and overlap with variants 10-50kb in size. Points indicated in red had a Z score > 2.996 and a F_{ST} > 0.431 (0.99 empirical cutoff). Heatmap below indicates the quantity of variants within a 500kb window.



Supplemental Figure 3.6: Proportion of variants divisible three for all variants and for variants found in protein coding exons.



Supplemental Figure 3.7: Gene trees for copies of *ethe1* for Poecilia mexicana individuals included in this study. For each tree, color represent the ecotype (yellow - sulfidic, blue - nonsulfidic). Shape represents drainage of origin (Pichulcalo - circle, Ixtapangajoya - diamond, Puyacatengo - triangle, Tacotalpa - square). Additionally, phylogenetic tree of *ethe1* paralogs is shown in the upper right, including a species without the duplication (*Danio rerio*).

Chapter 4 Gene flow facilitated the movement of adaptive alleles in extremophile fishes

Kara Ryan [1], Ryan Alex [1], Rishi De-Kayne [1,2], Lenin Arias-Rodriguez [3], Michael Tobler [4,5,6], Joanna L. Kelley [1]

1. Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95060, USA
2. Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA
3. División Académica de Ciencias Biológicas, Universidad Juárez Autónoma de Tabasco (UJAT), Villahermosa, México
4. Department of Biology, University of Missouri - St. Louis, St. Louis, MO, 63121, USA
5. Whitney R. Harris World Ecology Center, University of Missouri - St. Louis, St. Louis, MO, 63121, USA
6. WildCare Institute, Saint Louis Zoo, St. Louis, MO 63110, USA

Abstract

Understanding whether adaptation arises from new mutations, standing ancestral variation, or introgression is central to understanding the predictability of evolutionary outcomes. The *Poecilia mexicana* species complex provides a powerful system for testing these alternative sources of beneficial variation, as multiple lineages have independently adapted to toxic springs rich in hydrogen sulfide (H_2S). Using whole-genome resequencing of 160 individuals from ten sulfidic and nonsulfidic populations of *P. mexicana*, we combined analyses of population structure and gene flow with genome-wide selection scans and local ancestry analyses to test whether introgression facilitates the movement of adaptive alleles. We found evidence for gene flow across the phylogeny, especially between two major clades. Selection scans revealed hundreds of candidate genes, most of which were population

specific, but a subset of shared outlier windows containing key sulfide-related genes, including *sqor* and *ethe1*. Tests for excess allele sharing and analysis of haplotype structure suggest these loci exhibit signals consistent with a selective sweep after introgression. Our results suggest that gene flow has repeatedly contributed to convergent adaptation in sulfide spring fishes.

Introduction

Understanding how species adapt to novel environments is a central goal of evolutionary biology because it allows us to better understand the mechanisms that generate and maintain biodiversity. Adaptive evolution can proceed either through selection on *de novo* mutations or on standing genetic variation (Barrett & Schluter, 2008). *De novo* mutations accumulate slowly, are more often neutral or deleterious than beneficial (Kimura, 1968; Shen et al., 2022), and can be highly dependent on factors such as germline mutation rate (L. A. Bergeron et al., 2023). In contrast, standing variation can facilitate rapid adaptation (Bitter et al., 2019; Chaturvedi et al., 2021), because alleles with potentially beneficial effects are already available for selection to act upon (Barrett & Schluter, 2008; Hermisson & Pennings, 2005; Lande & Shannon, 1996; Orr & Betancourt, 2001). Despite large strides, the extent to which adaptation arises from *de novo* mutation versus standing genetic variation in natural populations and under which particular conditions each might play a role remains an open question (Agashe et al., 2023; Beatty, 2022; Bomblies & Peichel, 2022). There are striking examples of *de novo* mutation facilitating adaptation, such as the

recurrent deletion in the enhancer of the *Pitx1* gene driving pelvis reduction in stickleback (Chan et al., 2010), but growing evidence suggests adaptation is often a result of selection on standing variation (Bitter et al., 2019; Chaturvedi et al., 2021; Haenel et al., 2019; Lai et al., 2019; Ryan et al., 2023).

Understanding the origins of standing genetic variation is central to explaining how populations repeatedly adapt to new environments. Adaptive variation is often categorized as *de novo* mutation that arises after encountering a selective pressure or as pre-existing standing variation (Agashe et al., 2023; Barrett & Schluter, 2008), but pre-existing variation can exist in a population either as standing ancestral alleles maintained through speciation events or be introduced via gene flow (Agashe et al., 2023; Blanquart et al., 2012). Both theoretical and empirical studies have implicated both introgression and standing ancestral variation as the source of adaptive variation for various natural and experimental populations. For example, the role of recurrent hybridization in maintaining freshwater alleles in marine populations and thereby facilitating repeated freshwater adaptation in stickleback (Schluter & Conte, 2009) has been supported by both simulations (Galloway et al., 2020) and empirical data (Kingman et al., 2021). Additional evidence for hybridization and adaptive introgression as drivers of evolution is accumulating across diverse taxa. In *Picea* spruce trees, a suite of candidate genes associated with stress resilience and flowering time show signals of adaptive introgression between allopatric species pairs (Feng et al., 2025). In *Heliconius* butterflies, a gene with experimental evidence for determining mate preference for red patterns was introgressed from another species

(Rossi et al., 2024), highlighting a direct link between a gene and behavior and the resulting contribution to adaptation and speciation. In Malawi cichlids, hybridization and gene flow fuel the diversity driving ecological speciation and adaptive radiation (Blumer et al., 2025; Malinsky et al., 2018; Meier et al., 2017; Svardal et al., 2020). But in other systems, standing ancestral variation is implicated as the source of adaptive variation. For example, the haplotype associated with beak size variation across different species of finches is estimated to be over 1 million years old (Lamichhaney et al., 2016). In stickleback, neutral variants present in ancestral marine populations drove adaptation to acidic and basic environments (Haenel et al., 2019). Furthermore, standing variation has been shown to facilitate adaptation in a variety of experimental studies ranging from replicated field studies of *Drosophila* (Bitter et al., 2024; Rudman et al., 2022) to experiments of adaptation to ocean acidification in mussels (Bitter et al., 2019).

These studies emphasize that both standing ancestral variation and gene flow can provide beneficial alleles for adaptation and further underscore the need to understand their relative importance across taxa and contexts. But distinguishing these sources of beneficial alleles can be challenging. Incomplete lineage sorting and introgression can generate similar signals of allele sharing, such as discordant gene trees (Joly et al., 2009), making it difficult to disentangle these processes with limited genomic data or few individuals. Recent advances in whole-genome sequencing and statistical modeling enable the detection of candidate adaptive alleles and inference of

their evolutionary origins. These advances are especially useful for testing hypotheses about the source of adaptive alleles that underpin convergent phenotypes.

Extreme environments provide a strong model for testing hypotheses about mechanisms that facilitate adaptation, as strong selective pressures often result in convergent phenotypes. One such system is springs rich in toxic hydrogen sulfide (H_2S) in the Grijalva River basin in southern Mexico. While small amounts of H_2S are produced endogenously and are required for normal physiological function (L. Li et al., 2011; Peers et al., 2012; Stipanuk & Ueki, 2011; Wallace & Wang, 2015), slightly elevated levels inhibit mitochondrial Complex IV, halting oxidative phosphorylation and leading to cellular dysfunction and death (Bagarinao, 1992). Because H_2S reacts with oxygen, sulfide springs are also frequently hypoxic (Millero, 1986). Nevertheless, multiple lineages of live bearing fishes (Poeciliidae) have independently colonized springs with H_2S concentrations orders of magnitude higher than what is considered toxic to most metazoans (Tobler & Hastings, 2011; Tobler et al., 2018, 2016).

Adaptation to sulfide springs has been studied across multiple levels of biological organization in the family Poeciliidae, including behavior (Riesch et al., 2009; Sommer-Trembo et al., 2016), life history (Riesch et al., 2014), morphology (Schulz-Mirbach et al., 2016; Tobler & Hastings, 2011; Tobler et al., 2015), physiology (Greenway et al., 2020; Passow, Arias-Rodriguez, et al., 2017; Tobler et al., 2011), and genetics (Barts et al., 2018; A. P. Brown et al., 2018, 2017; De-Kayne et al., 2024; Greenway et al., 2020; Kelley et al., 2016; Passow, Brown, et al., 2017;

Passow, Henpita, et al., 2017; Pfenninger et al., 2014, 2015; Plath et al., 2013a; Ryan et al., 2023). Within the *P. mexicana* species complex, which includes a *P. sulphuraria* and *P. thermalis* clade, convergent transcriptional responses have been observed in genes involved in H_2S detoxification (A. P. Brown et al., 2018; Kelley et al., 2016; Passow, Brown, et al., 2017; Passow, Henpita, et al., 2017), with some of these transcriptional differences associated with variation in nascent transcription (Perry et al., 2024), DNA methylation (Kelley, Tobler, et al., 2021), microRNA expression (Kelley, Desvignes, et al., 2021), and structural variants (Ryan et al., *in prep*). Furthermore, studies have identified convergent *de novo* amino acid changes in subunits of mitochondrial Complex IV (Greenway et al., 2020; Pfenninger et al., 2014) as well as evidence for selection on standing variation in H_2S detoxification genes (Greenway et al., 2020). However, these previous studies were limited in their ability to distinguish whether these convergent patterns arose from incomplete lineage sorting or from introgression due to either sample size or sequencing approach. Notably, the clustering of H_2S adapted individuals in small regions of the genome (Brown et al. 2019, Greenway et al 2020) and long, shared haplotypes identified by Ryan et al. (2023) suggest a monophyletic origin of adaptive alleles in sulfide detoxification genes, raising the possibility that introgression has played a key role in their persistence and spread across sulfide springs.

Here, we test the hypothesis that gene flow facilitated repeated adaptation to extreme environments by providing a source of beneficial alleles that arose in previously adapted populations. Using genome-wide data from the *Poecilia mexicana*

species complex, we first reconstructed the demographic history to establish that gene flow occurred across the phylogeny. We then combined scans for genomic divergence and haplotype structure with tests for local introgression to identify candidate regions under selection and test for evidence of introgression. Our results suggest that adaptive variation in genes associated with H_2S detoxification has a monophyletic origin and was subsequently exported via gene flow from the *P. sulphuraria* and *P. thermalis* clade. These findings provide evidence that introgression provided a source of adaptive alleles for sulfide springs, facilitating repeated adaptation to extreme environments.

Methods

Focal populations

We analyzed ten populations of the *Poecilia mexicana* species complex distributed across four river drainages in the Río Grijalva Basin of Chiapas and Tabasco, Mexico: Pichucalco (Pich), Ixtapangajoya (Ixta), Puyacatengo (Puya), and Tacotalpa (Taco). These included five sulfide-spring (S) populations and five nonsulfidic (NS) populations. Sulfidic populations consisted of two *P. mexicana* populations (Taco S, Puya S), two *Poecilia sulphuraria* populations (Pich1 S, Pich2 S), and one *Poecilia thermalis* population (Ixta S). The closely related NS populations included five *P. mexicana* populations (Taco NS, Puya NS, Ixta NS, Pich1 NS, Pich2 NS) (Figure 4.1A, Table S4.1). For each population, we sequenced 16 individuals per population.

DNA extraction, library preparation, and sequencing

This study utilized previously extracted DNA, of which about half of the samples were previously used for capture sequencing in Ryan et al., 2023. To summarize, DNA was extracted using the Gentra Puregene Tissue Kit from 5 to 10 mg of muscle tissue. Due to the potential for reduced DNA quality due to freeze-thaw cycles, DNA quality was assessed using an Agilent TapeStation and quantified DNA using the Qubit high-sensitivity DNA kit. A total of 16 individuals per population were chosen for library preparation and shipped to the Genomics and Cell Characterization Core Facility (GC3F) at University of Oregon overnight on ice. At GC3F, 10 nanograms of DNA per sample was transferred to a 384-well plate and dried in a speed-vac. A Bead-Linked Transposase (Illumina) was diluted to 1:60 in a custom buffer (10 mM Tris-HCl, 5 mM MgCl2, 8%PEG), and 6µl of diluted beads were added to DNA and vortexed to resuspend DNA. Samples were incubated at 53°C for 30 minutes. After incubation, the plate was spun down and placed on a magnet to gather beads. While on the magnet, 5µl of supernatant was removed and discarded. Then, 1µl of barcoded Illumina Nextera adapter primers (IDT), 5µl 2x Equinox Master-mix (Watchmaker Genomics), and 3µl Nuclease Free Water (NEB) were added to each well. Samples were heated to 68 °C for 3 min, 98 °C for 3 min, then amplified for 10 cycles (98 °C for 45 s, 62 °C for 30 s, 68 °C for 2 min; 68 °C for 1 min) and held at 10 °C. After amplification, 20µl of Nuclease Free water was added to each well and mixed. 2µl of each sample was pooled and cleaned twice with 0.8x

volume Total Pure NGS beads (Omega Biosciences). The resulting library was run on Illumina MiSeq Nano PE150. Read counts from the MiSeq run were used to calculate a new pool with equal representation. The new pool was cleaned as above and sequenced on two lanes of Illumina NovaSeq S4 PE150.

In addition to generating whole-genome, short-read sequencing for the 160 individuals of *P. mexicana* generated in this study, we also downloaded similar data for 12 *Poecilia reticulata* individuals generated in (Fraser et al., 2015). To do so, we used the NCBI SRA Toolkit v3.0.0 to download Illumina short reads for IDs SAMEA8418972 - SAMEA8418983 that were sequenced to a similar depth as data generated in this study. These individuals served as an outgroup for downstream analyses.

Quality control, read mapping, and variant calling

We first assessed raw sequencing quality using FastQC v0.12.1 (Andrews, 2010) and summarized the output using MultiQC v1.27 (Ewels et al., 2016). Then, we used fastp v0.23.2 (Chen, 2023) with default parameters to trim the reads. We indexed and mapped reads to the *Poecilia reticulata* reference genome (GCF_000633615.1 (Künstner et al., 2016)) using BWA-MEM2 v2.2.1 (Vasimuddin et al., 2019). The resulting SAM files were converted to BAM files using Sambamba v0.8.2 (Tarasov et al., 2015). We used Picard v2.27.1 (Broad Institute, 2019) to fix mate-pair information, Sambamba to sort BAM files, and Picard again to mark duplicated reads. The resulting BAM files were indexed using Sambamba. Mapping

statistics were generated with Sambamba, and average genome-wide coverage was calculated with Mosdepth v0.3.4 (Pedersen & Quinlan, 2018). Variants were called per scaffold using BCFtools v1.21 (Petr Danecek et al., 2021) *mpileup* and *call* and *reheader* were used to adjust the header of the resulting VCF. We used VCFtools to generate variant and individual-level statistics from the resulting raw SNP calls, including depth, quality, allele frequency, missingness, and heterozygosity. Then, we retained biallelic SNPs with a minimum mean depth of 2, a maximum mean depth of 120, a minor allele count of less than 12, minimum genotype quality of 8 and retained sites that were genotyped in 80% or 100% of individuals for downstream analyses. Additionally, we generated a VCF of invariant sites (--mac=0) with similar filtering parameters without a minor allele frequency filter for downstream analysis with VCFtools (P. Danecek et al., 2011).

Analysis of population structure and gene flow

To explore the demographic history and current population structure of this species complex, we first generated a bootstrapped phylogeny using genome-wide SNPs. To do so, we first generated an individual pairwise distance matrix using Plink v1.90b6.16 (Chang et al., 2015). We then used Ape v5.0 (Paradis & Schliep, 2019) to generate a best supported Neighbor Joining tree rooted with *P. reticulata*. We then randomly subsampled the VCF file to retain 1% of the original input SNPs, recalculated the individual distance matrix for each subset with Plink and generated a Neighbor joining tree with Ape. Finally, bootstrap support was calculated in R and

visualized with ggtree. In addition to a nuclear phylogeny, we also explored the mitochondrial relationship among individuals. To do so, we assembled mitochondrial genomes per individual using NOVOplasty v4.3.5 using the *Gambusia affinis* cytochrome c oxidase subunit 1 sequence (NC_004388.1:5462 - 7018) as a seed sequence. Then, we modified the headers to reflect the individual ID and concatenated the fasta sequences together. From there, we used Mafft v.7.525 (Katoh & Standley, 2013) to align sequences and used IQTree v2.2.2.6 (Minh et al., 2020) to generate a phylogeny with 1000 ultrafast bootstrap (Hoang et al., 2018) replicates. The result tree and regional multiple sequence alignment was visualized in R using ggtree and plotmsa.

We used a Principal Components Analysis (PCA) to explore individual clustering. To do so, we utilized Plink v1.90b6.16 (Chang et al., 2015) --indep-pairwise to generate a linkage-pruned input (max $r^2 = 0.2$, 50 kilobase (kb) windows, step-size 10 basepairs (bp)) and then using Plink –pca and –make-bed to calculate eigenvalue and eigenvectors. We also used Plink to generate a file for downstream Admixture analysis. The resulting Principal Components were plotted using ggplot2. We explored individual and population ancestries using Admixture (Alexander et al., 2009), testing k clusters, 2 through 10, with a 10-fold cross-validation (CV) error.

To test for evidence of historical introgression, we calculated Patterson's D statistic (Patterson et al., 2012)) to detect excess allele sharing. We also used the f_4-ratio (Patterson et al., 2012)) to estimate the proportion of introgressed ancestry between populations using DSuite v0.5 r58 (Malinsky et al., 2021) using the

phylogeny generated above. We also calculated *f*-branch statistics ((Malinsky et al., 2018)), which extend D and f_4-ratio by assigning introgression signals to specific branches of the phylogeny. Because the direction of introgression can be challenging to detect in a trio-based structure, we calculated D_{FOIL} statistics (Pease & Hahn, 2015) for symmetrical five taxon phylogenies that places the *P. sulphuraria* clade populations as sister taxa, the *P. mexicana* clade populations as sister taxa and *P. reticulata* as the outgroup.

