5-2020

Evolutionary Inference from Admixed Genomes: Implications of Hybridization for Biodiversity Dynamics and Conservation

Tyler Chafin
University of Arkansas, Fayetteville

Citation

Chafin, T. (2020). Evolutionary Inference from Admixed Genomes: Implications of Hybridization for Biodiversity Dynamics and Conservation. *Theses and Dissertations* Retrieved from https://scholarworks.uark.edu/etd/3603

Evolutionary Inference from Admixed Genomes: Implications of Hybridization for Biodiversity Dynamics and Conservation

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Biology

by

Tyler Chafin
Southern Arkansas University
Bachelor of Science in Biology, 2012

May 2020
University of Arkansas

The dissertation is approved for recommendation to the Graduate Council.

Marlis R. Douglas, Ph.D.
Dissertation Director

Michael E. Douglas, Ph.D.
Dissertation Co-Director

Andrew J. Alverson, Ph.D.
Committee Member

Abraham E. Tucker, Ph.D.
Committee Member

ABSTRACT

Hybridization as a macroevolutionary mechanism has been historically underappreciated among vertebrate biologists. Yet, the advent and subsequent proliferation of next-generation sequencing methods has increasingly shown hybridization to be a pervasive agent influencing evolution in many branches of the Tree of Life (to include ancestral hominids). Despite this, the dynamics of hybridization with regards to speciation and extinction remain poorly understood. To this end, I here examine the role of hybridization in the context of historical divergence and contemporary decline of several threatened and endangered North American taxa, with the goal to illuminate implications of hybridization for promoting—or impeding—population persistence in a shifting adaptive landscape.

Chapter I employed population genomic approaches to examine potential effects of habitat modification on species boundary stability in co-occurring endemic fishes of the Colorado River basin (*Gila robusta* and *G. cypha*). Results showed how one potential outcome of hybridization might drive species decline: via a breakdown in selection against interspecific heterozygotes and subsequent genetic erosion of parental species.

Chapter II explored long-term contributions of hybridization in an evolutionarily recent species complex (*Gila*) using a combination of phylogenomic and phylogeographic modelling approaches. Massively parallel computational methods were developed (and so deployed) to categorize sources of phylogenetic discordance as drivers of systematic bias among a panel of species tree inference algorithms. Contrary to past evidence, we found that hypotheses of hybrid origin (excluding one notable example) were instead explained by gene-tree discordance driven by a rapid radiation.

Chapter III examined patterns of local ancestry in the endangered red wolf genome (*Canis rufus*) – a controversial taxon of a long-standing debate about the origin of the species. Analyses show how pervasive autosomal introgression served to mask signatures of prior isolation—in turn misleading analyses that led the species to be interpreted as of recent hybrid origin. Analyses also showed how recombination interacts with selection to create a non-random, structured genomic landscape of ancestries with, in the case of the red wolf, the 'original' species tree being retained only in low-recombination 'refugia' of the X chromosome.

The final three chapters present bioinformatic software that I developed for my dissertation research to facilitate molecular approaches and analyses presented in Chapters I–III. Chapter IV details an *in-silico* method for optimizing similar genomic methods as used herein (RADseq of reduced representation libraries) for other non-model organisms. Chapter V describes a method for parsing genomic datasets for elements of interest, either as a filtering mechanism for downstream analysis, or as a precursor to targeted-enrichment reduced-representation genomic sequencing. Chapter VI presents a rapid algorithm for the definition of a 'most parsimonious' set of recombinational breakpoints in genomic datasets, as a method promoting local ancestry analyses as utilized in Chapter III.

My three case studies and accompanying software promote three trajectories in modern hybridization research: How does hybridization impact short-term population persistence? How does hybridization drive macroevolutionary trends? and How do outcomes of hybridization vary in the genome? In so doing, my research promotes a deeper understanding of the role that hybridization has and will continue to play in governing the evolutionary fates of lineages at both contemporary and historic timescales.

ACKNOWLEDGEMENTS

None of the work described here would have been possible without the ongoing efforts of countless fisheries and wildlife researchers, managers, and advocates whose dedicated work has provided the foundation on which this dissertation is built, and without which these species would no longer be here to study. Special thanks in particular go to biologists of the Arizona Game and Fish Department, Colorado Division of Wildlife, Jicarilla Apache Game and Fish, Nevada Department of Wildlife, U. S. Fish and Wildlife Service, Forest Service, National Park Service, and Bureau of Reclamation, Utah Department of Natural Resources, Utah Division of Wildlife, and the Wyoming Game and Fish Department for generously providing expertise, sampling efforts, and funding over the years during and prior to this work.

Special thanks to my advisors, Drs. Michael and Marlis Douglas, for letting me waste resources on bad ideas (and some good ones). Many thanks are also due to my fellow graduate students and colleagues at the University of Arkansas for their support, in particular past lab members Drs. Max Bangs and Steve Mussmann for their guidance, and current members Bradley Martin and Zach Zbinden, for useful feedback and occasional commiseration.

This work would also not have been possible without resources and funding provided by the University of Arkansas. Funding was provided by several generous fellowships and endowments: A Distinguished Doctoral Fellowship, The Bruker Professorship in Life Sciences (Marlis Douglas), and the 21st Century Chair in Global Change Biology (Michael Douglas). Computational resources were provided by the NSF-funded cloud-computing resource JetStream (XSEDE Allocation TG-BIO160065) and the Arkansas High Performance Computing Center.

Last, but not least, I thank my family for their unconditional patience, encouragement, and support.

DEDICATION

Dedicated to Valerie, for not letting me quit.

TABLE OF CONTENTS

LIST OF PUBLISHED PAPERS

Chapter I:

Chafin TK, Douglas MR, Martin BT, Douglas ME. 2019. Hybridization drives genetic erosion in sympatric desert fishes of western North America. *Heredity*. 123:759-773

Chapter II:

Chafin TK, Douglas MR, Bangs MR, Mussmann SM, Douglas ME. 2020. Taxonomic uncertainty and the anomaly zone: Phylogenomics resolve rapid radiation and hybrid origin in a contentious species complex. *Systematic Biology*. In revision.

Chapter III:

Chafin TK, Douglas MR, Douglas ME. 2020. Discriminating homoploid hybrid speciation from secondary introgression using genome-wide local ancestry. *Molecular Biology and Evolution*. Submitted.

Chapter IV:

Chafin TK, Martin BT, Mussmann SM, Douglas MR, Douglas ME. 2018. FRAGMATIC: *in silico* locus prediction and its utility in optimizing ddRADseq projects. *Conservation Genetics Resources*. 10(3):325-328.

Chapter V:

Chafin TK, Douglas MR, Douglas ME. 2018. MrBait: Universal identification and design of targeted-enrichment capture probes. *Bioinformatics*. 34(24):4293-4296.