Patterns of genetic differentiation and haplotype-based selection analyses

To detect highly differentiated regions between sulfidic and nonsulfidic populations, we analyzed genome-wide differentiation between population pairs. To do so, we merged the invariant and variant sites VCF files using VCFtools (P. Danecek et al., 2011). Then, we used pixy (Korunes & Samuk, 2021) to calculate pairwise F_{ST} in 1 kb windows. These measures were then Z-transformed and plotted in R. Given that haplotype-based statistics can outperform F_{ST} scans under certain evolutionary histories (Szpiech et al., 2021) and have been valuable in identifying recent and ongoing hard and soft sweeps (Salmón et al., 2021), we used selscan v2.1 (Szpiech, 2024) to calculate the cross population nS_L (XP-nS_L) statistic, a comparison of the number of segregating sites per haplotype length between populations (Ferrer-Admetlla et al., 2014) and the cross population extended haplotype homozygosity (XP-EHH) statistic (Sabeti et al., 2007). To do so, we first split the VCF by population using VCFtools then phased variants using Beagle (B. L. Browning et al.,

2021). We then calculated XP-nS_L and XP-EHH for each population pair using selscan. We normalized these scores genome-wide using selscan and considered SNPs with an absolute Z-score > 3.3 as outliers. Additionally, we used rehh 3.2.2 (Gautier et al., 2017; Gautier & Vitalis, 2012) to calculate and visualize Extended Haplotype Homozygosity and haplotype length around regions of interest. To annotate regions with signatures of selection, we used R packages GenomicRanges, IRanges, and rtracklayer to annotate outlier regions with the *P. reticulata* GFF annotation. We used R package ComplexUpset to plot the overlap across drainages and used ggplot2 to plot genome-wide measures of differentiation and selection.

Analysis of localized signals of adaptive introgression

For trios that showed significant evidence for gene flow between *P. thermalis* (Ixta S) or *P. sulphuraria* (Pich1 S or Pich2 S) and *P. mexicana* (Puya S or Taco S) in the proper phylogenetic context, we calculated D and d_f (Pfeifer and Kapan 2019) in 10 SNP sliding windows with 1 SNP step size using Dsuite v0.5 r58. Outlier windows were identified using a 99% empirical cutoff. These regions were annotated and compared to regions under selection in R using GenomicRanges, IRanges, and rtracklayer

Results

DNA sequencing and variant calling

We generated an average of 28,861,309 ± 5,738,279 reads per sample per lane for 160 individuals (Table S4.2). The reads mapped to the *P. reticulata* genome with an average of 97.03 ± 0.34% and an average coverage of 5.20 ± 1.08X per sample per lane. After merging both lanes, average coverage was 10.39 ± 2.11X per sample (Table S4.2). For the previously generated *P. reticulata* sequencing data from (Fraser et al., 2015), there was an average of 51,082,276 ± 14,939,508 reads per sample (Table S4.2). The reads mapped to the *P. reticulata* genome with an average of 97.96 ± 0.11% and average coverage of 6.45 ± 2.43X per sample. Variant calling of single nucleotide polymorphisms (SNPs) identified over 16 million unfiltered SNPs across 172 individuals. After filtering for multiallelic sites, high missingness, low-quality genotypes, mean depth, and rare alleles, we retained 3,131,448 SNPs for downstream analyses.

Identification of an F1 hybrid

We identified one putative hybrid in our dataset that was phenotyped as nonsulfidic and collected from a nonsulfidic site in the Pichucalco drainage, but it landed directly between the sulfidic and nonsulfidic Pichucalco population clusters on Principal component 1 and 2 (Figure S4.1). Evidence for shared ancestry with sulfidic and nonsulfidic individuals was further supported by an Admixture analysis (Figure S4.2). The mitochondrial genome from this individual clusters with the *P. sulphuraria* individuals, suggesting the mother was a sulfidic female (Figure S4.3). Due to the potential for this putative F1 hybrid to influence the inference of

population evolutionary history, it was removed prior to the remainder of the analyses.

Population structure and discordance between autosomal and mitochondrial phylogeny

We first inferred the relationship among populations using genome-wide phylogenetic analyses. The best-supported phylogeny divided the complex into two major clades: one containing *P. sulphuraria* and *P. thermalis* populations (hereafter the *P. sulphuraria* clade) and another containing all *P. mexicana* populations (the *P. mexicana* clade; Figure 4.1A). This topology is broadly consistent with previous phylogenies generated from a single individual per population (Greenway et al., 2020) and from population-level target capture data (Ryan et al., 2023), with some notable differences. For example, all phylogenies support the Puya populations as sister taxa, and Greenway (2020) recovered a monophyletic grouping of Puya S, Puya NS, and Taco S. But these phylogenies vary in the relationship among nonsulfidic populations. For example, Greenway (2020) suggested that Pich1 NS and Pich2 NS were sister taxa that shared a more recent common ancestor not shared by other *P. mexicana* populations, which was not supported by this phylogeny. Instead, our data supports a stepwise branching of nonsulfidic populations (Figure 4.1A). All nodes in our genome-wide tree are highly supported (Figure 4.1A).

This genome-wide phylogeny varies from previous mitochondrial phylogeny generated in Brown (2017), which suggested that *P. sulphuraria* populations were not

sister taxa and did not support the Puya populations as sister taxa. Given this discordance, we assembled the mitochondrial genome for each individual, aligned these assemblies, and generated a mitochondrial phylogeny. Interestingly, the mitochondrial phylogeny is highly discordant from the autosomal phylogeny. The mitochondrial phylogeny suggests that *P. thermalis* clusters within the *P. sulphuraria* group (Figure S4.3). Furthermore, it does not support the monophyletic grouping of Puya S, Puya NS and Taco S (Figure S4.3). Additionally, there is stronger population-specific clustering for sulfidic individuals than nonsulfidic individuals (Figure S4.3).

Next, we assessed population structure using ~2.5 million LD-pruned SNPs in intergenic regions, excluding the putative F1 hybrid using a Principal Components Analysis (PCA). PCA revealed strong population-specific clustering across all populations (Figure 4.1B). The majority of genetic variation explained by PC1 (accounting for 23% total variance) splits the two clades (Figure 4.1B), while PC2 separates populations within the clades (accounting for ~9% of total variance).

Gene flow occurs across the species complex

Gene flow between non-sister taxa was pervasive across the species complex. There was significant evidence for gene flow between non-sister taxa in many trios, including across clades (Figure 4.1C & D). There were particularly strong signals involving *P. thermalis* (Ixta S) and nearly all *P. mexicana* populations, including Ixta NS, Taco NS, Taco S, Puya S and Puya NS (D = 0.211 - 0.0148; p-values = 2.3 x 10^{-}

^{16} to 1.4 x 10 $^{-12}$; Figure 4.1C, D). In contrast, evidence for gene flow from the *P. sulphuraria* clade into Puya S and Taco S was weaker when tests were conditioned on within-drainage nonsulfidic populations as the sister taxon. For example Puya S (Puya NS as sister taxa), evidence for gene flow was low but significantly (Puya: D = 0.0061 - 0.0056, p-values = 4.1E-7 - 4.8E-6) whereas for Taco S (Taco NS as sister taxa) there was no significant evidence (D = 0.0064 - 0.0051, p-values = 0.0029 - 0.0127).

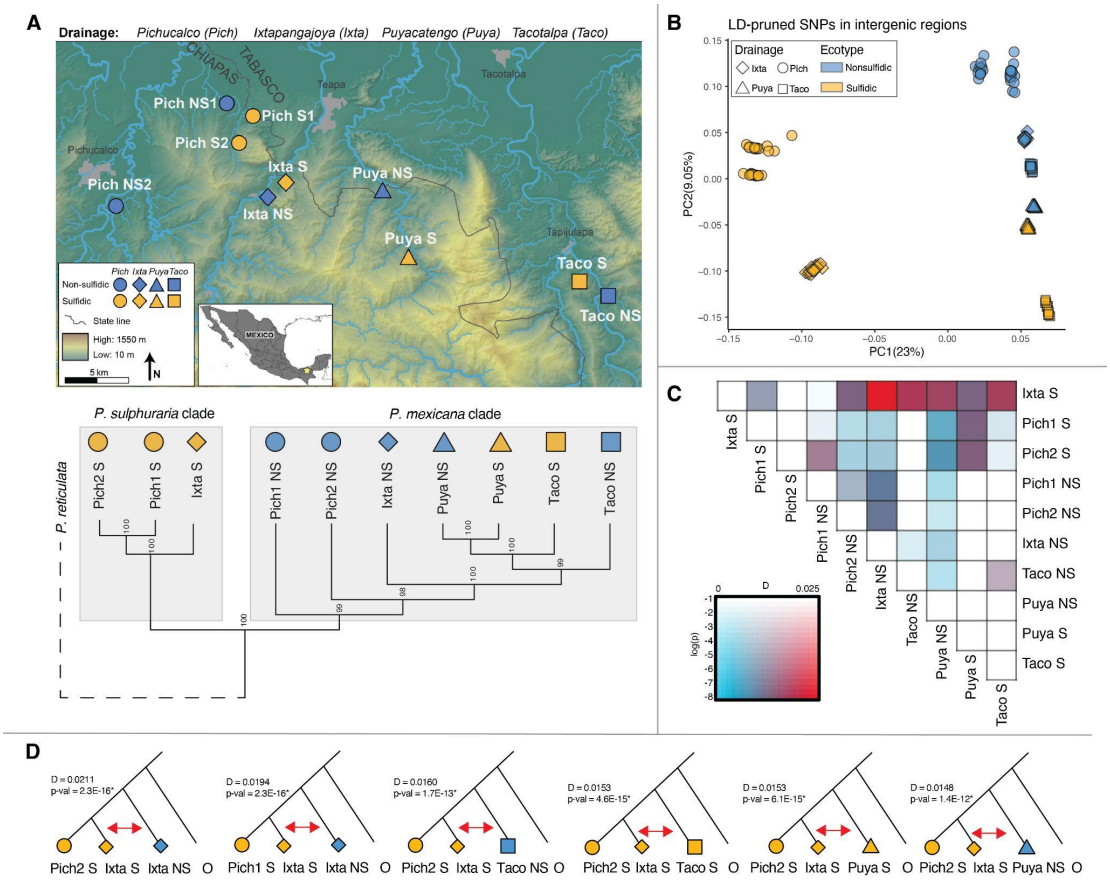


Figure 4.1: Population structure and evidence for gene flow among populations of the Poecilia mexicana species complex. (A) Map of sample sites for focal populations in this study that span four river drainages and a Neighbor Joining phylogeny with bootstrap support shown on each node. Shapes represent the drainage of origin (circle - Pich; diamond - Ixta; triangle - Puya; square - Taco) and color represents ecotype

(gold - sulfidic; blue - nonsulfidic) Figure adapted from (Hotaling et al., 2019) (B) Principal component analysis of linkage-pruned SNPs from intergenic regions. (C) Evidence of gene flow from Patterson's D statistic. Color represents D statistic value (blue = low, red = high) and opacity represents the p-value, generated from a Z-score. (D) Six trios with highest D statistic value across all trios tested.

Given that a single gene flow event can generate multiple significant D values among non-independent trios, we calculated *f* and the *f*-branch (f_b) metric to quantify excess allele sharing between both tips and internal branches of the phylogeny (Figure S4.4, Figure 4.2). This analysis showed a strong signal of allele sharing between Taco NS and Taco S (f_b = ~0.26). Taco NS also shows excessive allele sharing with Puya NS but not Puya S (Figure 4.2). We also identified a broad signal of excess allele sharing along the Ixta S branch, involving all *P. mexicana* populations except for Pich2 NS (Figure 4.2). On this branch, the highest f_b occurs between Ixta populations and decreases along the branch, suggesting a correlated, single event versus ongoing recent gene flow events. Finally, contrasting patterns emerged between the Puya ecotypes. Puya NS showed increased allele sharing with nonsulfidic *P. mexicana* populations whereas Puya S shared more with the sulfidic *P. mexicana* populations (Figure 4.2).

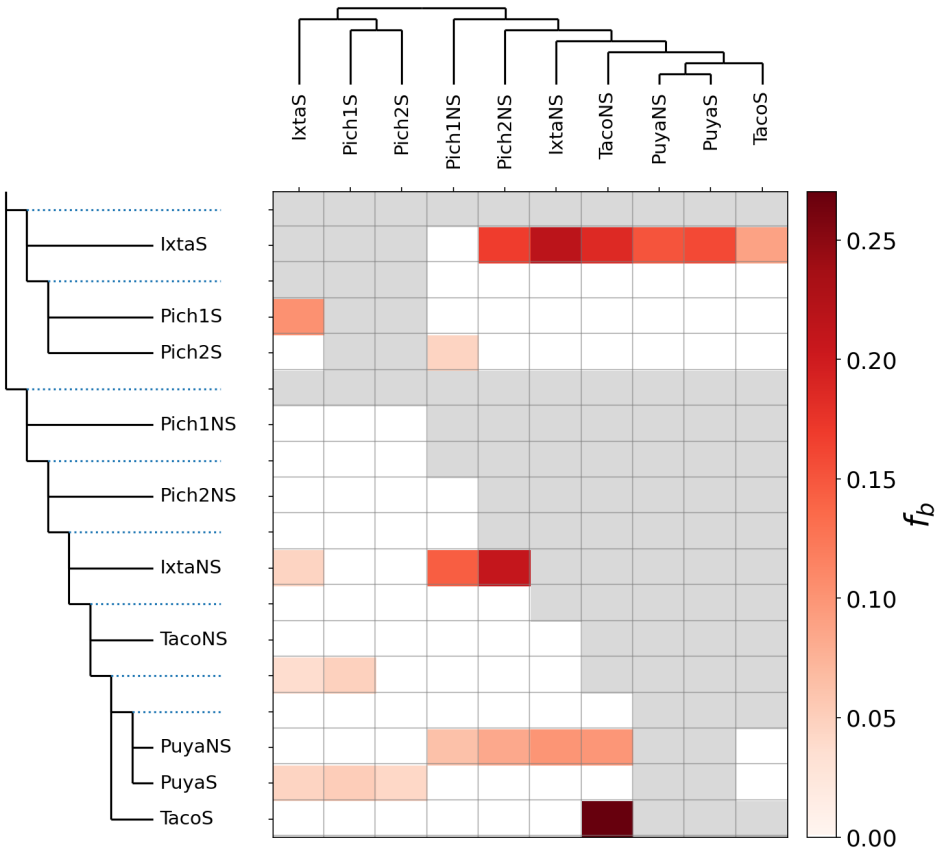


Figure 4.2: *f*-branch (f_b)statistics across the Poecilia mexicana species complex supports multiple gene flow events. Grey represents pairs not tests due to phylogenic relationship.

To test for both evidence and directionality of introgression, we calculated D_{FOIL} statistics for combinations of phylogenetically relevant five-taxa groups. Of the 63 combinations, 24 showed significant evidence for introgression on at least one chromosome (Table 4.1). The strongest signal of gene flow, supported by 13-15 chromosomes across all quartets was from Pich1 S (donor) into all *P. mexicana* populations when Pich1 NS is sister taxa to the acceptor but no evidence for Pich1 S into Pich1 NS in any quartets. This analysis almost always assigns Pich1 S as the donor into both ecotypes in *P. mexicana*. But there is also strong support for

introgression events from Ixta NS into Pich1 S and Taco NS into Pich1 S. Finally, there are four inferred introgression events from the ancestor of Pich1 S and Pich2 S into different *P. mexicana* populations (Table 4.1).

Table 4.1:Inferred introgression from DFOIL. Populations P1 through P4 set utilizing previously generated estimates of divergence times (Brown et al. 2017; Brown et al. 2018). The significant introgression event is labeled as the donor population -> acceptor population. Introgression that was inferred from a shared ancestor is labeled with both populations, for example 12 for the shared ancestor of P1 and P2. The number of chromosomes that show significant support for that introgression event as included.

P1	**P2**	**P3**	**P4**	**Introgression**	**Number of chromosomes that support introgression**
Pich1 S	Pich2 S	Pich1 NS	Pich2 NS	1->4	15
Pich1 S	Pich2 S	Pich1 NS	Ixta NS	1->4	14
Pich1 S	Pich2 S	Pich1 NS	Puya NS	1->4	13
Pich1 S	Pich2 S	Pich1 NS	Puya S	1->4	13
Pich1 S	Pich2 S	Pich1 NS	Taco NS	1->4	13
Pich1 S	Pich2 S	Pich1 NS	Taco S	1->4	13
Pich1 S	Pich2 S	Puya NS	Ixta NS	4->1	7
Pich1 S	Pich2 S	Pich2 NS	Taco NS	4->1	6
Pich1 S	Pich2 S	Pich2 NS	Ixta NS	1->4	4
Pich1 S	Pich2 S	Pich2 NS	Puya NS	1->4	4
Pich1 S	Pich2 S	Pich2 NS	Puya S	1->4	4
Pich1 S	Pich2 S	Puya S	Taco NS	1->3	4
Pich1 S	Pich2 S	Taco S	Ixta NS	1->4	4
Pich2 S	Ixta S	Pich2 NS	Taco NS	3->1	4
Pich1 S	Pich2 S	Pich2 NS	Taco S	12->3	3
Pich1 S	Pich2 S	Puya NS	Taco NS	1->3	3
Pich1 S	Pich2 S	PuyaN S	Taco S	1->3	3
Pich1 S	Pich2 S	Puya S	Taco S	1->3	3

Pich1 S	Pich2 S	Taco NS	Ixta NS	1->4	3
Pich1 S	Pich2 S	Taco S	Taco NS	12->4	3
Pich1 S	Ixta S	Pich2 NS	Taco NS	4->1	1
Pich1 S	Pich2 S	Pich2 NS	Taco NS	12->4	1
Pich1 S	Pich2 S	Pich2 NS	Taco NS	12->3	1
Pich2 S	Ixta S	Pich2 NS	Taco NS	4->1	1

Evidence for selection across the genome

We identified regions under selection in sulfidic populations using complementary approaches that capture both allele frequency differentiation and haplotype structure, given the spectrum of divergence time between locally adapted populations. First, we calculated F_{ST} between ecotypes within a drainage in 1kb windows. Relative divergence (mean weighted genome-wide Weir & Cockerham F_{ST}) was lowest in Puya (0.079), followed by Taco (0.225), Pich1 (0.275), Ixta (0.302) and highest in Pich2 (0.322). Outlier thresholds (99th percentile) ranged from (0.498 - 0.856) across comparisons (Figure S4.5). As expected, most outlier windows were unique to individual comparisons, consistent with largely independent adaptation within each drainage both to hydrogen sulfide and other drainage specific factors (Figure S4.6 & S4.7).

Despite this largely independent signal, 60 shared outlier windows overlapped with genes, including sulfide detoxification gene sulfide quinone oxidoreductase (*sqor*) and other genes involved in sulfur related pathways, such as adenosylhomocysteinase 2 and heparan sulfate glucosamine 3-O-sulfotransferase 5 (Figure S4.7). Interestingly, many of these remaining genes are involved in signaling,

transcription, and other biological processes not directly linked with H_2S detoxification or Oxidative Phosphorylation (Table S4.3). Beyond these, 162 genes were found as under selection in four of the five populations, excluding Taco, and 92 in four of the five populations, excluding Puya (Figure S4.7, Table S4.3). These included multiple genes involved in sulfur or thiol-related pathways (thiopurine S-methyltransferase (Taco), methionine sulfoxide reductase B3 (Puya), kynurenine-oxoglutarate transaminase 3 (Puya) and nucleoredoxin (Taco)) and sulfate transport (solute carrier family 13 member 1 (Puya) and family 26 member 1 (Puya). Additionally, these included multiple genes associated with OxPhos, including cytochrome C oxidase assembly factor 3b (Puya) and ubiquinol-cytochrome C reductase core protein 2b (Puya). Interestingly, 27 unique genes that code for subunits, assembly factors of cytochrome c oxidase, cytochrome b or the cytochrome b-c1 complex, are under selection, of which 14 are unique to a drainage (Table S4.3).