Chapter VI:

Chafin TK. 2020. FGTpartitioner: Parsimonious delimitation of ancestry breakpoints in large genome-wide SNP datasets. *Journal of Open Source Software.* 5(46): 2030.

INTRODUCTION

Hybridization (=gene flow between diverged lineages) has classically been considered both rare and unimportant as an evolutionary process in vertebrates (Hubbs 1955; Dowling and Secor 1997). From a biological standpoint, it was often viewed as wasted reproductive effort and thus largely antagonistic to speciation (Dobzhansky 1937; Mayr 1963). In recent years, spurned by increasing accessibility of genome-scale sequencing, hybridization has been shown to instead be relatively common in natural populations (Twyford and Ennos 2012; Abbott et al. 2013; Mallet et al. 2016; Taylor and Larson 2019). Despite this paradigm shift, theory defining the importance of hybridization as a macroevolutionary process remains under-developed (Folk et al. 2018). To this end, several overarching questions define modern trajectories in hybridization research: How does hybridization affect species response to environmental change? How does it vary phylogenetically, and what are the long-term macroevolutionary implications of this variation? and How does the genome shape outcomes of hybridization?

Q1: Hybridization and species persistence

The first trajectory concerns the role of hybridization in promoting–or impeding–species persistence. Under what circumstances is it adaptive *versus* maladaptive? How do extrinsic factors modulate this? These questions are often addressed within the context of anthropogenically-mediated hybridization, e.g. with regards to species boundary modulations in response to altered or changing environments (Grabenstein and Taylor 2018; Larson et al. 2019). Accumulating evidence highlights the role of hybridization in escalating 'adaptive potential' by expanding the pool of genetic variation from which diversifying lineages can draw (Meier et al. 2017; Grant and Grant 2019; Marques et al. 2019). Yet, hybridization is often interpreted negatively within the context of species conservation (vonHoldt et al. 2018). This is in large part due to substantial uncertainty involving how the interplay between the environment, gene flow, and selection defines hybrid outcomes.

In some circumstances, hybridization can facilitate adaptation faster than 'conventionally accepted' mechanisms (Barton 2001; Orr and Unckless 2014; Kokko et al. 2017; Marques et al. 2019), and thus an emerging role for introgression is seen as promoting population recovery after de-stabilizing events (Kanarek et al. 2014; Stelkens et al. 2014; Stewart et al. 2017). However, the capacity for hybridization to drive so-called 'evolutionary rescue' is contingent on a limited rate of environmental change (Lindsey et al. 2013), and a rate of gene flor which bolsters adaptive diversity (Fitzpatrick et al. 2016; Tomasini and Peischl 2020) without leading to total demographic replacement (e.g. Mussmann et al. 2017). Likewise, the probability of populations benefiting from hybridization depends on the fitness costs of hybridization (Buerkle et al. 2003; Owens and Samuk 2020). The outcome of hybridization is clearly context dependent, with possible positive (Hamilton and Miller 2016) or negative (Rhymer and Simberloff 1996; Muhlfeld et al. 2009) impacts with respect to net diversity.

Q2: Hybridization and macroevolutionary trends

Another outstanding question in hybridization research regards the manner by which population genetic outcomes of reticulation accumulate over phylogenetic timescales. To what degree have they contributed to extant biodiversity? Are there traits (e.g., life history) that affect the 'propensity' to hybridize? If so, does one outcome of hybridization (e.g. hybrid speciation) overshadow another (e.g. secondary introgression)? Theory again is relatively young [albeit notably less so in some taxonomic circles than others (e.g. Anderson 1948; Stebbins 1959)].

Early vertebrate zoologists envisioned two primary outcomes of hybridization as related to speciation: Fusion of lineages (e.g. reversing ‘progress’ towards biological speciation); or selection against hybridization causing a subsequent reinforcement of reproductive barriers (e.g. increasing 'progress'; Mayr 1963). While modern molecular methods have demonstrated cases of ‘speciation reversal’ by hybridization (Seehausen et al. 2008; Kearns et al. 2018), and offered a richer understanding of the process of reinforcement (Servedio and Noor 2003; Servedio et al. 2013), they have also exposed a more interesting and nuanced role for hybridization in the evolutionary theater (e.g. Abbott et al. 2013).

The view that is emerging instead suggests that ‘partial’ reproductive isolation is itself a stable evolutionary outcome (Servedio and Hermisson 2020). To suggest a potentially overly simplistic conceptual model of this: If we envision the adaptive landscape as rugged, hybridization could be viewed as a ‘query’ between peaks. Here, lineages may borrow elements of foreign genomes (e.g., horizontal gene transfer/ introgression), or colonize entirely new adaptive optima Logically, a relationship between such outcomes and relative probabilities of speciation and/or extinction (=net diversification) can be hypothesized. Maladaptive hybridization could be selected against, promoting the solidification of reproductive barriers and thereby ‘accelerating’ speciation. Entirely novel lineages may also be formed, having some intermediate or recombinant phenotype offering an ability to capitalize in dynamic adaptive space (Dittrich-Reed and Fitzpatrick 2013). Over time, this process may increase the speciation rate of the encompassing clade. Net diversification may also be guided via a manipulation of extinction probabilities: Just as a species merger drives extinction (Rhymer and Simberloff 1996), adaptive introgression acts to circumvent it (e.g. Oziolor et al. 2019).

A major barrier to understanding the prevalence of these outcomes as a large-scale driver of macroevolutionary rates, and hence interaction with trait evolution, is two-fold: In itself hybridization is difficult to conclusively detect; and the phylogenetic framework on which such questions are built generally assume that evolution is by-and-large non-reticulate. Together these impediments have not only biased our fundamental understanding of the process of speciation but limited our appreciation of what is likely a near-ubiquitous evolutionary force driving patterns of diversification throughout the Tree of Life.

Q3: Hybridization in the genome

Much progress is being made towards the detection of hybridization in phylogenies by harvesting phylogenetic signals from different regions in the genome (Payseur and Rieseberg 2016), although these often tend to overlook important signals in the quest for a 'resolved' tree or network (e.g. Hahn and Nakhleh 2016). This prompts the final question: Is the accumulation of introgression in the genome non-random? Or is retention of alien genetic material instead biased towards certain regions or features of the genome? A failure to consider variation in the outcomes of hybridization in the genome could be not only misleading, but generate downright false conclusions (e.g. Fontaine et al. 2015).

However, understanding hybrid outcomes along the genomic axis is hampered by theoretical and methodological inertia. The most common framework by which introgression is localized in the genome is based on the D-statistic (Patterson et al. 2012; Pease and Hahn 2015). This method relies on a baseline expectation of species relationships in order to find unexpectedly high similarity among lineages as evidence of gene flow. This is done on the basis of site-pattern counts, with an expectation that discordant patterns (e.g. those disagreeing with the ‘species tree’) would naturally occur stochastically; an over-abundance of any one pattern being then taken as evidence for introgression. Although numerous methods exist [e.g. based on ancestry block lengths (Hvala et al. 2018) or hidden Markov models (Liu et al. 2014)], relatively simplistic methods such as the D-statistic remain attractive due to their ability to perform with relatively information-sparse datasets, such as SNP data that have been widely-adopted for population-level studies involving non-model organisms. However, just as genomic heterogeneity (e.g. of mutation rates) has been shown to inflate the false positive rate in a similar approach for seeking ‘islands’ of divergence (Cruickshank and Hahn 2014), they may serve to invalidate this approach for localizing introgression (Blair and Ané 2019). A more appropriate framework better leverages the power of modern phylogenetic methods, as is already being done in some areas of genomic research (Pease et al. 2016; Smith et al. 2020). This would relax the assumption of character independence, and instead recognizes the genome as a series of genealogies, within which characters (=nucleotides) are correlated. Although delimiting such regions is problematic (Springer and Gatesy 2018), doing so provides a necessary advantage by allowing explicit consideration of locus-specific histories that could otherwise serve to confound ‘genome scanning’ approaches (e.g. Hoban et al. 2016).

Dissertation Objectives

My dissertation develops empirical systems situated within each of the above trajectories in hybridization research. By considering—and adapting as necessary—recent methodological advances in population- and phylo-genomics, I attempt to build adoptable analytical frameworks to explore similar questions in other empirical systems. Results from my analyses are then directly related to the three major themes defined above. Chapter I examines the stability of species boundaries in threatened/ endangered sympatric species of the Colorado River Basin (*Gila robusta* and *G. cypha*) as a function of anthropogenic habitat degradation in order to understand how hybridization and environmental change jointly influence species persistence.

Chapter II examines how information from ‘non-phylogenetic’ signals (*sensu* Philippe et al. 2011) in the genome can be used to categorize sources of model violation in phylogenies, and thereby categorizes hybridization and rapid radiation as major sources of phylogenetic discordance in *Gila*. This provides a necessary framework to explore the broader role of hybridization in generating ‘real’—as opposed to artefactual—patterns of diversification in phylogenies.

Chapter III expands on this theme by using sequential patterns of admixture-derived ancestry in the genome to distinguish different outcomes of hybridization. Specifically, using the genome of the enigmatic (and critically endangered) red wolf (*Canis rufus*) I discriminate between hybrid speciation and secondary (=post-speciation) introgression. Here, I identified the manner by which genomes stabilize after a hybridization event to create a non-random distribution of ancestries, as structured by recombination.

Chapters IV–VII describe new computational methods and software for developing or analyzing genome-wide SNP datasets. Chapter IV describes FRAGMATIC, a program written in Perl which enables the optimization (e.g. for cost-efficiency or throughput) of reduced-representation projects (ddRADseq) *in silico* using available genomic references for related species. Chapter V describes another method for reduced-representation genomic design, MRBAIT, which enables the universal design of targeted-enrichment protocols for non-model organisms within a diversity of data contexts. Chapter VI describes FGTPARTITIONER, a rapid Python program for delimiting a most-parsimonious set of recombinational breakpoints in genome-wide SNP datasets without requiring complex assumptions or extensive *a priori* genomic resources.

Combined, these case studies and affiliated methods cumulatively contribute to our understanding of the role of hybridization in creating biological diversity, as well as the manner by which hybridization might modulate the response of extant diversity to a dynamic future (e.g. governed by global-scale anthropogenic processes and climate change). My results show that species boundaries are not static, and moreover, that breaches therein have been important contributors of adaptive diversity throughout the evolutionary histories of the focal taxa. Furthermore, my analyses show how human-mediated dissolution of species boundaries is a non-trivial threat to extant biodiversity. My research also shows that the accumulation of hybrid histories in the genome is non-random, and that assuming otherwise may lead to erroneous conclusions both in terms of phylogenetic and demographic inference. These objectives together concretize a nuanced role for hybridization as a mediator of lineage evolution on both micro- and macro-evolutionary scales.

References

Abbott R., Albach D., Ansell S., et al. 2013. Hybridization and speciation. J. Evol. Biol. 26:229–246.

Anderson E. 1948. Hybridization of the habitat. Evolution. 2:1–9.

Barton N.H. 2001. The role of hybridization in evolution. Mol. Ecol. 10:551–568.

Blair C., Ané C. 2019. Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. Syst. Biol. 0:1–9.

Buerkle C.A., Wolf D.E., Loren H. 2003. The origin and extinction of species through hybridization. In: Brigham C.A., Schwartz M.W., editors. Population Viability in Plants: Conservation, Management, and Modelling of Rare Plants. p. 117–141.

Cruickshank T.E., Hahn M.W. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23:3133–3157.

Dittrich-Reed D.R., Fitzpatrick B.M. 2013. Transgressive Hybrids as Hopeful Monsters. Evol. Biol. 40:310–315.

Dobzhansky T. 1937. Genetic nature of species differences. Am. Nat. 71:404–420.

Dowling T.E., Secor C.L. 1997. The Role of Hybridization and Introgression in the Diversification of Animals. Annu. Rev. Ecol. Syst. 28:593–619.

Fitzpatrick S.W., Gerberich J.C., Angeloni L.M., et al. 2016. Gene flow from an adaptively divergent source causes rescue through genetic and demographic factors in two wild populations of Trinidadian guppies. Evol. Appl. 9:879–891.

Folk R.A., Soltis P.S., Soltis D.E., Guralnick R. 2018. New prospects in the detection and comparative analysis of hybridization in the tree of life. Am. J. Bot. 105:364–375.

Fontaine M.C., Pease J.B., Steele A., et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. 347:1258524.

Grabenstein K.C., Taylor S.A. 2018. Breaking barriers: Causes, consequences, and experimental utility of human-mediated hybridization. Trends Ecol. Evol. 33:198–212.

Grant P.R., Grant B.R. 2019. Hybridization increases population variation during adaptive radiation. Proc. Natl. Acad. Sci. U. S. A. 116:23216–23224.

Hahn M.W., Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution. 70:7–17.

Hamilton J.A., Miller J.M. 2016. Adaptive introgression as a resource for management and genetic conservation in a changing climate. Conserv. Biol. 30:33–41.

Hoban S., Kelley J.L., Lotterhos K.E., et al. 2016. Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. Am. Nat. 188(4):379-397.

Hubbs C.L. 1955. Hybridization between fish species in nature. Syst. Zool. 4:1–20.

Hvala J.A., Frayer M.E., Payseur B.A. 2018. Signatures of hybridization and speciation in genomic patterns of ancestry. Evolution. 72:1540–1552.