We also applied haplotype-based statistics (XP-EHH and XP-nS_L) to detect long, high-frequency haplotypes consistent with selective sweeps (Figure 4.3 & S4.8). The two approaches largely converged, with most XP-EHH outliers also detected by XP-nS_L (Table S4.4 & S4.5). Similar to the F_{ST} scans, most genes under selection were unique to a population, suggesting a large degree of independent adaptation and highly polygenic nature of adaptation to sulfide springs (153 - 1,293 genes) (Figure 4.3). Puya S shares more genes under selection with Pich1 S, Pich2 S and Ixt S (79 - 62 genes) than with the other *P. mexicana* sulfidic population (Taco, 12 genes) (Figure 4.3). Of note, the sulfidic Taco S has reduced heterozygosity compared to all

other populations in this study, and this reduced number of sharing in Tacotalpa compared to Puyacatengo could be the result of limitations in detection due to demographic history (Figure S4.9).

XP-EHH identified two genes under selection in all five populations (ecto-ADP- ribosyltransferase 5-like and leucine-zipper-like transcriptional regulator 1) neither of which were identified by F_{ST} scans (Table S4.4 & S4.5). Both XP-EHH and XP-nS_L also recovered H_2S detoxification and metabolism genes including persulfide dioxygenase (*ethe1*) (under selection in Pich1, Pich2, Ixta and Puya) and 3-mercaptopyruvate sulfurtransferase-like (under selection in Pich1, Pich2, Puya and Taco) (Table S4.4 & S4.5). These results highlight that while many adaptive responses are lineage-specific, repeated selection on H_2S detoxification genes recurs across populations.

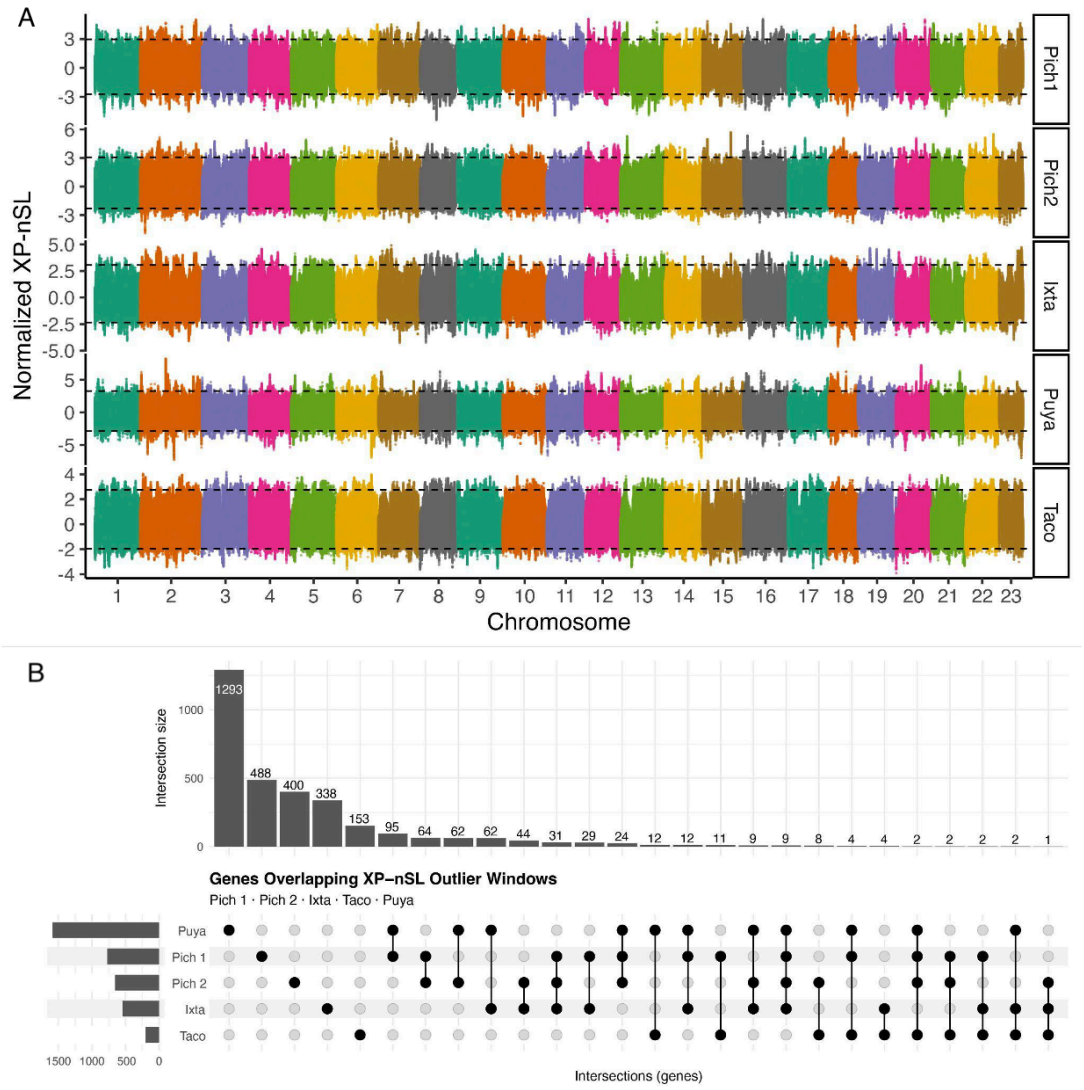


Figure 4.3: (A) Genome wide normalized Cross Population nSL (XP-nSL) between ecotypes in each drainage. The dotted line represents a |Z-score| > 3.3 outlier threshold. (B) The degree of overlap of outlier genes among population comparisons.

Adaptive introgression of candidate genes

We next tested whether regions under selection also exhibited signals of introgression from other sulfidic populations. For topologies supporting introgression, we identified 99th-percentile outlier windows for D and d_f statistics (Figure S4.12 and

S4.13). Two key H_2S detoxification genes, *sqor* and *ethe1*, consistently showed strong evidence for gene flow from the *P. sulphuraria* clade into Taco S and Puya S (Figure 4.4). In Taco S, the 300kb regions around *sqor* (27,077,021 – 27,317,942) shows evidence for introgression from Pich1 S, Pich2 S, and Ixta S, with the highest level of introgression from Ixta S (D = 0.61 - 0.71, d_f = 0.24 - 0.38; Figure 4.4). Interestingly, for the ~150kb region around *ethe1* (14,581,430 - 14,744,100), there is again evidence for Pich1 S and Pich2 S into Taco S but not Ixta S into Taco S (Figure 4.4). For Puyacatengo, a smaller region around *sqor* (27,150,254 - 27,304,838) also shows evidence for introgression from Pich1 S, Pich2 S and Ixta S (D = 0.25 - 0.28, d_f = 0.13 - 0.22) A similar pattern was seen for *ethe1*, with the strongest signal from Pich1 S (D = 0.29 - 0.41), although there is more variation among trios in this signal than around *sqor*.

Beyond these two genes showing evidence for adaptive introgression, we identified a strong signal shared by all six trios with introgression from the *P. sulphuraria* clade into Puya S and Taco S on chromosome 18 around the gene encoding protein translocase subunit SecA (LOC103480309) (XP-nS_L outlier in 4 of 5 populations). Additionally, there was evidence for adaptive introgression into Puya S around WD repeat domain 7 (*wdr7*) and mitochondrial matrix import factor 23 (*mix23*), but notably, these genes do not show evidence for adaptive introgression into Taco. Of the 26 genes showing evidence for selection in four of five populations via XP-EHH scans, only three additional genes show evidence for introgression. In Taco S, transmembrane protease serine 6 (LOC103477882) is under selection and shows

evidence for introgression (although distance to *ethe1* could be driving this signal, ~10kb distance) and cyclin G2 (*ccng2*) and shroom family member 3 (*shroom3*) in Puya S with Ixta S as the donor.

We explored extended haplotype homozygosity around these regions with signals of adaptive introgression. Consistently, there were longer haplotypes with lower heterozygosity in sulfidic populations compared to nonsulfidic populations around *sqor* and *ethe1*. Interestingly, the two copies of the *ethe1* gene had variable signals of selection. The 200kb region around these copies is considered an outlier in all populations. But the focal SNP with the highest haplotype-based outlier scores varies among populations. For Ixta S and Puya, the highest outlier SNP is present in exon 7, for Pich1 it is present in exon 1, for Pich2, intron 4 and Taco in exon 4. Interestingly, a limited number of outlier SNPs were present in the other copy of *ethe1* (LOC103477883) despite its proximity to its copy (<1kb between copies).

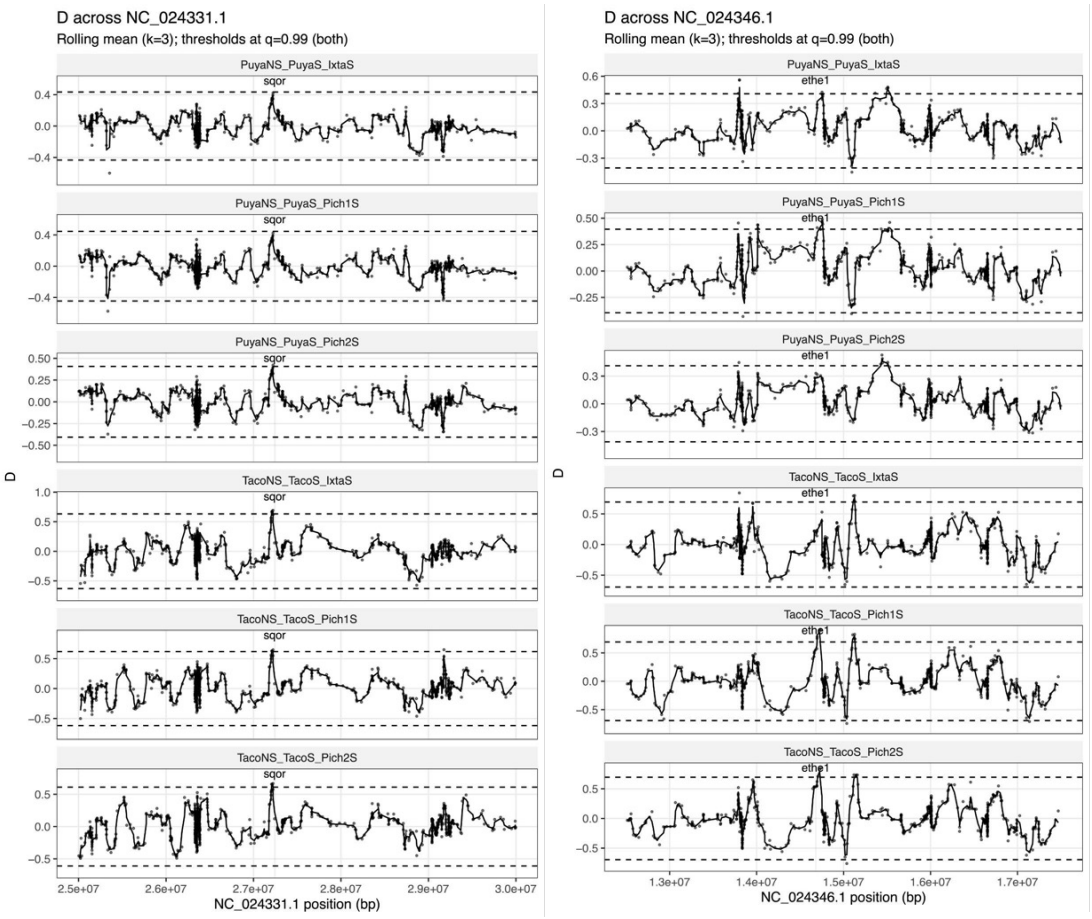


Figure 4.4:Windowed D statistic around candidate genes sqor and ethe1. Trios labeled above each plot are labeled as P1_P2_P3 with P. reticulata as the outgroup. Points represent the value in the sliding window, and the solid line represents the rolling mean. The d dotted line represents a 0.99% outlier threshold.

Discussion

An open question in evolutionary biology is how often adaptation via selection on *de novo* mutation, segregating standing ancestral variation, or on variation introduced via gene flow. Here, we analyzed whole-genomes of 160 *Poecilia mexicana* individuals across five sulfidic and nonsulfidic population pairs, together with the genomes of 12 outgroup individuals, to explore the role of gene

flow in facilitating adaptation to springs rich in hydrogen sulfide. We found that, while most regions under selection were unique to a drainage, suggesting a large degree of independent adaptation, a small number of genes – including the sulfide detoxification genes *sqor* and *ethe1* – showed both strong signatures of selection and evidence of introgression. These findings suggest that adaptive alleles originating in the *P. sulphuraria* clade were subsequently transferred among populations via gene flow, thereby facilitating repeated colonization of hydrogen sulfide rich springs.

Discordance between autosomal and mitochondrial tree

This study is the first to generate whole-genome sequencing data from multiple individuals per population across ten populations for the *P. mexicana* species complex, complementing earlier phylogenetic analyses that relied on a small number of loci or mitochondrial genes (A. P. Brown et al., 2017; Pfenninger et al., 2014; Ryan et al., 2023) or a single individual per population (A. P. Brown et al., 2019; Greenway et al., 2020). While all studies support a split between the *P. sulphuraria* and *P. mexicana* clade, the finer-scale relationship within the clade differs depending on the marker set. Given the high level of bootstrap support across a vast number of markers, our phylogeny, which supports a stepwise branching pattern among *P. mexicana* populations as well as a monophyletic group that consists of Puya S, Puya NS, and Taco S provides the highest resolution to date of the relationship among these populations (Figure 4.1A).

The mitochondrial phylogenies recovered here and in previous work (A. P. Brown et al., 2017) underscore the substantial discordance with autosomal histories, highlighting the complexity of evolutionary relationships in this species complex. In our analyses, *P. sulphuraria* (Pich1 S) cluster with *P. thermalis* individuals (Ixta S) instead of *P. sulphuraria* (Pich2 S) individuals (Figure S4.3). Interestingly, there are three nonsulfidic individuals that are basal to the remaining *P. mexicana* individuals (two Puya NS and one Pich1 S). Across the phylogeny, sulfidic individuals from the same population formed a cluster while nonsulfidic individuals were more dispersed across the tree. Moreover, Puya S and Puya NS did not form a monophyletic group, despite estimated divergence time as little as 200 years (A. P. Brown et al., 2018). Together, these patterns are consistent with a complex history of incomplete lineage sorting or introgression, though the limited number of variable sites in the mitochondrial genome necessitate further analyses specific to mitochondrial evolution.

These results provide an opportunity to link mitochondrial evolution to functional divergence. Convergent amino acid substitutions previously identified in cytochrome c oxidase genes (Greenway et al., 2020; Pfenninger et al., 2014) may have been shaped by the complex phylogenetic history observed here. Notably, the variants shared in *cox1* between all sulfidic populations except for Taco (Figure S4.13) could explain previous physiological work that showed Taco S has lower relative COX activity compared to other sulfidic populations and it more closely resembles a nonsulfidic population at increasing levels of H_2S exposure (Greenway et

al., 2020). Future work to model the implication of this variant in subunit 1 of cytochrome c oxidase on protein shape, function, and complex assembly could be crucial in exploring convergent amino acid changes across time scales, given that this variant is present in other sulfide adapted species across the family Poeciliidae (Greenway et al., 2020).

Evidence for gene flow in the species complex

Previous clustering analyses using Saguaro with single representatives per population suggested that only a small fraction of the genome (~1.2%) support topologies clustering at least four sulfidic populations together (A. P. Brown et al., 2018). Of these topologies, the most common pattern (0.71% of the genome) groups Puya S — though notability not Taco S or Puya NS — with *P. thermalis* and *P. sulphuraria,* despite very recent divergence times between Puya S and Puya NS (~265 years; (A. P. Brown et al., 2018)). Other topologies placed different subsets of sulfidic populations together, but represented a decreasing proportion of the genome (A. P. Brown et al., 2018). Our results expand upon these results and reveal that gene flow between lineages is more pervasive than previously recognized. D-statistics confirmed widespread introgression both within and across drainages (Figure 4.1). The *f*-branch statistic further supported excess allele sharing between *P. thermalis* and the ancestor of Puya S, Puya NS and Taco S as well as *P. thermalis* and *P. sulphuraria* into Puya S (Figure 4.2). But, given that horizontal lines across the plot can be the result of shared ancestry with the population where gene flow occurred

(Malinsky et al., 2021), we used the D_{FOIL} method to explore directionality of introgression, including assigning donor and acceptor populations. Interestingly, D_{FOIL} showed strong support for introgression from *P. sulphuraria* (Pich1 S) into all *P. mexicana* populations except for Pich1 NS (Table 4.1), suggesting that an introgression event could have occurred in the ancestor of these *P. mexicana* populations. Using previous estimates of divergence times — approximately 18,900 years ago between *P. mexicana* and *P. sulphuraria* (Greenway et al., 2024), 10,500 years ago between Taco NS and Taco S (A. P. Brown et al., 2018), 7,500 years ago between *P. themalis* and *P. sulphuraria* (Brown et al. 2017) and 265 years ago between Puya S and Puya NS (Brown et al. 2018) —the inferred timing suggests a complex history of introgression. If these estimates are accurate, then this suggests that introgression from a shared common ancestor may have facilitated the initial Tacotalpa adaptation, followed by a more recent introgression event from *P. thermalis* or *P. sulphuraria* that contributed to adaptation in Puyacatengo. But there is significant evidence for other gene flow scenarios, such as in the opposite direction from nonsulfidic *P. mexicana* populations into *P. sulphuraria* (Pich1 S), suggesting gene flow may be ongoing at low levels between ecotypes in both directions.

The presence of gene flow between ecotypes and across drainages is striking given the ecological and physical barriers that limit movement between drainages (Figure 4.1A) and the strong divergent natural and sexual selection documented between sulfidic and nonsulfidic habitats (Bierbach et al., 2012; Plath et al., 2013a; Tobler et al., 2016). Reciprocal transplant experiments have shown that all

nonsulfidic individuals exhibit extremely reduced fitness in sulfidic waters and sulfidic individuals experience reduced fitness in nonsulfidic waters, with the strength of selection varying among drainages (Plath et al., 2013a). Our results therefore raise an important question about the evolution of reproductive isolation under strong divergent selection, particularly the limited evidence for the development of mito-nuclear incompatibilities between ecotypes, gene extensive gene flow between clades. Furthermore, the identification of a putative F1 hybrid between *P. sulphuraria* and *P. mexicana* further underscores the limited reproductive barriers, given their divergence time is estimated to be ~18,900 years ago (Greenway et al., 2024). The potential for hybridization between these two species provides an opportunity to test how maternal inheritance influences the evolutionary trajectory of repeated adaptation.