Kanarek A.R., Webb C.T., Barfield M., Holt R.D. 2014. Overcoming allee effects through evolutionary, genetic, and demographic rescue. J. Biol. Dyn. 9:15–33.

Kearns A.M., Restani M., Szabo I., et al. 2018. Genomic evidence of speciation reversal in ravens. Nat. Commun. 9:906.

Kokko H., Chaturvedi A., Croll D., Fischer M.C., Guillaume F., Karrenberg S., Kerr B., Rolshausen G., Stapley J. 2017. Can Evolution Supply What Ecology Demands? Trends

Ecol. Evol. 32:187–197.

Larson E.L., Tinghitella R.M., Taylor S.A. 2019. Insect Hybridization and Climate Change. Front. Ecol. Evol. 7:1–11.

Lindsey H.A., Gallie J., Taylor S., Kerr B. 2013. Evolutionary rescue from extinction is contingent on a lower rate of environmental change. Nature. 494:463–467.

Liu K.J., Dai J., Truong K., Song Y., Kohn M.H., Nakhleh L. 2014. An HMM-based comparative genomic framework for detecting introgression in eukaryotes. PLoS Comput. Biol. 10:e1003649.

Mallet J., Besansky N., Hahn M.W. 2016. How reticulated are species? BioEssays. 38:140–149.

Marques D.A., Meier J.I., Seehausen O. 2019. A combinatorial view on speciation and adaptive radiation. Trends Ecol. Evol. 34:531–544.

Mayr E. 1963. Animal species and evolution. Belknap Press of Harvard University Press.

Meier J.I., Marques D.A., Mwaiko S., Wagner C.E., Excoffier L., Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat. Commun. 8:1–11.

Muhlfeld C.C., Kalinowski S.T., McMahon T.E., Taper M.L., Painter S., Leary R.F., Allendorf F.W. 2009. Hybridization rapidly reduces fitness of a native trout in the wild. Biol. Lett. 5:328–331.

Mussmann S.M., Douglas M.R., Anthonysamy W.J.B., Davis M.A., Simpson S.A., Louis W., Douglas M.E. 2017. Genetic rescue, the greater prairie chicken and the problem of conservation reliance in the Anthropocene. R. Soc. Open Sci. 4:160736.

Orr H.A., Unckless R.L. 2014. The population genetics of evolutionary rescue. PLoS Genet. 10:1–9.

Owens G.L., Samuk K. 2020. Adaptive introgression during environmental change can weaken reproductive isolation. Nat. Clim. Chang. 10:58–62.

Oziolor E.M., Reid N.M., Yair S., Lee K.M., Guberman VerPloeg S., Bruns P.C., Shaw J.R., Whitehead A., Matson C.W. 2019. Adaptive introgression enables evolutionary rescue from extreme environmental pollution. Science. 364:455–457.

Patterson N., Moorjani P., Luo Y., Mallick S., Rohland N., Zhan Y. 2012. Ancient admixture in human history. Genetics. 192:1065–1093.

Payseur B.A., Rieseberg L.H. 2016. A genomic perspective on hybridization and speciation. Mol. Ecol. 25:2337–2360.

Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14:1–24.

Pease J.B., Hahn M.W. 2015. Detection and polarization of introgression in a five-taxon phylogeny. Syst. Biol. 64:651–662.

Philippe H., Brinkmann H., Lavrov D. V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol. 9:e1000602.

Rhymer J.M., Simberloff D. 1996. Extinction by hybridization and introgression. Annu. Rev. Ecol. Syst. 27:83–109.

Seehausen O., Takimoto G., Roy D., Jokela J. 2008. Speciation reversal and biodiversity dynamics with hybridization in changing environments. Mol. Ecol. 17:30–44.

Servedio M.R., Hermisson J. 2020. The evolution of partial reproductive isolation as an adaptive optimum. Evolution. 74:4–14.

Servedio M.R., Hermisson J., Van Doorn G.S. 2013. Hybridization may rarely promote speciation. J. Evol. Biol. 26:282–285.

Servedio M.R., Noor M.A.F. 2003. The role of reinforcement in speciation: Theory and data. Annu. Rev. Ecol. Evol. Syst. 34:339–364.

Smith S.D., Pennell M.W., Dunn C.W., Edwards S. V. 2020. Phylogenetics is the new genetics (for most of biodiversity). Trends Ecol. Evol. In press.

Springer M.S., Gatesy J. 2018. Delimiting coalescence genes (C-genes) in phylogenomic data sets. Genes. 9:1–19.

Stebbins G.L. 1959. The role of hybridization in evolution. Proc. Am. Philos. Soc. 103:231–251.

Stelkens R.B., Brockhurst M.A., Hurst G.D.D., Greig D. 2014. Hybridization facilitates evolutionary rescue. Evol. Appl. 7:1209–1217.

Stewart G.S., Morris M.R., Genis A.B., Szűcs M., Melbourne B.A., Tavener S.J., Hufbauer R.A. 2017. The power of evolutionary rescue is constrained by genetic load. Evol. Appl. 10:731–741.

Taylor S.A., Larson E.L. 2019. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. Nat. Ecol. Evol. 3:170–177.

Tomasini M., Peischl S. 2020. When does gene flow facilitate evolutionary rescue ? bioRxiv. 622142.

Twyford A.D., Ennos R.A. 2012. Next-generation hybridization and introgression. Heredity. 108:179–189.

vonHoldt B.M., Brzeski K.E., Wilcove D.S., Rutledge L.Y. 2018. Redefining the role of admixture and genomics in species conservation. Conserv. Lett. 11:1–6.

CHAPTER I

Hybridization drives genetic erosion in sympatric desert fishes of western North America

Chafin TK, Douglas MR, Martin BT, Douglas ME

Abstract

Many species have evolved or currently coexist in sympatry due to differential adaptation in a heterogeneous environment. However, anthropogenic habitat modifications can either disrupt reproductive barriers or obscure environmental conditions which underlie fitness gradients. In this study, we evaluated the potential for an anthropogenically-mediated shift in reproductive boundaries that separate two historically sympatric fish species (*Gila cypha* and *G. robusta*) endemic to the Colorado River Basin using ddRAD sequencing of 368 individuals. We first examined the integrity of reproductive isolation while in sympatry and allopatry, then characterized hybrid ancestries using genealogical assignment tests. We tested for localized erosion of reproductive isolation by comparing site-wise genomic clines against global patterns and identified a breakdown in the drainage-wide pattern of selection against interspecific heterozygotes. This, in turn, allowed for the formation of a hybrid swarm in one tributary, and asymmetric introgression where species co-occur. We also detected a weak but significant relationship between genetic purity and degree of consumptive water removal, suggesting a role for anthropogenic habitat modifications in undermining species boundaries. In addition, results from basin-wide genomic clines suggested that hybrids and parental forms are adaptively non-equivalent. If so, then a failure to manage for hybridization will exacerbate the long-term extinction risk in parental populations. These results reinforce the role of anthropogenic habitat modification in promoting interspecific introgression in sympatric species by relaxing divergent

selection. This, in turn, underscores a broader role for hybridization in decreasing global biodiversity within rapidly deteriorating environments.