Shared and unique regions under selection in sulfidic populations

Genome-wide scans for selection revealed that adaptation to sulfide springs is highly polygenic, with most outlier regions unique to individual drainages. The pattern is consistent with largely independent responses to hydrogen sulfide, as expected given its widespread direct and indirect effects on cellular pathways and physiological processes (L. Li et al., 2011; Peers et al., 2012). Nonetheless, a subset of genes shows repeated evidence for selection across most or all sulfide adapted populations. These include both *sqor* and *ethe1*, both of which play central roles in sulfide detoxification (Landry et al., 2021; Ono et al., 2014), as well as genes associated with oxidative phosphorylation and sulfur metabolism (Table S4.3-S4.5)

consistently emerged as candidates. The recurrence of these genes across independent comparisons suggests that, despite the genome-wide architecture of adaptation, certain core pathways are repeatedly targeted by selection.

Interestingly, many outlier regions fell outside of annotated gene regions. While this could reflect linkage to causal variants or incomplete gene annotations, it may also point to the presence of regulatory mechanisms. This possibility is particularly compelling in light of extensive evidence for convergent shifts in gene expression changes among sulfide adapted fishes (A. P. Brown et al., 2018; Greenway et al., 2020; Kelley et al., 2016; Passow, Brown, et al., 2017; Passow, Henpita, et al., 2017). Selection acting on noncoding regions could therefore underlie changes in transcriptional regulation. Future work integrating assays of chromatin accessibility (e.g., ATA-seq) and histone modifications (e.g., ChIP-seq) will be essential to identify regulatory elements that contribute to adaptation and will complement recent work that utilized capped-small RNA sequencing to explore promoter and enhancer activity (Perry et al., 2024). Such approaches are especially relevant for understanding the functional difference between copies of duplicated genes, like *ethe1*, where one paralog consistently exhibits stronger evidence for selection than the other.

The use of whole-genome data underscores the limitations of targeted sequencing approaches. Many of the genes we identified as under selection in this study were not included in Ryan (2023), primarily because they lack obvious ties to sulfide processes or oxidative phosphorylation. Genes under selection identified in

this study were involved in diverse processes such as signaling, transport, and transcription. Collectively, these results suggest that adaptation extends beyond sulfide metabolism and mitochondrial pathways, encompassing a much broader set of biological functions than previously appreciated.

The role of introgression in adaptation

This work adds to the growing body of evidence that gene flow can be an important source of adaptive variation. Studies across diverse taxonomic groups, including humans (Sankararaman et al., 2014), butterflies (Edelman et al., 2019; Rossi et al., 2024), nematodes (Lewis et al., 2023), spruce trees (Feng et al., 2025) and cichlids (Malinsky et al., 2018; Selz & Seehausen, 2019; Svardal et al., 2020), have demonstrated that hybridization and subsequent introgression can introduce alleles that enhance fitness in novel environments. Consistent with these patterns, we find evidence that introgression among lineages in the *P. mexicana* species complex facilitated the spread of adaptive alleles. These findings highlight that introgression can be both a homogenizing and diversifying force in evolution, potentially enhancing the capacity of natural populations to respond to strong, environmental challenges like those present in extreme environments.

Supplementary Material

Supplemental Table 4.1: Focal populations included in this study including the species name, site name, site ID used in the manuscript, the drainage of origin, the habitat type, either sulfidic or nonsulfidic springs, latitude and longitude of the population.

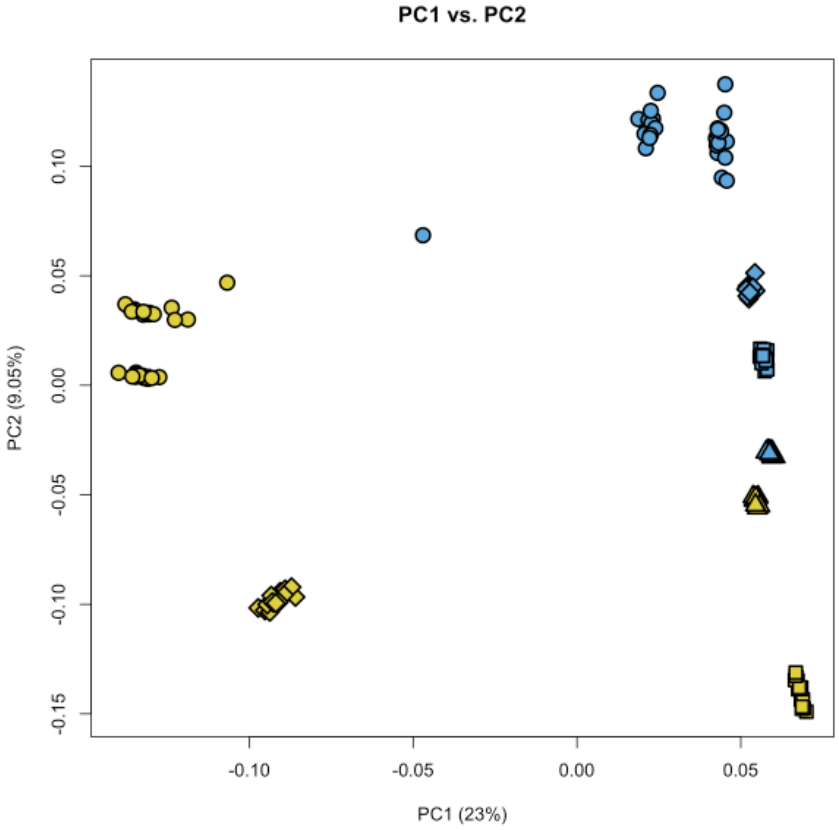
Species	Site name	Site ID	Drainage	Habitat	Lat & Lon
P. sulphuraria	Baños del Azufre	Pich 1 S	Pichucalco	Sulfidic	17.552, -92.998
P. mexicana	Vet Station	Pich 1 NS	Pichucalco	Nonsulfidic	17.556, -93.008
P. sulphuraria	La Gloria	Pich 2 S	Pichucalco	Sulfidic	17.532, -93.015
P. mexicana	Rosita	Pich 2 NS	Pichucalco	Nonsulfidic	17.485 -93.103
P. thermalis	La Esperanza	Ixta S	Ixtapangajoya	Sulfidic	17.511, -92.980
P. mexicana	Rio Ixtapangajoya	Ixta NS	Ixtapangajoya	Nonsulfidic	17.495, -92.998
P. mexicana	La Lluvia, small spring	Puya S	Puyacatengo	Sulfidic	17.463, -92.895
P. mexicana	Vicente Guerrero	Puya NS	Puyacatengo	Nonsulfidic	17.510, -92.914
P. mexicana	El Azufre I	Taco S	Tacotalpa	Sulfidic	17.442, -92.774
P. mexicana	Bonita	Taco NS	Tacotalpa	Nonsulfidic	17.426, -92.752

Supplemental Table 4.2: Sequencing and mapping results for samples included in this study. Table is included as a "Gene_flow_Table_S2_squencing_mapping.xlxs"

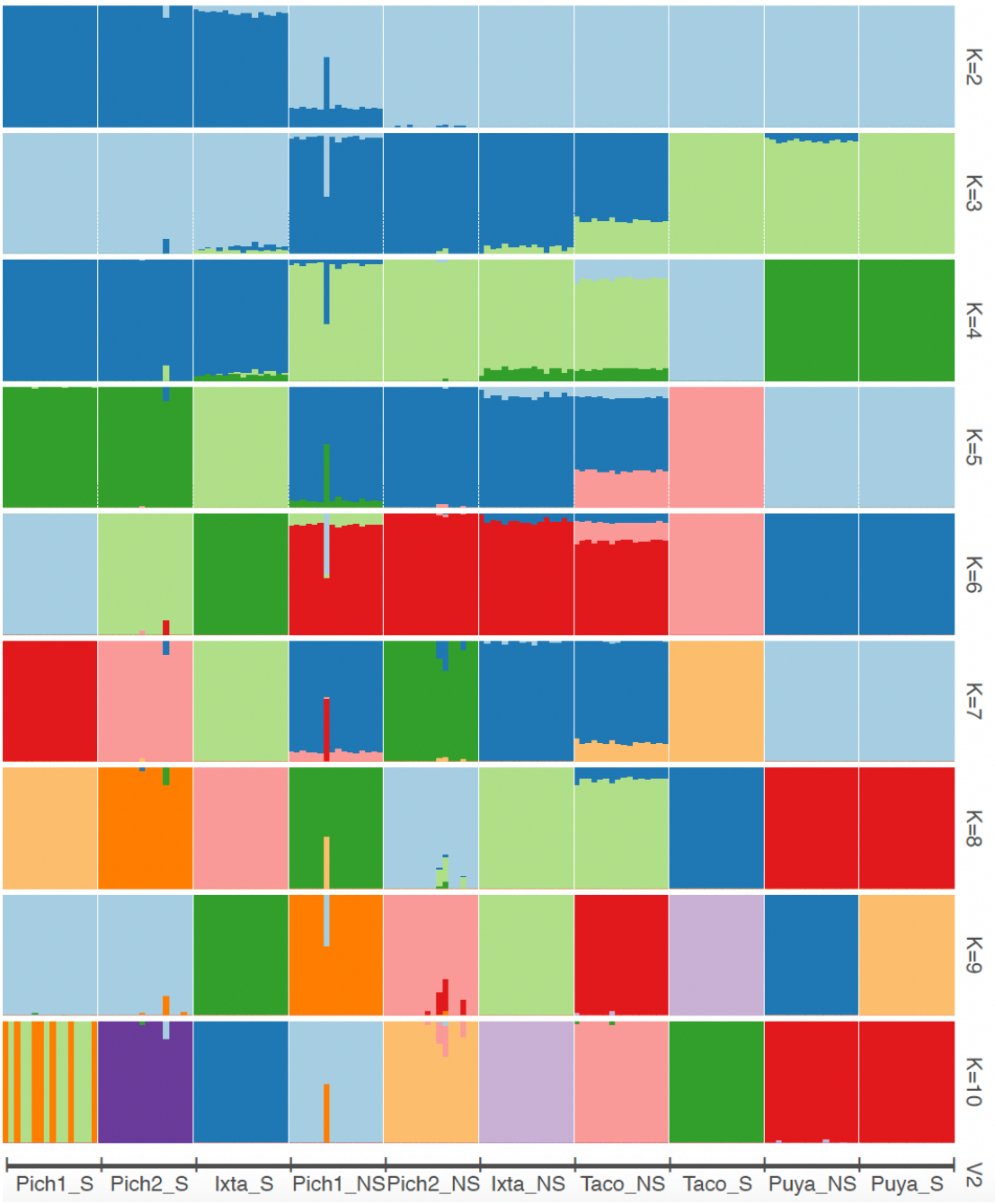
Supplemental Table 4.3: Outlier genes identified by F_{ST} scans including the gene symbol, the comparisons in which it was an outlier in, the number of comparisons it was considered an outlier in and a description of the gene. The table is included as "Gene_flow_Table_S3_FST_outliers.xlxs"

Supplemental Table 4.4: Outlier genes identified by XP-EHH scans including the gene symbol, the comparisons in which it was an outlier in, the number of comparisons it was considered an outlier in and a description of the gene. The table is included as "Gene_flow_Table_S4_xpehh_outliers.xlxs"

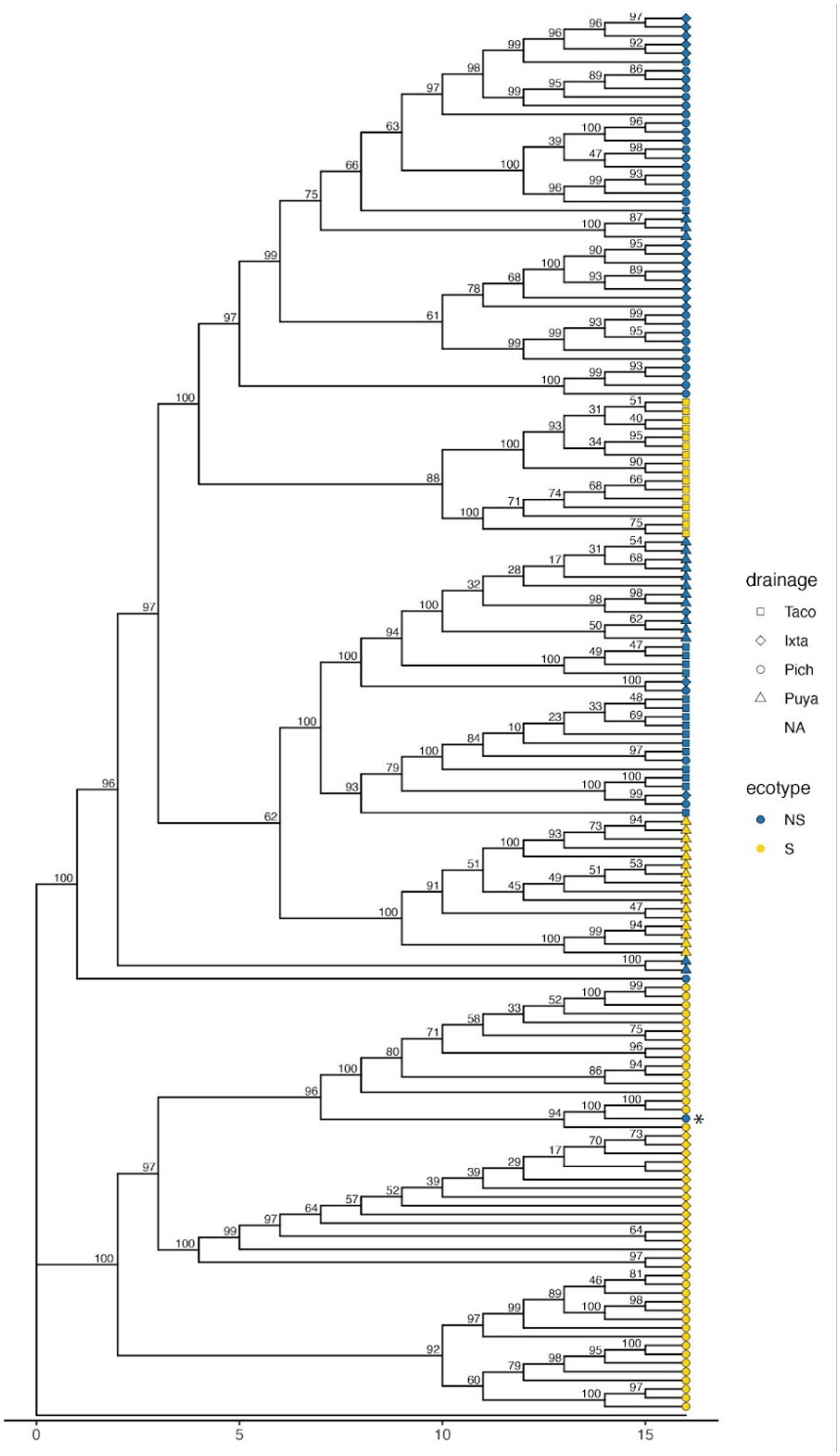
Supplemental Table 4.5: Outlier genes identified by XP-nS_L scans including the gene symbol, the comparisons in which it was an outlier in, the number of comparisons it was considered an outlier in and a description of the gene. The table is included as "Gene_flow_Table_S5_xpnsl_outliers.xlxs"



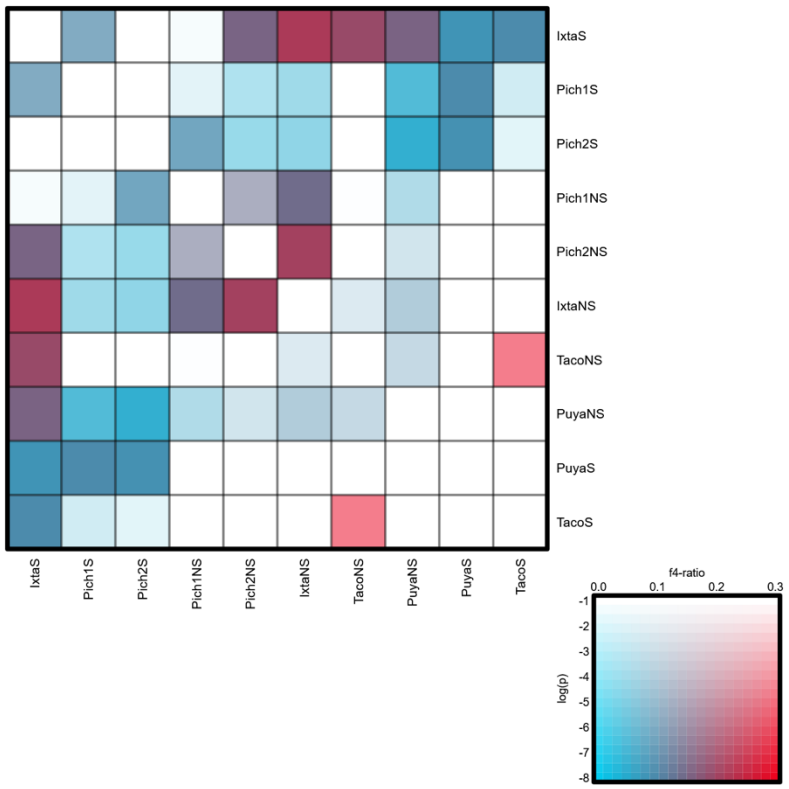
Supplemental Figure 4.0.1: Principal components analysis of all individuals, including a putative hybrid. Color represents ecotype (blue = nonsulfidic, yellow = sulfidic).