Introduction

Many natural populations respond to anthropogenic change by either shifting geographic distributions or adjusting life histories so as to 'track' optimal conditions (Hoffmann and Sgrò 2011; Pecl et al. 2017). However, the ability of organisms to track changing environments is conditioned upon the rate of environmental change (Lindsey et al. 2013) and the rate at which adaptive machinery can act (Orr and Unckless, 2014). This evolutionary caveat creates an incentive for hybridization, in that recombinant genotypes might more rapidly establish in a dynamic adaptive landscape (Klonner et al. 2017). Widespread hybridization thus may provide an effective mechanism of population persistence in changing or novel conditions (Pease et al. 2016; Meier et al. 2017). Introgressed alleles that are beneficial under novel conditions can then be driven to fixation by the combined action of recombination and selection (Arnold and Martin 2010).

However, the relationship between hybridization and extinction is not well established under contemporary timescales. On one hand, hybrid lineages might facilitate adaptation by providing access to a greater pool of genetic variation (Dittrich-Reed and Fitzpatrick 2013; Schumer et al. 2018), whereas on the other, diversity might diminish as species boundaries dissolve (Buerkle et al. 2003; Kearns et al. 2018). Often, results are a combination of the above. Introgressed genotypes may initially compensate for erratic conditions and facilitate population persistence in the near term, but with lineages eventually merging if environmental change is prolonged (Seehausen et al. 2008). This presents an obvious paradox for conservation efforts, in that the permeability of species-boundaries may be seen as promoting both persistence and extinction.

Hybridization also represents a legacy issue for conservation policy (Allendorf et al. 2001), due primarily to its conflict with a species-centric management paradigm (Fitzpatrick et al. 2015; Hamilton and Miller 2016). Although the reticulate nature of speciation has become a contemporary research focus (e.g. Mallet et al. 2016), it has yet to gain consensus among managers (vonHoldt et al. 2018). This unanimity is required to understand the manner by which anthropogenic modifications disrupt species boundaries (Grabenstein and Taylor 2018; Ryan et al. 2018). However, predicting the outcome of hybridization in a changing environment requires an understanding of both the temporal and spatial stability of the mechanisms (e.g. intrinsic and extrinsic) that are responsible for maintaining species boundaries. In this sense, consistent patterns can often be obscured by local context (e.g. individual behaviors, population demographics; Klein et al. 2017). Hence, there remains a need to quantify the manner by which species boundaries in diverse taxa respond to rapid environmental change. We applied these perspectives to endemic, large-bodied and long-lived minnows that exist within the Colorado River, one of the most impacted riverine ecosystems of the Anthropocene (Hughes et al. 2007). Because of the pervasive human impacts therein, the Colorado River provides a natural laboratory within which to examine the stability of species undergoing rapid, anthropogenically-induced environmental change.

Hybridization in Gila

Hybridization has long been recognized as an evolutionary process in fishes (Hubbs, 1955), and as such, has been hypothesized as a mechanism for native fish diversification in western North

America (e.g. DeMarais et al. 1992). An inseparable link also exists between fishes and their environment, such that opportunities for migration or hybridization can be substantially influenced by characteristics of the riverscape (Hopken et al. 2013; Thomaz et al. 2016). The instability produced by modified flows may compromise boundaries between historically coexisting species, or provide ecological opportunities within which hybrid lineages might capitalize (Dowling and Secor 1997). The fact that habitats in western North America have a dynamic history including tectonism and progressive aridity also provides one potential causative factor for introgressive hybridization (e.g. Mandeville et al. 2017; Bangs et al. 2018). However, more contemporary anthropogenic modifications are also prominent and widespread, most apparent in the form of water acquisition and retention (Cayan et al. 2010). As a result, niche gradients that historically segregated species are now seriously perturbed. This, in turn, can promote hybridization by effectively removing selection against hybrid phenotypes, and by disrupting the phenology and reproductive cues that discourage heterospecific mating (Grabenstein and Taylor 2018).

We applied these perspectives to three species of conservation concern endemic to the Colorado River Basin: Humpback chub [*Gila cypha* (IUCN status=Endangered)], Roundtail chub [*G. robusta* (Near Threatened)], and Bonytail [*G. elegans* (Critically Endangered)]. All are hypothesized as exhibiting various levels of historic hybridization, with contemporary populations shaped by geologic processes and anthropogenic interventions. *Gila cypha* and *G. robusta*, display not only morphological intergradation (McElroy and Douglas 1995) but also taxonomic ambiguity (Douglas et al. 1989) and cannot be distinguished on the basis of mitochondrial (mt)DNA (Douglas and Douglas 2007; Dowling and DeMarais 1993), despite numerous lines of evidence supporting evolutionary independence [genetic structuring in nuclear markers (microsatellites: Douglas and Douglas 2007); discrete persistence in the fossil record (e.g. Uyeno and Miller 1963, 1965); pre-mating isolation in the form of exclusive reproductive ecology and phenology (Kaeding et al. 1990); and divergent phenotypic evolution (Smith et al. 1979; Valdez et al. 1990; Portz and Tyus 2004)]. Also, McElroy and Douglas (1995) and Douglas et al. (1998) found clear species-level differentiation in discriminant and geometric morphometric space, respectively, while the former also reported species-intermediacy at two sympatric localities (Desolation and Cataract canyons).

A likely explanation for this mosaic pattern would invoke historic separation followed by hybridization. We examine this possibility herein and framed our results within the context of change both on geologic and contemporary timescales.

Methods

Sampling

Fin tissue was non-lethally sampled from 368 specimens across three native *Gila* of the Colorado River Basin (*G. cypha*, *G. elegans*, and *G. robusta*; Table 1), collected primarily by state/ federal agencies between 1997-2017 (see Acknowledgements). One location, at the San Rafael River (hereafter RSRR), was sampled both in 2009 and 2017. Given the conservation status of these fishes, we minimized impacts on already-stressed populations by opportunistic sampling which took advantage of monitoring activities by agencies.

Gila cypha is constrained within five known aggregates associated with specific geomorphic features: Black Rocks, Cataract, Desolation, Grand, Westwater, and Yampa canyons (Fig. 1; USFWS, 2011), all of which were sampled save Cataract Canyon. Westwater (HWWC) and Black Rocks (HBKR) were treated separately, despite their potential for connectivity

(Francis et al. 2016). Due to its range-wide extirpation, samples of *G. elegans* were obtained from the Southwestern Native Aquatic Resources and Recovery Center, Dexter, NM (formerly the Dexter National Fish Hatchery). Our sampling of *G. robusta* encompassed its entire range, to include pre-defined MUs (=Management Units; Douglas and Douglas 2007) and represented wild populations, with the exception the Mancos River (RMCO), which was obtained from the Colorado Department of Wildlife Native Aquatic Species Restoration Facility. *Gila robusta* from the lower basin Bill Williams and Gila River drainages was not included, given its known polyphyly (Dowling and DeMarais 1993; Chafin et al. unpubl.).