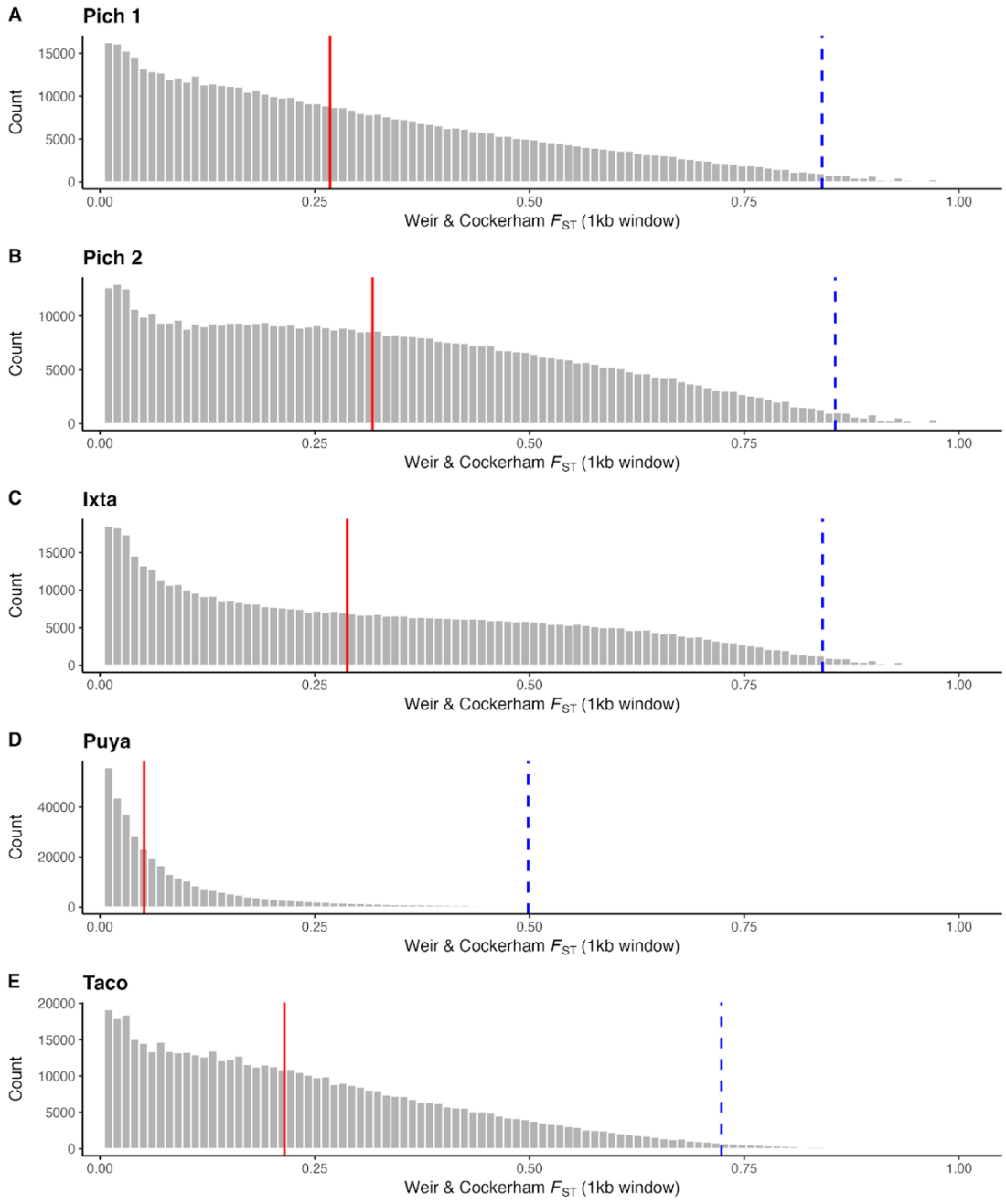
Supplemental Figure 4.0.2: Admixture plot of all individuals including putative F1 hybrid identified in the nonsulfidic Pichucalco population.



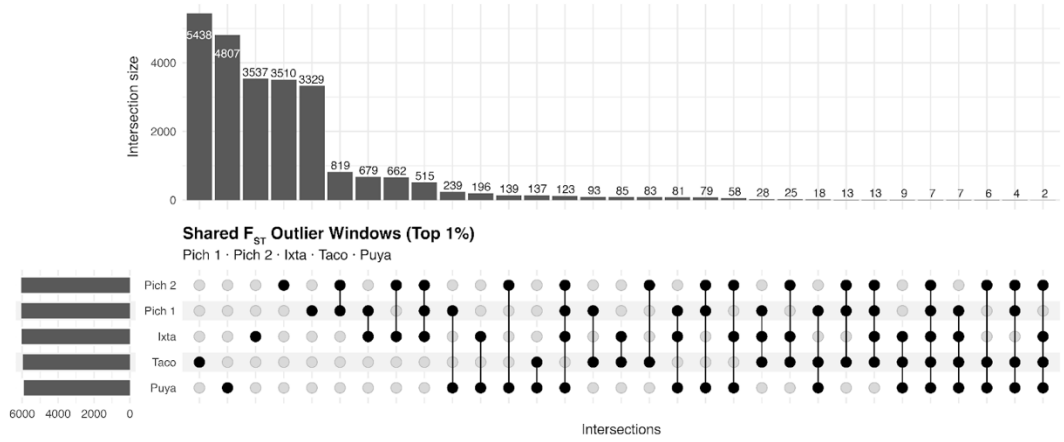
Supplemental Figure 4.0.3: Maximum likelihood mitochondrial phylogeny and associated bootstrap support scores. Shape at the tips denotes population and color represents habitat type. * denotes the putative F1 hybrid identified in the analysis.



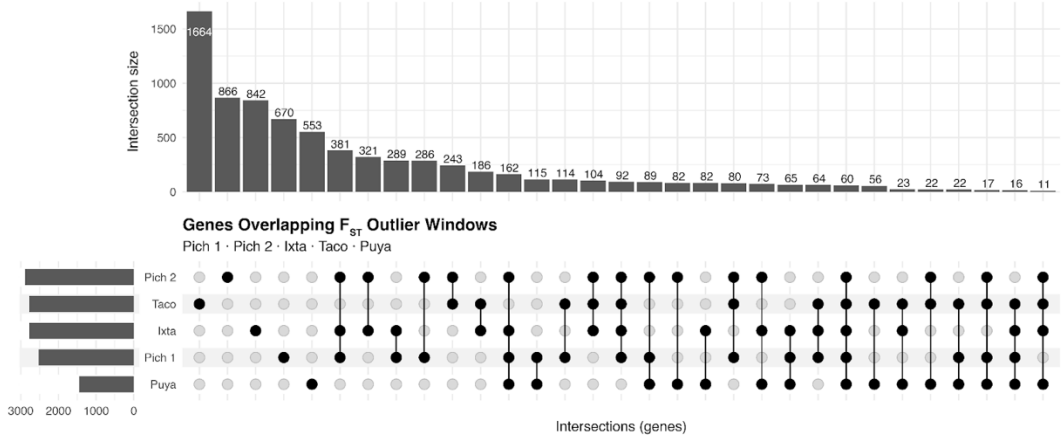
Supplemental Figure 4.0.4: f4-ratio plot showing evidence of introgression. The color presents the f4 ratio, where blue represents lower values and red represents higher values. The opacity represents significance, with more opaque colors representing more significant f4 ratios.



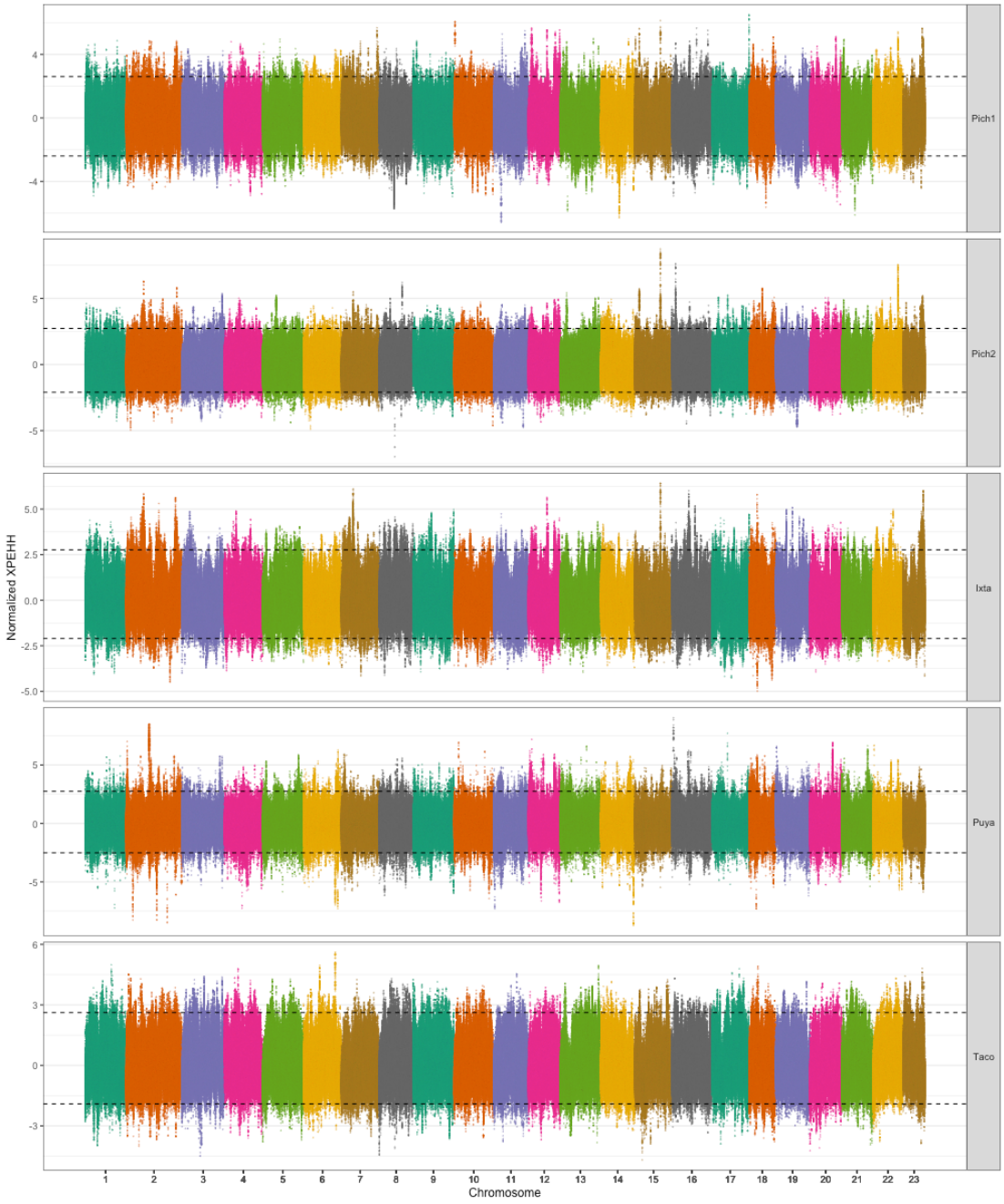
Supplemental Figure 4.0.5: Distribution of F_{ST} in 1,000bp windows between locally adapted populations within the same drainage in 1kb windows. The vertical red line represents the mean weighted F_{ST} and the blue dashed line represents the 99% outlier threshold for each comparison.



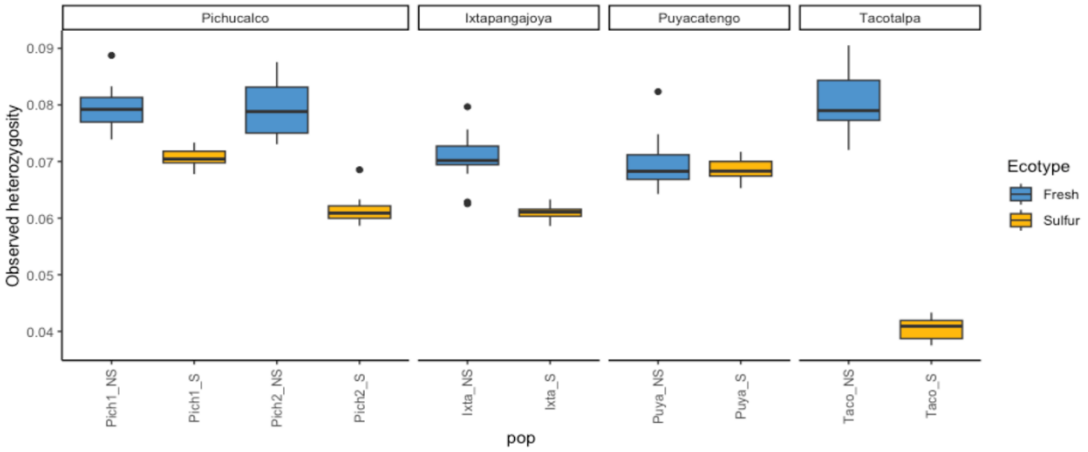
Supplemental Figure 4.0.6: Overlap of outlier F_{ST} windows among populations.



Supplemental Figure 4.0.7: Overlap of F_{ST} outlier genes among populations.



Supplemental Figure 4.0.8: Genome-wide cross population extended haplotype homozygosity (XP-EHH). 99% outlier threshold denoted with dashed line.



Supplemental Figure 4.0.9: Observed heterozygosity counted as the number of heterozygous sites over the total number of genotypes.

SYNTHESIS

Evolutionary biologists have long debated if evolution is predictable or idiosyncratic (Darwin, 1859; Jonathan B. Losos, 2011, 2017; Rosenblum et al., 2014). This question centers on whether repeated evolution is inevitable – driven by similar selective pressures leading to predictable, convergent outcomes – or whether it is contingent on historical and stochastic processes. My dissertation addresses this question by investigating the genetic underpinnings of convergent evolution in the *Poecilia mexicana* species complex, which has repeatedly adapted to toxic streams rich in hydrogen sulfide. Collectively, these studies add to this larger debate about the predictability of evolution, exploring how often convergent phenotypes are underpinned by similar genomic changes and the evolutionary processes that result in convergent phenotypes.

In Chapter One, I identified strong signals of predictability, including parallel allele frequency changes in single nucleotide polymorphisms (SNPs) in sulfide detoxification genes across all five sulfide adapted populations (Ryan et al., 2023). Chapter Two then provided a broader comparative framework, revealing conserved patterns of synteny across genera and highlighting how demographic histories differ among sulfide-adapted populations, shaping the genomic landscape on which selection acts. In Chapter Three, I also detected a predictable signal of allele frequency shifts in structural variants (Ryan et al., *in review*). Yet, phylogenetic analyses revealed that much of this adaptive variation underlying these predictable patterns had a monophyletic origin, suggesting that shared ancestry or historical

introgression – not independent mutation – contributed to these repeated outcomes. Chapter Four then showed that gene flow among populations facilitates the spread of adaptive alleles, adding a layer of contingency to seemingly predictable outcomes (Ryan et al., *in prep*).

The role of gene flow in shaping repeated adaptation is especially interesting given previous working using microsatellites showed strongly reduced gene flow between adjacent sulfidic and nonsulfidic populations, attributed to both natural and sexual selection (Plath et al., 2013b). However, more recent analyses have shown there is low, asymmetric gene flow between some sulfidic and nonsulfidic populations (A. P. Brown et al., 2018; Greenway et al., 2024). Expanding sampling near potential hybrid zones will be crucial to test whether adaptive alleles are already present in nonsulfidic populations. These data could distinguish whether sulfide springs are colonized by multiple nonsulfidic individuals, each carrying a subset of adaptive alleles (i.e. the transporter hypothesis (Schluter & Conte, 2009)) or by rare 'jackpot carriers' harboring large haploblocks of adaptive alleles, as proposed and debated to explain repeated freshwater adaptation in stickleback (Bassham et al., 2018; Galloway et al., 2020; Kingman et al., 2021; Kwakye et al., 2025).

While my findings provide insight into population-level predictability, they do not explain the broader macroevolutionary pattern of repeated adaptation to sulfide springs across the family Poeciliidae. In this family, at least 24 lineages have independently colonized sulfide springs in North American, the Neotropics and the Middle East, spanning 40 million years of evolution (Tobler et al., 2018). Some

lineages contain both sulfidic and nonsulfidic populations, while others are endemics restricted to a single sulfide spring (Greenway et al., 2014). Nearly all lineages share viviparity, a potential contingency for colonizing sulfide springs (Tobler et al., 2018). But it remains unclear whether predictable outcomes are shaped primarily by shared ancestry, gene flow, or parallel selection on the same detoxification genes. Indeed, sulfide detoxification genes consistently show predictable shifts in gene expression across independent lineages (Greenway et al., 2020), but the complex environmental and regulatory architecture underlying these shifts highlights the need for future studies. Such work will be essential to determine whether gene flow explains predictability within species, whereas independent molecular solutions dominate across deeper evolutionary scales.

Finally, this work brings to light questions regarding the evolution of reproductive barriers. Theory predicts that strong divergent selection can generate behavioral and genetic differences that reduce hybrid fitness and promote speciation (Dobzhansky, 1937; Muller, 1942; Nosil, 2012; Tobler et al., 2016). Although reduced gene flow between ecotypes is documented (Plath et al., 2013b), genomic data reveal that gene flow persists (A. P. Brown et al., 2018; Greenway et al., 2024) despite theoretical reproductive isolation. This suggests that hybrids, while having reduced fitness, may occasionally backcross and transmit adaptive alleles into nonsulfidic populations. Although detecting these incompatibilities are extremely difficult, the genetic basis of incompatibilities have been identified in other Poeciliid fishes (B. M. Moran et al., 2024; Robles et al., 2025), and recent work has suggested

that rapid sex chromosome turnover may mitigate incompatibility effects (R. L. Moran et al., 2025). Given the striking diversity in sex chromosome systems across the family Poeciliidae (Darolti et al., 2019; Fong et al., 2023; Kottler et al., 2020; Volff & Schartl, 2001), these dynamics may strongly influence both the predictability of reproductive isolation and the maintenance of gene flow.

Together, these results suggest that predictable evolutionary outcomes can emerge from contingent processes. The interplay of shared ancestry, gene flow, selection, and mutation shapes the paths available to adaptation, suggesting that predictability in evolution is not absolute, but conditional on the historical and genomic context. Ultimately, this dissertation demonstrates that while the genomic basis of adaptation can be strikingly predictable, the routes evolution takes to reach those outcomes are profoundly contingent – tracing not just where lineages end up, but who they got there.

REFERENCES

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Et al [Current Protocols in Human Genetics]*, *Chapter 7*, Unit7.20.

Agashe, D., Sane, M., & Singhal, S. (2023). Revisiting the role of genetic variation in adaptation. *The American Naturalist*, *202*(4), 486–502.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664.

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology*, *23*(1), 258.

Alper, S. L., & Sharma, A. K. (2013). The SLC26 gene family of anion transporters and channels. *Molecular Aspects of Medicine*, *34*(2), 494–515.

Alves, J. M., Carneiro, M., Cheng, J. Y., Lemos de Matos, A., Rahman, M. M., Loog, L., Campos, P. F., Wales, N., Eriksson, A., Manica, A., Strive, T., Graham, S. C., Afonso, S., Bell, D. J., Belmont, L., Day, J. P., Fuller, S. J., Marchandeau, S., Palmer, W. J., … Jiggins, F. M. (2019). Parallel adaptation of rabbit populations to myxoma virus. *Science (New York, N.Y.)*, *363*(6433), 1319–1326.

Amsterdam, A., Nissen, R. M., Sun, Z., Swindell, E. C., Farrington, S., & Hopkins, N. (2004). Identification of 315 genes essential for early zebrafish

development. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(35), 12792–12797.

Andrews, S. (2010). *FastQC A Quality Control tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bagarinao, T. (1992). Sulfide as an environmental factor and toxicant: tolerance and adaptations in aquatic organisms. *Aquatic Toxicology (Amsterdam, Netherlands)*, *24*(1), 21–62.

Balart-García, P., Aristide, L., Bradford, T. M., Beasley-Hall, P. G., Polak, S., Cooper, S. J. B., & Fernández, R. (2023). Parallel and convergent genomic changes underlie independent subterranean colonization across beetles. *Nature Communications*, *14*(1), 3842.

Barrett, R. D. H., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, *23*(1), 38–44.

Barts, N., Greenway, R., Passow, C. N., Arias-Rodriguez, L., Kelley, J. L., & Tobler, M. (2018). Molecular evolution and expression of oxygen transport genes in livebearing fishes (Poeciliidae) from hydrogen sulfide rich springs. *Genome*, *61*(4), 273–286.

Bassham, S., Catchen, J., Lescak, E., von Hippel, F. A., & Cresko, W. A. (2018). Repeated selection of alternatively adapted haplotypes creates sweeping genomic remodeling in stickleback. *Genetics*, *209*(3), 921–939.

Beatty, J. (2022). The synthesis and the two scenarios. *Evolution; International Journal of Organic Evolution*, *76*(S1), 6–14.

Begum, G., Albanna, A., Bankapur, A., Nassir, N., Tambi, R., Berdiev, B. K., Akter, H., Karuvantevida, N., Kellam, B., Alhashmi, D., Sung, W. W. L., Thiruvahindrapuram, B., Alsheikh-Ali, A., Scherer, S. W., & Uddin, M. (2021). Long-read sequencing improves the detection of structural variations impacting complex non-coding elements of the genome. *International Journal of Molecular Sciences*, *22*(4), 2060.

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics (Oxford, England)*, *32*(18), 2817–2823.

Bell, E. (2012). *Life at Extremes: Environments, Organisms, and Strategies for Survival*. CABI.

Berdan, E. L., Aubier, T. G., Cozzolino, S., Faria, R., Feder, J. L., Giménez, M. D., Joron, M., Searle, J. B., & Mérot, C. (2024). Structural Variants and Speciation: Multiple Processes at Play. *Cold Spring Harbor Perspectives in Biology*, *16*(3), a041446.

Bergeron, L. A., Besenbacher, S., Zheng, J., Li, P., Bertelsen, M. F., Quintard, B., Hoffman, J. I., Li, Z., St. Leger, J., Shao, C., Stiller, J., Gilbert, M. T. P., Schierup, M. H., & Zhang, G. (2023). Evolution of the germline mutation rate across vertebrates. *Nature*, *615*(7951), 285–291.

Bergeron, M. J., Clémençon, B., Hediger, M. A., & Markovich, D. (2013). SLC13 family of Na+-coupled di- and tri-carboxylate/sulfate transporters. *Molecular Aspects of Medicine*, *34*(2), 299–312.

Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., Robledo, D., Kent, M. P., Røsæg, L. L., Holen, M. M., Mulugeta, T. D., Ashton, T. J., Hindar, K., Sægrov, H., Florø-Larsen, B., Erkinaro, J., Primmer, C. R., Bernatchez, L., Martin, S. A. M., … Macqueen, D. J. (2020). The structural variation landscape in 492 Atlantic salmon genomes. *Nature Communications*, *11*(1), 5176.

Bierbach, D., Arias-Rodriguez, L., & Plath, M. (2018). Intrasexual competition enhances reproductive isolation between locally adapted populations. *Current Zoology*, *64*(1), 125–133.

Bierbach, D., Klein, M., Sassmannshausen, V., Schlupp, I., Riesch, R., Parzefall, J., & Plath, M. (2012). Divergent Evolution of Male Aggressive Behaviour: Another Reproductive Isolation Barrier in Extremophile Poeciliid Fishes? *International Journal of Evolutionary Biology*, *2012*(1), 148745.

Bitter, M. C., Berardi, S., Oken, H., Huynh, A., Lappo, E., Schmidt, P., & Petrov, D. A. (2024). Continuously fluctuating selection reveals fine granularity of adaptation. *Nature*, *634*(8033), 389–396.

Bitter, M. C., Kapsenberg, L., Gattuso, J.-P., & Pfister, C. A. (2019). Standing genetic variation fuels rapid adaptation to ocean acidification. *Nature Communications*, *10*(1), 5821.

Blanquart, F., Gandon, S., & Nuismer, S. L. (2012). The effects of migration and drift on local adaptation to a heterogeneous environment: Drift and migration and LA. *Journal of Evolutionary Biology*, *25*(7), 1351–1363.

Blount, Z. D., Borland, C. Z., & Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(23), 7899–7906.

Blount, Z. D., Lenski, R. E., & Losos, J. B. (2018). Contingency and determinism in evolution: Replaying life's tape. *Science (New York, N.Y.)*, *362*(6415), eaam5979.

Blumer, L. M., Burskaia, V., Artiushin, I., Saha, J., Garcia, J. C., Jiménez, F. C., Hooft van Huysdynen, A., Elkin, J., Fischer, B., Van Houtte, N., Zhou, C., Gresham, S., Malinsky, M., Linderoth, T., Sawasawa, W., Vernaz, G., Bista, I., Hickey, A., Kucka, M., … Svardal, H. (2025). Introgression dynamics of sex-linked chromosomal inversions shape the Malawi cichlid radiation. *Science (New York, N.Y.)*, *388*(6752), eadr9961.

Bomblies, K., & Peichel, C. L. (2022). Genetics of adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(30), e2122152119.

Borowsky, R. (2018). Cavefishes. *Current Biology: CB*, *28*(2), R60–R64.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo

methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), 2047-217X-2–10.

Brown, A. P., Arias-Rodriguez, L., Yee, M.-C., Tobler, M., & Kelley, J. L. (2018). Concordant Changes in Gene Expression and Nucleotides Underlie Independent Adaptation to Hydrogen-Sulfide-Rich Environments. *Genome Biology and Evolution*, *10*(11), 2867–2881.

Brown, A. P., Greenway, R., Morgan, S., Quackenbush, C. R., Giordani, L., Arias-Rodriguez, L., Tobler, M., & Kelley, J. L. (2017). Genome-scale data reveal that endemic Poecilia populations from small sulphidic springs display no evidence of inbreeding. *Molecular Ecology*, *26*(19), 4920–4934.

Brown, A. P., McGowan, K. L., Schwarzkopf, E. J., Greenway, R., Rodriguez, L. A., Tobler, M., & Kelley, J. L. (2019). Local ancestry analysis reveals genomic convergence in extremophile fishes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *374*(1777), 20180240.

Brown, M. R., Manuel Gonzalez de La Rosa, P., & Blaxter, M. (2025). tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics (Oxford, England)*, *41*(2), btaf049.

Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, *108*(10), 1880–1890.

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics*, *103*(3), 338–348.

Browning, S. R., & Browning, B. L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics*, *81*(5), 1084–1097.

Budde, M. W., & Roth, M. B. (2010). Hydrogen sulfide increases hypoxia-inducible factor-1 activity independently of von Hippel-Lindau tumor suppressor-1 in C. elegans. *Molecular Biology of the Cell*, *21*(1), 212–217.

Calvert, J. W., Jha, S., Gundewar, S., Elrod, J. W., Ramachandran, A., Pattillo, C. B., Kevil, C. G., & Lefer, D. J. (2009). Hydrogen sulfide mediates cardioprotection through Nrf2 signaling. *Circulation Research*, *105*(4), 365–374.

Calvo, S. E., Clauser, K. R., & Mootha, V. K. (2016). MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Research*, *44*(D1), D1251–D1257.

Camarillo, H., Arias Rodriguez, L., & Tobler, M. (2020). Functional consequences of phenotypic variation between locally adapted populations: Swimming performance and ventilation in extremophile fish. *Journal of Evolutionary Biology*, *33*(4), 512–523.

Cao, X., Ding, L., Xie, Z.-Z., Yang, Y., Whiteman, M., Moore, P. K., & Bian, J.-S. (2019). A Review of Hydrogen Sulfide Synthesis, Metabolism, and Measurement: Is Modulation of Hydrogen Sulfide a Novel Therapeutic for Cancer? *Antioxidants & Redox Signaling*, *31*(1), 1–38.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140.

Chaisson, M. J. P., Wilson, R. K., & Eichler, E. E. (2015). Genetic variation and the de novo assembly of human genomes. *Nature Reviews. Genetics*, *16*(11), 627–640.

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., Southwick, A. M., Absher, D. M., Grimwood, J., Schmutz, J., Myers, R. M., Petrov, D., Jónsson, B., Schluter, D., Bell, M. A., & Kingsley, D. M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science (New York, N.Y.)*, *327*(5963), 302–305.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), s13742-015-0047–0048.

Chaturvedi, A., Zhou, J., Raeymaekers, J. A. M., Czypionka, T., Orsini, L., Jackson, C. E., Spanier, K. I., Shaw, J. R., Colbourne, J. K., & De Meester, L. (2021). Extensive standing genetic variation from a small number of founders enables rapid adaptation in Daphnia. *Nature Communications*, *12*(1), 4306.

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *IMeta*, *2*(2), e107.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)*, *34*(17), i884–i890.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175.

Cheng, H., Jarvis, E. D., Fedrigo, O., Koepfli, K.-P., Urban, L., Gemmell, N. J., & Li, H. (2022). Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology*, *40*(9), 1332–1335.

Christiansen, K. (1961). Convergence and Parallelism in Cave Entomobryinae. *Evolution; International Journal of Organic Evolution*, *15*(3), 288.

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

Clement, M., Posada, D., & Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, *9*(10), 1657–1659.

Coffin, J. L., Kelley, J. L., Jeyasingh, P. D., & Tobler, M. (2022). Impacts of heavy metal pollution on the ionomes and transcriptomes of Western mosquitofish (Gambusia affinis). *Molecular Ecology*, *31*(5), 1527–1542.

Colosimo, P. F., Hosemann, K. E., Balabhadra, S., Villarreal, G., Jr, Dickson, M., Grimwood, J., Schmutz, J., Myers, R. M., Schluter, D., & Kingsley, D. M. (2005). Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science (New York, N.Y.)*, *307*(5717), 1928–1933.

Coombe, L., Kazemi, P., Wong, J., Birol, I., & Warren, R. L. (2024). Multi-genome synteny detection using minimizer graph mappings. *BioRxiv: The Preprint Server for Biology*, 2024.02.07.579356.

Coombe, L., Warren, R. L., & Birol, I. (2025). ntSynt-viz: Visualizing synteny patterns across multiple genomes. *BioRxiv: The Preprint Server for Biology*, 2025.01.15.633221.

Cooper, C. E., & Brown, G. C. (2008). The inhibition of mitochondrial cytochrome oxidase by the gases carbon monoxide, nitric oxide, hydrogen cyanide and hydrogen sulfide: chemical mechanism and physiological significance. *Journal of Bioenergetics and Biomembranes*, *40*(5), 533–539.

Corbett-Detig, R. B., Russell, S. L., Nielsen, R., & Losos, J. (2020). Phenotypic convergence is not mirrored at the protein level in a lizard adaptive radiation. *Molecular Biology and Evolution*, *37*(6), 1604–1614.

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13), 3133–3157.

Culver, D. C., & Pipan, T. (2009). *The Biology of Caves and Other Subterranean Habitats*. Oxford University Press.

Dainat, J. (n.d.). *AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format* (0.7.0) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3552717

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158.

Danecek, Petr, Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008.

Darolti, I., Wright, A. E., Sandkam, B. A., Morris, J., Bloch, N. I., Farré, M., Fuller, R. C., Bourne, G. R., Larkin, D. M., Breden, F., & Mank, J. E. (2019). Extreme heterogeneity in sex chromosome differentiation and dosage compensation in livebearers. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(38), 19031–19036.

Darwin, C. R. (1859). *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray.

De-Kayne, R., Perry, B. W., McGowan, K. L., Landers, J., Arias-Rodriguez, L., Greenway, R., Peña, C. M. R., Tobler, M., & Kelley, J. L. (2024). Evolutionary Rate Shifts in Coding and Regulatory Regions Underpin Repeated Adaptation to Sulfidic Streams in Poeciliid Fishes. *Genome Biology and Evolution*, *16*(5), evae087.

Derkarabetian, S., Steinmann, D. B., & Hedin, M. (2010). Repeated and time-correlated morphological convergence in cave-dwelling harvestmen (Opiliones, Laniatores) from Montane Western North America. *PloS One*, *5*(5), e10388.

Dobzhansky, T. (1937). Genetic nature of species differences. *The American Naturalist*, *71*(735), 404–420.

Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., García-Accinelli, G., Van Belleghem, S. M., Patterson, N., Neafsey, D. E., Challis, R., Kumar, S., Moreira, G. R. P., Salazar, C., Chouteau, M., Counterman, B. A., Papa, R., Blaxter, M., Reed, R. D., … Mallet, J. (2019). Genomic architecture and introgression shape a butterfly radiation. *Science*. https://doi.org/10.1126/science.aaw2090

Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M. T., Gudbjartsson, D. F., Stefansson, K., Halldorsson, B. V., & Melsted, P. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, *10*(1), 5402.

English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A., & Sedlazeck, F. J. (2022). Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biology*, *23*(1), 271.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)*, *32*(19), 3047–3048.

Feng, S., Ma, H., Yin, Y., Wan, W., Mao, K., & Ru, D. (2025). A complex interplay of genetic introgression and local adaptation during the evolutionary history of three closely related spruce species. *Plant Diversity*, *47*(4), 620–632.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, *31*(5), 1275–1291.

Feulner, P. G. D., & De-Kayne, R. (2017). Genome evolution, structural rearrangements and speciation. *Journal of Evolutionary Biology*, *30*(8), 1488–1490.

Fong, L. J. M., Darolti, I., Metzger, D. C. H., Morris, J., Lin, Y., Sandkam, B. A., & Mank, J. E. (2023). Parsimony and poeciliid sex chromosome evolution. *Genome Biology and Evolution*, *15*(9), evad128.

Fraser, B. A., Künstner, A., Reznick, D. N., Dreyer, C., & Weigel, D. (2015). Population genomics of natural and experimental populations of guppies (Poecilia reticulata). *Molecular Ecology*, *24*(2), 389–408.

Fraser, B. A., Whiting, J. R., Paris, J. R., Weadick, C. J., Parsons, P. J., Charlesworth, D., Bergero, R., Bemm, F., Hoffmann, M., Kottler, V. A., Liu, C., Dreyer, C., & Weigel, D. (2020). Improved Reference Genome Uncovers Novel Sex-

Linked Regions in the Guppy (Poecilia reticulata). *Genome Biology and Evolution*, *12*(10), 1789–1805.

Galloway, J., Cresko, W. A., & Ralph, P. (2020). A Few Stickleback Suffice for the Transport of Alleles to New Lakes. *G3 (Bethesda, Md.)*, *10*(2), 505–514.

García-Bereguiaín, M. A., Samhan-Arias, A. K., Martín-Romero, F. J., & Gutiérrez-Merino, C. (2008). Hydrogen sulfide raises cytosolic calcium in neurons through activation of L-type Ca2+ channels. *Antioxidants & Redox Signaling*, *10*(1), 31–42.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*(9), 875–879.

Gautier, M., Klassmann, A., & Vitalis, R. (2017). rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Molecular Ecology Resources*, *17*(1), 78–90.

Gautier, M., & Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics (Oxford, England)*, *28*(8), 1176–1177.

Gould, S. J. (1990). *Wonderful Life: The Burgess Shale and the Nature of History*. W. W. Norton.

Greenway, R., Arias-Rodriguez, L., Diaz, P., & Tobler, M. (2014). Patterns of Macroinvertebrate and Fish Diversity in Freshwater Sulphide Springs. *Diversity*, *6*(3), 597–632.

Greenway, R., Barts, N., Henpita, C., Brown, A. P., Rodriguez, L. A., Peña, C. M. R., Arndt, S., Lau, G. Y., Murphy, M. P., Wu, L., Lin, D., Shaw, J. H., Kelley, J. L., & Tobler, M. (2020). Convergent evolution of conserved mitochondrial pathways underlies repeated adaptation to extreme environments. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(28), 16424–16430.

Greenway, R., Brown, A. P., Camarillo, H., Delich, C., McGowan, K. L., Nelson, J., Arias-Rodriguez, L., Kelley, J. L., & Tobler, M. (2021). Convergent adaptation and ecological speciation result from unique genomic mechanisms in sympatric extremophile fishes. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/2021.06.28.450104

Greenway, R., De-Kayne, R., Brown, A. P., Camarillo, H., Delich, C., McGowan, K. L., Nelson, J., Arias-Rodriguez, L., Kelley, J. L., & Tobler, M. (2024). Integrative analyses of convergent adaptation in sympatric extremophile fishes. *Current Biology: CB*. https://doi.org/10.1016/j.cub.2024.09.027

Greenway, R., McNemee, R., Okamoto, A., Plath, M., Arias-Rodriguez, L., & Tobler, M. (2019). Correlated divergence of female and male genitalia in replicated lineages with ongoing ecological speciation. *Evolution; International Journal of Organic Evolution*, *73*(6), 1200–1212.

Haenel, Q., Roesti, M., Moser, D., MacColl, A. D. C., & Berner, D. (2019). Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish. *Evolution Letters*, *3*(1), 28–42.

Harewood, L., Chaignat, E., & Reymond, A. (2012). Structural variation and its effect on expression. *Methods in Molecular Biology (Clifton, N.J.)*, *838*, 173–186.

Hermisson, J., & Pennings, P. S. (2005). Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. *Genetics*, *169*(4), 2335–2352.

Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., Abel, H. J., Antonacci-Fulton, L. L., Asri, M., Baid, G., Baker, C. A., Belyaeva, A., Billis, K., Bourque, G., Buonaiuto, S., Carroll, A., Chaisson, M. J. P., Chang, P.-C., Chang, X. H., … Human Pangenome Reference Consortium. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, *42*(4), 663–673.

Hildebrandt, T. M., & Grieshaber, M. K. (2008). Three enzymatic activities catalyze the oxidation of sulfide to thiosulfate in mammalian and invertebrate mitochondria. *The FEBS Journal*, *275*(13), 3352–3361.

Hilgers, L., Liu, S., Jensen, A., Brown, T., Cousins, T., Schweiger, R., Guschanski, K., & Hiller, M. (2024). Avoidable false PSMC population size peaks occur across numerous studies. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/2024.06.17.599025

Hill, B. C., Woon, T. C., Nicholls, P., Peterson, J., Greenwood, C., & Thomson, A. J. (1984). Interactions of sulphide and other ligands with cytochrome c oxidase. An electron-paramagnetic-resonance study. *The Biochemical Journal*, *224*(2), 591–600.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522.