Data collection

Genomic DNA was extracted using either PureGene® or DNeasy® kits (Qiagen Inc.), with electrophoresis (2% agarose gel) confirming presence of sufficiently high molecular weight DNA. Our ddRAD library preparations were modified from previous protocols (Peterson et al. 2012). Restriction enzyme pairings and size-selection ranges were optimized using an *in silico* procedure (FRAGMATIC; Chafin et al. 2018). Samples were digested with *Msp*I (5'-CCGG-3') and *Pst*I (5'-CTGCAG-3') following manufacturer's protocols (New England Biosciences). Fragments were then purified using Ampure XP beads (Beckman-Coulter Inc.) and concentrations standardized at 100ng per sample. Custom adapters containing in-line barcodes were ligated with T4 Ligase (New England Biosciences), pooled in sets of 48, and size-selected with the Pippin Prep (Sage Sciences) at 250-350bp prior to adjusting for adapter length (=gDNA length). We then utilized a 12-cycle PCR to extend adapters with indexed Tru-Seq primers and Phusion high-fidelity DNA polymerase (manufacturer protocols; New England Biosciences). Final libraries were visualized on the Agilent 2200 TapeStation fragment analyzer and pooled for

100bp read length single-end sequencing (Illumina HiSeq 2500; University of Wisconsin/Madison).

Assembly and filtering of genomic data

Data assembly was performed using computing resources at the Arkansas High Performance Computing Center (AHPCC), and the XSEDE-funded cloud computing resource JetStream (co-managed by the Pervasive Technology Institute/Indiana University, and the Texas Advanced Computing Center/Austin).

Raw Illumina reads were demultiplexed and filtered using the PYRAD pipeline (Eaton 2014). Discarded reads exhibited >1 mismatch in the barcode sequence or >5 nucleotides with Phred quality <20. Loci were clustered *de novo* within and among samples using a distance threshold of 80%. We then removed loci with: >5 ambiguous nucleotides; >10 heterozygous sites in the consensus sequence; >2 haplotypes per individual; <20X and >500X coverage per individual; >70% heterozygosity per-site among individuals; or presence in <50% of individuals. Individuals with >50% missing data were also discarded. Scripts for post-assembly filtering and file conversion are available as open-source (github.com/tkchafin/scripts).

Estimating population and individual ancestry

Hypotheses of admixture and hybridization were based on genetic differentiation, as visualized using Discriminate Analysis of Principal Components (DAPC; R-package *adegenet;* Jombart, 2008). Discriminant functions combine principal components (PCs) so as to maximally separate hypothesized groups. Importantly, sufficient PC axes must be retained so as to summarize the high-dimensional input, yet also avoid over-fitting. We accomplished this using the following cross-validation procedure: Stratified random sampling defined 80% of samples per population as a "training set," with the remaining 20% then classified. PC retention was optimized by minimizing root-mean-square error (RMSE) while maximizing classification success across analyses.

These results were contrasted with model-based assignment tests (STRUCTURE, Pritchard et al. 2000; ADMIXTURE, Alexander and Novembre 2009). A shared assumption is that populations can be divided into *K*-clusters identified by permuting membership so as to minimize linkage disequilibrium and departure from Hardy-Weinberg expectations. Given excessive runtimes in STRUCTURE, we first applied ADMIXTURE to evaluate a broader range of models (i.e., *K*=1-20, using 20 replicates), followed by STRUCTURE on a reduced range (*K*=1-10, using 10 replicates with 500,000 MCMC iterations following a burn-in period of 200,000).

Model selection followed a cross-validation procedure in ADMIXTURE where assignment error was minimized by optimal choice of *K*, with results parsed using available pipelines (github.com/mussmann82/admixturePipeline). We used the delta *K* method (Evanno et al. 2005) to define the proper model in STRUCTURE (CLUMPAK; Kopelman et al. 2015).

We identified putative admixed individuals using Bayesian genealogical assignment (NEWHYBRIDS, Anderson and Thompson 2002) that assessed the posterior probability of assignment to genealogical classes (e.g. F_1, F_2), as defined by expected genotype frequency distributions. This component is vital, in that mixed probability of assignment in STRUCTURE and ADMIXTURE can stem from weakly differentiated gene pools. The MCMC procedure in NEWHYBRIDS was run for 4,000,000 iterations following 1,000,000 burn-in, using a panel of 200 loci containing the highest among-population differentiation (F_{ST}) and lowest linkage disequilibrium ($r^2 < 0.2$), as calculated in GENEPOPEDIT (Stanley et al. 2017). To ensure accuracy of this method as applied to our data, we performed a power analysis using the HybridDetective workflow (Wringe et al. 2017). We first generated simulated multi-generational hybrids using 50% as a training dataset, and analyzed classification success across replicated simulations using the remaining 50% of samples as a validation set. To examine convergence, simulations were run across three replicates, each with three independent MCMC chains. Final runs were used to categorize individuals to genealogical class, using a posterior probability threshold of 0.90.

Spatial and genomic heterogeneity in introgression

We tested for signatures of reproductive isolation by examining clinal patterns in locus-specific ancestry across hybrid genomes, using multinomial regression to predict genotypes as a function of genome-wide ancestry. Analyses were performed in the R-package *introgress* (Gompert and Buerkle 2010). Putatively 'pure' populations of *G. robusta* and *G. cypha* were diagnosed from results generated by NewHybrids. We first filtered loci to include those with allele frequencies that differed in the reference populations (as defined by δ>0.8, where δ is the allele frequency differential at a given locus; Gregorius and Roberds 1986). We generated a null distribution by randomly re-assigning genotypes across 1,000 permutations, so as to test for deviations from neutral expectations. The significance of locus-specific clines (fit via multinomial regression) was then determined by computing a log-likelihood ratio of inferred clinal models versus the null model (at P<0.001).

To test for localized breakdown in reproductive barriers, we examined congruence of locus-specific introgression among sampling localities. We did so by deriving site-wise genomic clines within species, then subsequently contrasting the fit of site-wise regression models to the global pattern for each locus. This was accomplished by estimating probabilities of the observed genotypes for each site ($X_{i,j}$ where X=genotypic data over i sites for each locus j) given the site-specific models ($M_{i,j}$) versus the range-wide model ($M_{global,j}$). Concordance was reported as the log-likelihood ratio of $L(M_{global,j} \mid X_{i,j})$ to $L(M_{i,j} \mid X_{i,j})$ computed per-locus (Gompert and Buerkle 2009).

Testing effects of anthropogenic pressures

To test correlations between anthropogenic pressures on rates of hybridization, we parsed pressure indices per river reach for four dimensions of human impact from the global stream classifications of Grill et al. (2019). These were: 1) River fragmentation (=degree of fragmentation; DOF); 2) Flow regulation (=degree of regulation; DOR); 3) Sediment trapping (=SED); and 4) Water consumption (=USE), from the global stream classifications of Grill et al. (2019). We also tested predictive capacity of an integrated multi-criterion connectivity status index (=CSI), also from the free-flowing river assessments of Grill et al. (2019). Briefly, the DOF index (from 0 to 100) represents the flow disruption on a reach from dams, while also considering natural barriers such as waterfalls. The DOR index is derived from the relationship between storage volumes of reservoirs and annual river flows and is expressed as the percentage of total river flow that can be withheld in the reservoirs of a river reach. SED and USE quantify the potential sediment load trapped by dams, and the long-term average anthropogenic water consumption as a percentage of natural flow, respectively. The CSI index is a weighted average of these pressure indicators, while also considering road densities and degrees of urbanization [see Grill et al. (2019) for details regarding derivation of these indices and their underlying data sources].

We assigned pressure index values for all sites containing at least 1 hybrid (as classified using a 0.90 posterior probability threshold), and tested the predictive power of each pressure dimension on 'genetic purity' (calculated via linear regression as the proportion of individuals per population assigned to either P_0 or P_1).

Results

A mean of 106,061 loci were assembled per sample (σ=42,689). Following quality/depth filtering, and with mean coverage of 88X, this yielded 16,001 per sample (σ=6427). Loci were removed if absent in <50% of individuals, with paralog filtering performed on the basis of allele count and excess heterozygosity. This resulted in 13,538 loci (μ=10,202; σ=3601), and 1,257,356 nucleotides. Putative orthologs contained 62,552 SNPs, of which 38,750 were parsimony-informative, corresponding to 4.9% and 3% of sampled nucleotides. We retained one SNP per locus, with a final dataset comprising12,478 unlinked SNPs.

Population structure

Choice of *K* varied by assignment test, with *K*=8 (ADMIXTURE; Fig. S1, S2, and *K*=5 (STRUCTURE; Fig. S2). We thus retained *K*-values from 5-8.

The discriminant function axis with the greatest differentiation (DA1) primarily segregated *G. robusta* in the Little Colorado River (RECC) from the remaining *G. robusta* and *G. cypha*, with DA2 differentiating Upper Colorado *G. robusta* from *G. cypha* (Fig. 2A). Interestingly, DA3 (Fig. 2B) seemingly identified structure within *G. cypha* as well as potentially admixed populations of *G. robusta*. Both assignment tests (Fig. 3) differentiated RECC from conspecifics, with *G. elegans* also forming a discrete cluster in all cases. We interpret neither of these results as surprising, given the substantial anthropogenic (Glen Canyon dam) and natural (Little Colorado River Grand Falls) barriers separating the former from conspecifics, and the phylogenetic distinction of the former (Chafin *et al*. 2019). Additionally, no signal of contemporary mixture of *G. robusta* or *G. cypha* with *G. elegans* was detected. STRUCTURE models with *K*>8 and ADMIXTURE *K*>5 showed similar restrictions in gene flow between Grand Canyon *G. cypha versus* upper basin sites. Desolation Canyon (HDES) showed the highest probability of assignment to an 'upper basin' *G. cypha* cluster. Within *G. robusta*, a weak signal of differential assignment was apparent when Green River tributaries (RUGR and RMGR) were compared with the mainstem Colorado River, suggesting either reduced intraspecific gene flow, or an artefact of demographic processes.

DAPC and assignment tests each indicated potential hybridization among *G. cypha* and *G. robusta*, most prominently in regions of sympatry [i.e. Black Rocks (RBKR/HBKR), Westwater (RWWC/ HWWC), and Yampa (RYAM/HYAM) canyons; Fig. 3]. Those *G. robusta* sites most 'distant' in multivariate space (Fig. 2B) were also those which showed the least probability of interspecific assignment (Fig. 3). Signals of asymmetric introgression were apparent when sympatric localities were examined, with *G. cypha* generally having higher levels of heterospecific assignment.

One exception was RDES, where all specimens phenotypically identified as "*G. robusta*" were genetically indistinguishable from those designated as *G. cypha*. Misidentifications at time of capture is a likely cause, owing to the morphological intermediacy of *Gila spp.* at this site (i.e. McElroy and Douglas 1995).

Allopatric populations of *G. robusta* showed less interspecific ancestry, with the exception being the San Rafael River (RSSR), where samples had 30-50% assignment to *G.*

cypha ancestry (Fig. 3), a pattern supported by the weak differentiation of RSSR in DAPC analyses. Allopatric *G. robusta* from RNJA also showed mixed probability of assignment to *G. cypha*, albeit with low probability and consistency.

Hybrid detection and genealogical assignment

Genealogical assignment in NEWHYBRIDS was used to parse STRUCTURE and ADMIXTURE results for contemporary hybridization. We first defined a prior probability of genetic purity for *G. robusta* as being the upper-most Little Snake River tributaries (RLSR), and for *G. cypha* as the Little Colorado River confluence in Grand Canyon (HLCR). Both were chosen based on STRUCTURE and ADMIXTURE results (Fig. 3), and additionally informed by prior studies of natural recruitment (Douglas and Douglas 2010; Kaeding and Zimmerman 1983). Because of the lack of any signal of interspecific admixture in *G. elegans*, they were omitted from these analyses.

Introgressive hybridization at sympatric locations was found to be asymmetric (Fig. 4). In cases of mixed assignment, individuals were classified as either "late-generation hybrid," or "of uncertain status" (Table 2). *Gila robusta* were largely classified as pure in both sympatric and allopatric sites, with a few exceptions (outlined below). Samples assigned to hybrid classes tended to be *robusta*-backcrossed (6–12.5%) or late-generation/ uncertain (4.5–37.5%). In contrast, *G. cypha* at sympatric localities had comparatively low purity (0–61%), with most hybrids categorized as either F_2, *cypha*-backcrossed, or unclassifiable (Table 2). The genetic effects of hybridization are thus inferred as asymmetric, with a greater penetration of *G. robusta* alleles into *G. cypha* populations. RDES and HDES samples were mostly classified as either late-generation or unassignable. F_1 hybrids were notably absent at all localities, suggesting hybridization occurred over multiple generations and ongoing introgression (i.e., hybrids fertile and reproductively successful).