Hotaling, S., Quackenbush, C. R., Bennett-Ponsford, J., New, D. D., Arias-Rodriguez, L., Tobler, M., & Kelley, J. L. (2019). Bacterial Diversity in Replicated Hydrogen Sulfide-Rich Streams. *Microbial Ecology*, *77*(3), 559–573.

Ishikawa, A., Stuart, Y. E., Bolnick, D. I., & Kitano, J. (2021). Copy number variation of a fatty acid desaturase gene Fads2 associated with ecological divergence in freshwater stickleback populations. *Biology Letters*, *17*(8), 20210204.

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, *8*(1), 14061.

Jiang, J., Chan, A., Ali, S., Saha, A., Haushalter, K. J., Lam, W.-L. M., Glasheen, M., Parker, J., Brenner, M., Mahon, S. B., Patel, H. H., Ambasudhan, R., Lipton,

S. A., Pilz, R. B., & Boss, G. R. (2016). Hydrogen Sulfide—Mechanisms of Toxicity and Development of an Antidote. *Scientific Reports*, *6*(1), 20831.

Jilani, M., Turcan, A., Haspel, N., & Jagodzinski, F. (2022). Elucidating the structural impacts of protein InDels. *Biomolecules*, *12*(10), 1435.

Joly, S., McLenachan, P. A., & Lockhart, P. J. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*, *174*(2), E54-70.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61.

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., Whibley, A., Becuwe, M., Baxter, S. W., Ferguson, L., Wilkinson, P. A., Salazar, C., Davidson, C., Clark, R., Quail, M. A., Beasley, H., Glithero, R., Lloyd, C., Sims, S., … ffrench-Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, *477*(7363), 203–206.

Jourdan, J., Bierbach, D., Riesch, R., Schießl, A., Wigh, A., Arias-Rodriguez, L., Indy, J. R., Klaus, S., Zimmer, C., & Plath, M. (2014). Microhabitat use,

population densities, and size distributions of sulfur cave-dwelling Poecilia mexicana. *PeerJ*, *2*, e490.

Jourdan, J., Krause, S. T., Lazar, V. M., Zimmer, C., Sommer-Trembo, C., Arias-Rodriguez, L., Klaus, S., Riesch, R., & Plath, M. (2016). Shared and unique patterns of phenotypic diversification along a stream gradient in two congeneric species. *Scientific Reports*, *6*(1), 38971.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Kabil, O., Motl, N., & Banerjee, R. (2014). H2S and its role in redox signaling. *Biochimica et Biophysica Acta. Proteins and Proteomics*, *1844*(8), 1355–1366.

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Kelley, J. L., Arias-Rodriguez, L., Patacsil Martin, D., Yee, M.-C., Bustamante, C. D., & Tobler, M. (2016). Mechanisms Underlying Adaptation to Life in Hydrogen Sulfide–Rich Environments. *Molecular Biology and Evolution*, *33*(6), 1419–1434.

Kelley, J. L., Desvignes, T., McGowan, K. L., Perez, M., Rodriguez, L. A., Brown, A. P., Culumber, Z., & Tobler, M. (2021). microRNA expression variation as a potential molecular mechanism contributing to adaptation to hydrogen sulphide. *Journal of Evolutionary Biology*, *34*(6), 977–988.

Kelley, J. L., Tobler, M., Beck, D., Sadler-Riggleman, I., Quackenbush, C. R., Arias Rodriguez, L., & Skinner, M. K. (2021). Epigenetic inheritance of DNA methylation changes in fish living in hydrogen sulfide–rich springs. *Proceedings of the National Academy of Sciences*, *118*(26), e2014929118.

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915.

Kim, K.-W., De-Kayne, R., Gordon, I. J., Omufwoko, K. S., Martins, D. J., ffrench-Constant, R., & Martin, S. H. (2022). Stepwise evolution of a butterfly supergene via duplication and inversion. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *377*(1856), 20210207.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, *217*(5129), 624–626.

Kingman, G. A. R., Vyas, D. N., Jones, F. C., Brady, S. D., Chen, H. I., Reid, K., Milhaven, M., Bertino, T. S., Aguirre, W. E., Heins, D. C., Hippel, F. A. von, Park, P. J., Kirch, M., Absher, D. M., Myers, R. M., Palma, F. D., Bell, M. A., Kingsley, D. M., & Veeramah, K. R. (2021). Predicting future from past: The

genomic basis of recurrent and rapid stickleback evolution. *Science Advances*. https://doi.org/10.1126/sciadv.abg5285

Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Battle, A., Aganezov, S., & Schatz, M. C. (2023). Jasmine and Iris: population-scale structural variant comparison and analysis. *Nature Methods*, *20*(3), 408–417.

Kirubakaran, T. G., Grove, H., Kent, M. P., Sandve, S. R., Baranski, M., Nome, T., De Rosa, M. C., Righino, B., Johansen, T., Otterå, H., Sonesson, A., Lien, S., & Andersen, Ø. (2016). Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*, *25*(10), 2130–2143.

Kitano, J., Ishikawa, A., Ravinet, M., & Courtier-Orgogozo, V. (2022). Genetic basis of speciation and adaptation: from loci to causative mutations. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *377*(1855), 20200503.

Korunes, K. L., & Samuk, K. (2021). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*(4), 1359–1368.

Kottler, V. A., Feron, R., Nanda, I., Klopp, C., Du, K., Kneitz, S., Helmprobst, F., Lamatsch, D. K., Lopez-Roques, C., Lluch, J., Journot, L., Parrinello, H., Guiguen, Y., & Schartl, M. (2020). Independent origin of XY and ZW sex determination mechanisms in mosquitofish sister species. *Genetics*, *214*(1), 193–209.

Künstner, A., Hoffmann, M., Fraser, B. A., Kottler, V. A., Sharma, E., Weigel, D., & Dreyer, C. (2016). The Genome of the Trinidadian Guppy, Poecilia reticulata, and Variation in the Guanapo Population. *PloS One*, *11*(12), e0169087.

Küpper, C., Stocks, M., Risse, J. E., dos Remedios, N., Farrell, L. L., McRae, S. B., Morgan, T. C., Karlionova, N., Pinchuk, P., Verkuil, Y. I., Kitaysky, A. S., Wingfield, J. C., Piersma, T., Zeng, K., Slate, J., Blaxter, M., Lank, D. B., & Burke, T. (2016). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics*, *48*(1), 79–83.

Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E. V., & Zdobnov, E. M. (2023). OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research*, *51*(D1), D445–D451.

Kwakye, A., Reid, K., Wund, M. A., Heins, D. C., Bell, M. A., & Veeramah, K. R. (2025). Rare "jackpot" individuals drive rapid adaptation in Threespine Stickleback. In *bioRxivorg* (p. 2025.03.25.642177). https://doi.org/10.1101/2025.03.25.642177

Lagoutte, E., Mimoun, S., Andriamihaja, M., Chaumontet, C., Blachier, F., & Bouillaud, F. (2010). Oxidation of hydrogen sulfide remains a priority in mammalian cells and causes reverse electron transfer in colonocytes. *Biochimica et Biophysica Acta*, *1797*(8), 1500–1511.

Lai, Y.-T., Yeung, C. K. L., Omland, K. E., Pang, E.-L., Hao, Y., Liao, B.-Y., Cao, H.-F., Zhang, B.-W., Yeh, C.-F., Hung, C.-M., Hung, H.-Y., Yang, M.-Y.,

Liang, W., Hsu, Y.-C., Yao, C.-T., Dong, L., Lin, K., & Li, S.-H. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proceedings of the National Academy of Sciences*, *116*(6), 2152–2157.

Lamichhaney, S., Han, F., Berglund, J., Wang, C., Almén, M. S., Webster, M. T., Grant, B. R., Grant, P. R., & Andersson, L. (2016). A beak size locus in Darwin's finches facilitated character displacement during a drought. *Science (New York, N.Y.)*, *352*(6284), 470–474.

Lande, R., & Shannon, S. (1996). The role of genetic variation in adaptation and population persistence in a changing environment. *Evolution; International Journal of Organic Evolution*, *50*(1), 434.

Landry, A. P., Ballou, D. P., & Banerjee, R. (2017). H2S oxidation by nanodisc-embedded human sulfide quinone oxidoreductase. *The Journal of Biological Chemistry*, *292*(28), 11641–11649.

Landry, A. P., Ballou, D. P., & Banerjee, R. (2021). Hydrogen sulfide oxidation by sulfide quinone oxidoreductase. *Chembiochem: A European Journal of Chemical Biology*, *22*(6), 949–960.

Lars Gabriel, Tomáš Brůna, Katharina J. Hoff, Matthis Ebel, Alexandre Lomsadze, Mark Borodovsky, & Mario Stanke. (2023). BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *BioRxiv: The Preprint Server for Biology*, 2023.06.10.544449.

Lenski, R., Rose, M., Simpson, S. C., & Tadler, S. (1991). Long-term experimental evolution in Escherichia coli. I. adaptation and divergence during 2,000 generations. *The American Naturalist*, *138*(6), 1315–1341.

Lewis, J. A., Kandala, P., Penley, M. J., & Morran, L. T. (2023). Gene flow accelerates adaptation to a parasite. *Evolution; International Journal of Organic Evolution*, *77*(6), 1468–1478.

Li, F., Qiao, Z., Duan, Q., & Nevo, E. (2021). Adaptation of mammals to hypoxia. *Animal Models and Experimental Medicine*, *4*(4), 311–318.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997 [q-Bio]*. http://arxiv.org/abs/1303.3997

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, *34*(18), 3094–3100.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496.

Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, *21*(1), 265.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079.

Li, L., Rose, P., & Moore, P. K. (2011). Hydrogen Sulfide and Cell Signaling. *Annual Review of Pharmacology and Toxicology*, *51*(1), 169–187.

Libiad, M., Yadav, P. K., Vitvitsky, V., Martinov, M., & Banerjee, R. (2014). Organization of the Human Mitochondrial Hydrogen Sulfide Oxidation Pathway *♦. *The Journal of Biological Chemistry*, *289*(45), 30901–30910.

Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., & Guo, J.-T. (2017). Effects of short indels on protein structure and function in human genomes. *Scientific Reports*, *7*(1), 9313.

Lin, Y., Darolti, I., Bijl, W. van der, Morris, J., & Mank, J. E. (2023). Extensive variation in germline de novo mutations in Poecilia reticulata. *Genome Research*, *33*(8), 1317–1324.

Linnen, C. R., Kingsley, E. P., Jensen, J. D., & Hoekstra, H. E. (2009). On the Origin and Spread of an Adaptive Allele in Deer Mice. *Science*, *325*(5944), 1095–1098.

Liu, Y., Cotton, J. A., Shen, B., Han, X., Rossiter, S. J., & Zhang, S. (2010). Convergent sequence evolution between echolocating bats and dolphins. *Current Biology: CB*, *20*(2), R53–R54.

Liu, Z., Qi, F.-Y., Zhou, X., Ren, H.-Q., & Shi, P. (2014). Parallel Sites Implicate Functional Convergence of the Hearing Gene Prestin among Echolocating Mammals. *Molecular Biology and Evolution*, *31*(9), 2415–2424.

Losos, J. B., Jackman, T. R., Larson, A., Queiroz, K., & Rodriguez-Schettino, L. (1998). Contingency and determinism in replicated adaptive radiations of island lizards. *Science (New York, N.Y.)*, *279*(5359), 2115–2118.

Losos, Jonathan B. (2011). Convergence, Adaptation, and Constraint. *Evolution; International Journal of Organic Evolution*, *65*(7), 1827–1840.

Losos, Jonathan B. (2017). *Improbable destinies*. Riverhead Books.

Lyčka, M., Bubeník, M., Závodník, M., Peska, V., Fajkus, P., Demko, M., Fajkus, J., & Fojtová, M. (2024). TeloBase: a community-curated database of telomere sequences across the tree of life. *Nucleic Acids Research*, *52*(D1), D311–D321.

Maček, I., Vodnik, D., Pfanz, H., Low-Décarie, E., & Dumbrell, A. J. (2016). Locally extreme environments as natural long-term experiments in ecology. In *Advances in Ecological Research* (Vol. 55, pp. 283–323). Elsevier.

Mahalanobis, P. C. (1936). On the Generalized Distance in Statistics. Journal of Genetics. *Journal of Genetics*, *41*, 159–193.

Mahler, D. L., Ingram, T., Revell, L. J., & Losos, J. B. (2013). Exceptional convergence on the macroevolutionary landscape in island lizard radiations. *Science (New York, N.Y.)*, *341*(6143), 292–295.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, *20*(1), 246.

Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, *21*(2), 584–595.

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

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654.

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)*, *27*(6), 764–770.

Martin, A., & Orgogozo, V. (2013). THE LOCI OF REPEATED EVOLUTION: A CATALOG OF GENETIC HOTSPOTS OF PHENOTYPIC VARIATION. *Evolution; International Journal of Organic Evolution*, *67*(5), 1235–1250.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12.

Mayr, C. (2019). What Are 3' UTRs Doing? *Cold Spring Harbor Perspectives in Biology*, *11*(10), a034728.

McClintock, B. (1931). *Cytological observations of deficiencies involving known genes, translocations and an inversion in Zea mays*. University of Missouri, College of Agriculture, Agricultural Experiment Station.

McGowan, K. L., Passow, C. N., Arias-Rodriguez, L., Tobler, M., & Kelley, J. L. (2019). Expression analyses of cave mollies (Poecilia mexicana) reveal key genes involved in the early evolution of eye regression. *Biology Letters*, *15*(10), 20190554.

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, *8*(1), 14363.

Mérot, C., Llaurens, V., Normandeau, E., Bernatchez, L., & Wellenreuther, M. (2020). Balancing selection via life-history trade-offs maintains an inversion polymorphism in a seaweed fly. *Nature Communications*, *11*(1), 670.

Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends in Ecology & Evolution*, *35*(7), 561–572.

Mérot, C., Stenløkk, K. S. R., Venney, C., Laporte, M., Moser, M., Normandeau, E., Árnyasi, M., Kent, M., Rougeux, C., Flynn, J. M., Lien, S., & Bernatchez, L. (2023). Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (Coregonus sp.) with long and short reads. *Molecular Ecology*, *32*(6), 1458–1477.

Millero, F. J. (1986). The thermodynamics and kinetics of the hydrogen sulfide system in natural waters. *Marine Chemistry*, *18*(2), 121–147.

Minh, B. Q., Nguyen, M. A. T., & von Haeseler, A. (2013). Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution*, *30*(5), 1188–1195.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534.

Morales, A. E., Burbrink, F. T., Segall, M., Meza, M., Munegowda, C., Webala, P. W., Patterson, B. D., Thong, V. D., Ruedi, M., Hiller, M., & Simmons, N. B. (2024). Distinct genes with similar functions underlie convergent evolution in Myotis bat ecomorphs. *Molecular Biology and Evolution*, *41*(9), msae165.

Moran, B. M., Payne, C. Y., Powell, D. L., Iverson, E. N. K., Donny, A. E., Banerjee, S. M., Langdon, Q. K., Gunn, T. R., Rodriguez-Soto, R. A., Madero, A., Baczenas, J. J., Kleczko, K. M., Liu, F., Matney, R., Singhal, K., Leib, R. D., Hernandez-Perez, O., Corbett-Detig, R., Frydman, J., … Schumer, M. (2024). A lethal mitonuclear incompatibility in complex I of natural hybrids. *Nature*, *626*(7997), 119–127.

Moran, R. L., Radcliffe, W. V., Dye, M., Kim, D., & Muralidhar, P. L. (2025). Sex chromosome turnover and mitonuclear conflict drive reproductive isolation. In *bioRxiv* (p. 2025.08.29.673102). https://doi.org/10.1101/2025.08.29.673102

Muller, H. J. (1942). Isolating mechanisms, evolution, and temperature. *Biol. Symp.* https://cir.nii.ac.jp/crid/1370290617850391810

Mustafa, A. K., Gadalla, M. M., Sen, N., Kim, S., Mu, W., Gazi, S. K., Barrow, R. K., Yang, G., Wang, R., & Snyder, S. H. (2009). H2S signals through protein S-sulfhydration. *Science Signaling*, *2*(96), ra72.

Natarajan, C., Hoffmann, F. G., Weber, R. E., Fago, A., Witt, C. C., & Storz, J. F. (2016). Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. *Science*, *354*(6310), 336–339.

Nei M, & Li W H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, *76*(10), 5269–5273.

Nosil, P. (2012). *Ecological speciation*. https://toc.library.ethz.ch/objects/pdf/e05_978-0-19-958710-0_01.pdf

Olson, K. R. (2018). H2S and polysulfide metabolism: Conventional and unconventional pathways. *Biochemical Pharmacology*, *149*, 77–90.

Ono, K., Akaike, T., Sawa, T., Kumagai, Y., Wink, D. A., Tantillo, D. J., Hobbs, A. J., Nagy, P., Xian, M., Lin, J., & Fukuto, J. M. (2014). Redox chemistry and chemical biology of H2S, hydropersulfides, and derived species: implications of their possible biological activity and utility. *Free Radical Biology & Medicine*, *77*, 82–94.

Orr, H. A., & Betancourt, A. J. (2001). Haldane's sieve and adaptation from the standing genetic variation. *Genetics*, *157*(2), 875–884.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M.

B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, *20*(1), 275.

Pagliarini, D. J., Calvo, S. E., Chang, B., Sheth, S. A., Vafai, S. B., Ong, S.-E., Walford, G. A., Sugiana, C., Boneh, A., Chen, W. K., Hill, D. E., Vidal, M., Evans, J. G., Thorburn, D. R., Carr, S. A., & Mootha, V. K. (2008). A mitochondrial protein compendium elucidates complex I disease biology. *Cell*, *134*(1), 112–123.

Palacios, M., Arias-Rodriguez, L., Plath, M., Eifert, C., Lerp, H., Lamboj, A., Voelker, G., & Tobler, M. (2013). The Rediscovery of a Long Described Species Reveals Additional Complexity in Speciation Patterns of Poeciliid Fishes in Sulfide Springs. *PloS One*, *8*(8), e71069.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics (Oxford, England)*, *26*(3), 419–420.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)*, *35*(3), 526–528.

Passow, C. N., Arias-Rodriguez, L., & Tobler, M. (2017). Convergent evolution of reduced energy demands in extremophile fish. *PloS One*, *12*(10), e0186935.

Passow, C. N., Brown, A. P., Arias-Rodriguez, L., Yee, M.-C., Sockell, A., Schartl, M., Warren, W. C., Bustamante, C., Kelley, J. L., & Tobler, M. (2017). Complexities of gene expression patterns in natural populations of an

extremophile fish (Poecilia mexicana, Poeciliidae). *Molecular Ecology*, *26*(16), 4211–4225.

Passow, C. N., Henpita, C., Shaw, J. H., Quackenbush, C. R., Warren, W. C., Schartl, M., Arias-Rodriguez, L., Kelley, J. L., & Tobler, M. (2017). The roles of plasticity and evolutionary change in shaping gene expression variation in natural populations of extremophile fish. *Molecular Ecology*, *26*(22), 6384–6399.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient Admixture in Human History. *Genetics*, *192*(3), 1065–1093.

Pease, J. B., & Hahn, M. W. (2015). Detection and polarization of introgression in a five-taxon phylogeny. *Systematic Biology*, *64*(4), 651–662.

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics (Oxford, England)*, *34*(5), 867–868.

Peers, C., Bauer, C. C., Boyle, J. P., Scragg, J. L., & Dallas, M. L. (2012). Modulation of Ion Channels by Hydrogen Sulfide. *Antioxidants & Redox Signaling*, *17*(1), 95–105.

Perry, B. W., McGowan, K. L., Arias-Rodriguez, L., Duttke, S. H., Tobler, M., & Kelley, J. L. (2024). Nascent transcription reveals regulatory changes in extremophile fishes inhabiting hydrogen sulfide-rich environments. *Proceedings. Biological Sciences*, *291*(2025), 20240412.

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare [version 1; peer review: 3 approved]. *F1000Research*, *9*(304). https://doi.org/10.12688/f1000research.23297.1

Pfenninger, M., Lerp, H., Tobler, M., Passow, C., Kelley, J. L., Funke, E., Greshake, B., Erkoc, U. K., Berberich, T., & Plath, M. (2014). Parallel evolution of cox genes in H2S-tolerant fish as key adaptation to a toxic environment. *Nature Communications*, *5*(1), 3873.

Pfenninger, M., Patel, S., Arias-Rodriguez, L., Feldmeyer, B., Riesch, R., & Plath, M. (2015). Unique evolutionary trajectories in repeated adaptation to hydrogen sulphide-toxic habitats of a neotropical fish (Poecilia mexicana). *Molecular Ecology*, *24*(21), 5446–5459.

Pietri, R., Román-Morales, E., & López-Garriga, J. (2011). Hydrogen sulfide and hemeproteins: knowledge and mysteries. *Antioxidants & Redox Signaling*, *15*(2), 393–404.

Plath, M., Pfenninger, M., Lerp, H., Riesch, R., Eschenbrenner, C., Slattery, P. A., Bierbach, D., Herrmann, N., Schulte, M., Arias-Rodriguez, L., Rimber Indy, J., Passow, C., & Tobler, M. (2013a). Genetic differentiation and selection against migrants in evolutionarily replicated extreme environments. *Evolution; International Journal of Organic Evolution*, *67*(9), 2647–2661.

Plath, M., Pfenninger, M., Lerp, H., Riesch, R., Eschenbrenner, C., Slattery, P. A., Bierbach, D., Herrmann, N., Schulte, M., Arias-Rodriguez, L., Rimber Indy, J., Passow, C., & Tobler, M. (2013b). Genetic differentiation and selection

against migrants in evolutionarily replicated extreme environments: Genetic differentiation and selection against migrants. *Evolution; International Journal of Organic Evolution*, *67*(9), 2647–2661.

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, *2*, e431.

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014). Demystifying the RAD fad. *Molecular Ecology*, *23*(24), 5937–5942.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842.

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1432.

Rappaport, H. B., & Oliverio, A. M. (2024). Lessons from Extremophiles: Functional Adaptations and Genomic Innovations across the Eukaryotic Tree of Life. *Genome Biology and Evolution*, *16*(8), evae160.

Reid, K., Bell, M. A., & Veeramah, K. R. (2021). Threespine Stickleback: A Model System For Evolutionary Genomics. *Annual Review of Genomics and Human Genetics*, *22*(1), 357–383.

Reid, N. M., Proestou, D. A., Clark, B. W., Warren, W. C., Colbourne, J. K., Shaw, J. R., Karchner, S. I., Hahn, M. E., Nacci, D., Oleksiak, M. F., Crawford, D. L.,

& Whitehead, A. (2016). The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science*, *354*(6317), 1305–1308.

Riesch, R., Duwe, V., Herrmann, N., Padur, L., Ramm, A., Scharnweber, K., Schulte, M., Schulz-Mirbach, T., Ziege, M., & Plath, M. (2009). Variation along the shy–bold continuum in extremophile fishes (Poecilia mexicana, Poecilia sulphuraria). *Behavioral Ecology and Sociobiology*, *63*(10), 1515–1526.

Riesch, R., Plath, M., García de León, F. J., & Schlupp, I. (2010). Convergent life-history shifts: toxic environments result in big babies in two clades of poeciliids. *The Science of Nature*, *97*(2), 133–141.

Riesch, R., Plath, M., Schlupp, I., Tobler, M., & Brian Langerhans, R. (2014). Colonisation of toxic environments drives predictable life-history evolution in livebearing fishes (Poeciliidae). *Ecology Letters*, *17*(1), 65–71.

Riesch, R., Tobler, M., Lerp, H., Jourdan, J., Doumas, T., Nosil, P., Langerhans, R. B., & Plath, M. (2016). Extremophile Poeciliidae: multivariate insights into the complexity of speciation along replicated ecological gradients. *BMC Evolutionary Biology*, *16*(1), 136.

Riesch, R., Tobler, M., & Plath, M. (Eds.). (2015). *Extremophile Fishes: Ecology, Evolution, and Physiology of Teleosts in Extreme Environments*. Springer International Publishing.

Robles, N. V., Moran, B. M., Rodríguez Barrera, M. J., Jofre, G. I., Gunn, T., Iverson, E. N. K., Beskid, S., Baczenas, J. J., Sedghifar, A., Andolfatto, P.,

Powell, D. L., Brandvain, Y., Havird, J. C., Rosenthal, G. G., & Schumer, M. (2025). Admixture mapping reveals evidence for multiple mitonuclear incompatibilities in swordtail fish hybrids. *Molecular Ecology*, *e70106*. https://doi.org/10.1111/mec.70106

Rosenblum, E. B., Parent Christine E., & Brandt, E. E. (2014). The Molecular Basis of Phenotypic Convergence. *Annual Review of Ecology, Evolution, and Systematics*, *45*, 203–226.

Rossi, M., Hausmann, A. E., Alcami, P., Moest, M., Roussou, R., Van Belleghem, S. M., Wright, D. S., Kuo, C.-Y., Lozano-Urrego, D., Maulana, A., Melo-Flórez, L., Rueda-Muñoz, G., McMahon, S., Linares, M., Osman, C., McMillan, W. O., Pardo-Diaz, C., Salazar, C., & Merrill, R. M. (2024). Adaptive introgression of a visual preference gene. *Science (New York, N.Y.)*, *383*(6689), 1368–1373.

Rossiter, S. J., Zhang, S., & Liu, Y. (2011). Prestin and high frequency hearing in mammals. *Communicative & Integrative Biology*, *4*(2), 236–239.

Rudman, S. M., Greenblum, S. I., Rajpurohit, S., Betancourt, N. J., Hanna, J., Tilk, S., Yokoyama, T., Petrov, D. A., & Schmidt, P. (2022). Direct observation of adaptive tracking on ecological time scales in Drosophila. *Science (New York, N.Y.)*, *375*(6586), eabj7484.

Ryan, K., De-Kayne, R., Davis, J., Arias-Rodriguez, L., Tobler, M., & Kelley, J. L. (2025). New genome assemblies for Poeciliidae: A foundation for adaptation studies. *Genome Biology and Evolution*, *17*(6), evaf111.

Ryan, K., Greenway, R., Landers, J., Arias-Rodriguez, L., Tobler, M., & Kelley, J. L. (2023). Selection on standing genetic variation mediates convergent evolution in extremophile fish. *Molecular Ecology*, *32*(18), 5042–5054.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., & Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918.

Salmón, P., Jacobs, A., Ahrén, D., Biard, C., Dingemanse, N. J., Dominoni, D. M., Helm, B., Lundberg, M., Senar, J. C., Sprau, P., Visser, M. E., & Isaksson, C. (2021). Continent-wide genomic signatures of adaptation to urbanisation in a songbird across Europe. *Nature Communications*, *12*(1), 2983.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., & Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, *507*(7492), 354–357.

Saunders, C. T., Holt, J. M., Baker, D. N., Lake, J. A., Belyeu, J. R., Kronenberg, Z., Rowell, W. J., & Eberle, M. A. (2025). Sawfish: improving long-read structural variant discovery and genotyping with local haplotype modeling. *Bioinformatics (Oxford, England)*, *41*(4), btaf136.

Schluter, D., & Conte, G. L. (2009). Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(Supplement 1), 9955–9962.

Schulz-Mirbach, T., Eifert, C., Riesch, R., Farnworth, M. S., Zimmer, C., Bierbach, D., Klaus, S., Tobler, M., Streit, B., Indy, J. R., Arias-Rodriguez, L., & Plath, M. (2016). Toxic hydrogen sulphide shapes brain anatomy: a comparative study of sulphide-adapted ecotypes in the Poecilia mexicana complex. *Journal of Zoology (London, England: 1987)*, *300*(3), 163–176.

Scott, A. J., Chiang, C., & Hall, I. M. (2021). Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *BioRxiv*, *31*, 2249–2257.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468.

Selz, O. M., & Seehausen, O. (2019). Interspecific hybridization can generate functional novelty in cichlid fish. *Proceedings. Biological Sciences*, *286*(1913), 20191621.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504.

Shaul, O. (2017). How introns enhance gene expression. *The International Journal of Biochemistry & Cell Biology*, *91*(Pt B), 145–155.

Shen, X., Song, S., Li, C., & Zhang, J. (2022). Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature*, *606*(7915), 725–731.

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science (New York, N.Y.)*, *236*(4803), 787–792.

Smolka, M., Paulin, L. F., Grochowski, C. M., Horner, D. W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., Scholz, S. W., Carvalho, C. M. B., Proukakis, C., & Sedlazeck, F. J. (2024). Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology*, *42*(10), 1571–1580.

Sommer-Trembo, C., Bierbach, D., Arias-Rodriguez, L., Verel, Y., Jourdan, J., Zimmer, C., Riesch, R., Streit, B., & Plath, M. (2016). Does personality affect premating isolation between locally-adapted populations? *BMC Evolutionary Biology*, *16*(1), 138.

Stern, D. L. (2013). The genetic causes of convergent evolution. *Nature Reviews. Genetics*, *14*(11), 751–764.

Stipanuk, M. H., & Ueki, I. (2011). Dealing with methionine/homocysteine sulfur: cysteine metabolism to taurine and inorganic sulfur. *Journal of Inherited Metabolic Disease*, *34*(1), 17–32.

Stuart, K. C., Edwards, R. J., Sherwin, W. B., & Rollins, L. A. (2023). Contrasting Patterns of Single Nucleotide Polymorphisms and Structural Variation Across Multiple Invasions. *Molecular Biology and Evolution*, *40*(3), msad046.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *The Journal of Experimental Zoology*, *14*(1), 43–59.

Svardal, H., Quah, F. X., Malinsky, M., Ngatunga, B. P., Miska, E. A., Salzburger, W., Genner, M. J., Turner, G. F., & Durbin, R. (2020). Ancestral hybridization facilitated species diversification in the Lake Malawi cichlid fish adaptive radiation. *Molecular Biology and Evolution*, *37*(4), 1100–1113.

Switzer, C. H., Kasamatsu, S., Ihara, H., & Eaton, P. (2023). SOD1 is an essential H2S detoxifying enzyme. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(3), e2205044120.

Szabo, C. (2018). A timeline of hydrogen sulfide (H2S) research: from environmental toxin to biological mediator. *Biochemical Pharmacology*, *149*, 5–19.

Szpiech, Z. A. (2024). Selscan 2.0: Scanning for sweeps in unphased data. *Bioinformatics (Oxford, England)*, *40*(1), btae006.

Szpiech, Z. A., Novak, T. E., Bailey, N. P., & Stevison, L. S. (2021). Application of a novel haplotype-based scan for local adaptation to study high-altitude adaptation in rhesus macaques. *Evolution Letters*, *5*(4), 408–421.

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics (Oxford, England)*, *31*(12), 2032–2034.

Templeton, A. R., Crandall, K. A., & Sing, C. F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction

endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, *132*(2), 619–633.

Thorstensen, M. J., Euclide, P. T., Jeffrey, J. D., Shi, Y., Treberg, J. R., Watkinson, D. A., Enders, E. C., Larson, W. A., Kobayashi, Y., & Jeffries, K. M. (2022). A chromosomal inversion may facilitate adaptation despite periodic gene flow in a freshwater fish. *Ecology and Evolution*, *12*(5), e8898.

Tobler, M., & Hastings, L. (2011). Convergent Patterns of Body Shape Differentiation in Four Different Clades of Poeciliid Fishes Inhabiting Sulfide Springs. *Evolutionary Biology*, *38*(4), 412–421.

Tobler, M., Henpita, C., Bassett, B., Kelley, J. L., & Shaw, J. H. (2014). H2S exposure elicits differential expression of candidate genes in fish adapted to sulfidic and non-sulfidic environments. *Comparative Biochemistry and Physiology. Part A, Molecular & Integrative Physiology*, *175*, 7–14.

Tobler, M., Kelley, J. L., Plath, M., & Riesch, R. (2018). Extreme environments and the origins of biodiversity: Adaptation and speciation in sulphide spring fishes. *Molecular Ecology*, *27*(4), 843–859.

Tobler, M., Palacios, M., Chapman, L. J., Mitrofanov, I., Bierbach, D., Plath, M., Arias-Rodriguez, L., García de León, F. J., & Mateos, M. (2011). Evolution in Extreme Environments: Replicated Phenotypic Differentiation in Livebearing Fish Inhabiting Sulfidic Springs. *Evolution; International Journal of Organic Evolution*, *65*(8), 2213–2228.

Tobler, M., Passow, C. N., Greenway, R., Kelley, J. L., & Shaw, J. H. (2016). The Evolutionary Ecology of Animals Inhabiting Hydrogen Sulfide–Rich Environments. *Annual Review of Ecology, Evolution, and Systematics*, *47*(1), 239–262.

Tobler, M., Scharnweber, K., Greenway, R., Passow, C. N., Arias-Rodriguez, L., & García-De-León, F. J. (2015). Convergent changes in the trophic ecology of extremophile fish along replicated environmental gradients. *Freshwater Biology*, *60*(4), 768–780.

Uliano-Silva, M., Ferreira, J. G. R. N., Krasheninnikova, K., Darwin Tree of Life Consortium, Formenti, G., Abueg, L., Torrance, J., Myers, E. W., Durbin, R., Blaxter, M., & McCarthy, S. A. (2023). MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics*, *24*(1), 288.

Van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the Cloud – Using Docker, GATK, and WDL in Terra*. O'Reilly Media, Inc.

van Kruistum, H., Guernsey, M. W., Baker, J. C., Kloet, S. L., Groenen, M. A. M., Pollux, B. J. A., & Megens, H.-J. (2020). The Genomes of the Livebearing Fish Species Poeciliopsis retropinna and Poeciliopsis turrubarensis Reflect Their Different Reproductive Strategies. *Molecular Biology and Evolution*, *37*(5), 1376–1386.

Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019, May). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. *2019 IEEE*

International Parallel and Distributed Processing Symposium (IPDPS). 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), Rio de Janeiro, Brazil. https://doi.org/10.1109/ipdps.2019.00041

vcfbub: use variant nesting information to flter overlapping sites from vg deconstruct output. (n.d.). Github. Retrieved July 19, 2025, from https://github.com/pangenome/vcfbub

Verity, R., Collins, C., Card, D. C., Schaal, S. M., Wang, L., & Lotterhos, K. E. (2017). minotaur: A platform for the analysis and visualization of multivariate results from genome scans with R Shiny. *Molecular Ecology Resources*, *17*(1), 33–43.

Volff, J.-N., & Schartl, M. (2001). Variability of genetic sex determination in poeciliid fishes. *Genetica*, *111*(1–3), 101–110.

Wallace, J. L., & Wang, R. (2015). Hydrogen sulfide-based therapeutics: exploiting a unique but ubiquitous gasotransmitter. *Nature Reviews. Drug Discovery*, *14*(5), 329–345.

Warren, W. C., García-Pérez, R., Xu, S., Lampert, K. P., Chalopin, D., Stöck, M., Loewe, L., Lu, Y., Kuderna, L., Minx, P., Montague, M. J., Tomlinson, C., Hillier, L. W., Murphy, D. N., Wang, J., Wang, Z., Garcia, C. M., Thomas, G. C. W., Volff, J.-N., … Schartl, M. (2018). Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. *Nature Ecology & Evolution*, *2*(4), 669–679.

Warrington, J. A., Nair, A., Mahadevappa, M., & Tsyganskaya, M. (2000). Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiological Genomics*, *2*(3), 143–147.

Waters, J. M., & McCulloch, G. A. (2021). Reinventing the wheel? Reassessing the roles of gene flow, sorting and convergence in repeated evolution. *Molecular Ecology*, *30*(17), 4162–4172.

Weinreich, D. M., Delaney, N. F., DePristo, M. A., & Hartl, D. L. (2006). Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science*, *312*(5770), 111–114.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution; International Journal of Organic Evolution*, *38*(6), 1358–1370.

Xu, S., Wang, J., Guo, Z., He, Z., & Shi, S. (2020). Genomic Convergence in the Adaptation to Extreme Environments. *Plant Communications*, *1*(6), 100117.

Yan, Z., Martin, S. H., Gotzek, D., Arsenault, S. V., Duchen, P., Helleu, Q., Riba-Grognuz, O., Hunt, B. G., Salamin, N., Shoemaker, D., Ross, K. G., & Keller, L. (2020). Evolution of a supergene that regulates a trans-species social polymorphism. *Nature Ecology & Evolution*, *4*(2), 240–249.

Zhang, F., & Broughton, R. E. (2013). Mitochondrial–Nuclear Interactions: Compensatory Evolution or Variable Functional Constraint among Vertebrate Oxidative Phosphorylation Genes? *Genome Biology and Evolution*, *5*(10), 1781–1791.

Zhang, L., Reifová, R., Halenková, Z., & Gompert, Z. (2021). How Important Are Structural Variants for Speciation? *Genes*, *12*(7), 1084.

Zhang, Y., Reynoso, Y., Reznick, D., & Wang, X. (2023). Whole Genome Assembly and Annotation of Blackstripe Livebearer Poeciliopsis prolifica. *Genome Biology and Evolution*, *15*(11), evad195.

Zong, S., Li, Y.-L., & Liu, J.-X. (2020). Genomic architecture of rapid parallel adaptation to fresh water in a wild fish. *Molecular Biology and Evolution*, *38*, 1317–1329.