Both species showed little signal of hybridization at allopatric locations, but with notable exceptions being the San Rafael River and, to a lesser extent, RMCO. Nearly all RSSR samples were assigned with high probability as either F_2 or *G. robusta*-backcrossed hybrids, a pattern consistent across years (2009 *versus* 2017), and regardless of priors used. Samples from 2009 were mostly classified as F_2 (45%) or *robusta*-backcrosses (45%). However, the greatest proportion of 2017 samples were *robusta*-backcrosses (67%) or late-generation hybrids (25%), suggesting an increase of admixture over time (although increased sampling is needed to verify this trend; two-tailed Fisher's exact test p=0.0967; Table 2). The RMCO samples, composed of 20% hybrids (Table 2), were derived from hatchery stock, not a natural population. Thus, we cannot say if our results represent natural or accidental hybridization that coincided with, or was subsequent to, stock establishment.

Genomic clines

We also examined how introgression varied across significantly differentiated genomic SNPs and species-diagnostic markers. Here, we considered locus-specific ancestry as the probability of sampling a homozygous *G. cypha* genotype [i.e. P(AA)] as a function of genome-wide ancestry, with the expectation that scant bias should occur if fitness is independent of hybrid ancestry. All loci exhibited clinal patterns that deviated significantly from neutral expectations (p<0.001, estimated via permutation; Fig. 5A). The majority displayed coincident sigmoidal relationships between genome-wide ancestry (hybrid index; h) and locus-specific ancestry (ϕ). The dominant sigmoidal pattern is suggestive of a deficiency in interspecific heterozygosity, presumably reflecting heterozygote disadvantage (Fitzpatrick 2013). Notably, some locus-specific clines deviated from this trend (Fig. 5A), suggesting that underdominance is not ubiquitous. For example, many loci show alternative cline shapes suggestive of either over- or under-representation of parental genotypes in hybrids, suggestive of a selective advantage in these or linked genomic regions. However, lacking a suitable genomic reference, the phenotypic implications of these alternative cline forms are not explored herein.

We also examined the observed genotypes at each locus, given expectations from the range-wide model and site-specific regression models. These were reported as a log-likelihood ratio *per* locus and *within* each sampling locality (Fig. 5B). We found the 'fit' of the range-wide clinal models was rather variable, although with the majority of loci showing little deviation. One notable exception was RSSR, where an exceptionally flattened distribution of the locus-specific log-likelihood ratios was apparent. This in turn suggested that the global expectation was a poor predictor of within-population genotypes. Thus, while most loci reflected patterns consistent with selection against hybrids, the same cannot be said for the RSSR population. It also displayed a strong signal of interspecific admixture in the Bayesian and ML assignment tests (Fig. 3), and variable assignment to >2nd generation hybrid classes in NEWHYBRIDS (Fig. 4), as well as greater intermediacy in multivariate genotypic clustering (Fig. 2). Several other sites also showed a 'flattened' distribution of clinal fit among loci (e.g., HWWC, HBKR, HDES, RDES, RUGR, RMCO; see Fig. 5B). This could be a response to elevated introgression and a relative breakdown of heterozygote disadvantage in the sympatric localities (HWWC, HBKR, HDES, RDES), especially where previous analyses indicated admixture (Fig. 2-4), or as an artefact of reduced sampling (RUGR, RMCO).

Although not possible for other localities, our temporal sampling for RSSR allowed us to further explore this discrepancy across different time periods: 2009 (*N*=11) and 2017 (*N*=12). We then fitted locus-specific clines among years (Fig. 5C) and compared those to range-wide expectations (Fig. 5B). We found little qualitative change in the overall distribution, save for four outliers in 2017, suggesting further breakdown of clinal expectations over time. Even though we cannot evaluate this trend range-wide, we suggest that further study examine the persistence vs. breakdown of genetic purity by using a consistent genetic assay as a part of ongoing monitoring efforts.

Testing dimensions of anthropogenic pressure

Among sites showing varying levels of hybridization (N=10; Table 2), only the consumptive water use (=USE) was significantly corelated with a decline in genetic purity (r^2=0.444; adjusted r^2=0.375; *p*=0.035; see Fig. S5). The connectivity status index (CSI) showed a weak but non-significant positive relationship (i.e. increased connectivity = increased genetic purity; r^2=0.02; adjusted r^2=-0.103; *p*=0.7). However, we caution that sample sizes were notably low (N=10) after sites were reduced to those containing hybrids and which could be assigned pressure indices (see Fig. S6 for reach assignments), thus urge that these results be interpreted accordingly.

Discussion

We found strong evidence for contemporary hybridization among *G. cypha* and *G. robusta* extending beyond their regions of sympatry. These results refine rather than conflict with previous studies employing ‘legacy’ genetic markers (Douglas and Douglas 2007; Dowling and DeMarais 1993; Gerber et al. 2001), and complement contemporary work (Bohn *et al*. 2019). In addition, these results broaden our understanding of each species and their evolutionary histories, as well as the trajectory of their ongoing evolutionary change in the face of extensive anthropogenic modifications.

Species boundaries and reproductive isolation in Gila

Our survey of the nuclear genome suggested that contemporary hybridization between our study species is occurring where sympatric, as interpreted from several lines of evidence: The coincidence and shape of our genomic clines; the pervasive signal of genealogical assignment to early-generation hybrid classes; and signatures of selection antagonistic to interspecific heterozygous genotypes.

We interpret this hybridization as following historical isolation, particularly given the coexistence of study species since at least the mid-Pliocene (Uyeno and Miller, 1965; Spencer et al. 2008). In addition, past studies have shown sustained morphological divergence displayed in sympatry (Douglas et al. 1989; McElroy and Douglas 1995), although we note that contemporary evaluations are conspicuously absent. This suggests that genetic exchange is ongoing despite, rather than in the absence of, reproductive isolation.

Dowling and DeMarais (1993) suggested that hybridization between *G. cypha* and *G. robusta* may have contributed to the evolutionary persistence of each species by providing necessary adaptive genetic variation so as to withstand environmental fluctuations. We concur, and further note that this exchange is ongoing, with a substantial risk that contemporary habitat change will outpace the rate at which introgressed alleles are selectively "filtered." If so, then continued habitat alteration could lead to a scenario in which genetic/demographic swamping contributes to local extirpations, or to eventual genetic homogenization of one species by the other (Todesco et al. 2016). This pattern is particularly evident in the asymmetric levels of introgression into *G. cypha*, a species of particular concern given its fragmentary distribution and reduced densities within the upper Colorado River Basin (e.g. Badame 2008; Franci et al. 2016; USFWS 2017).

To consider the plausibility of such a scenario in which environmental change leads to the dissolution of a species boundary, we must first consider how this boundary is itself structured. We do so by considering results of our genetic data within the context of those derived from species-specific morphology and life history. Several morphological evaluations have demonstrated that morphological distinctions among species can be blurred in sympatry (Douglas and Douglas 2007; Douglas et al. 2001; McElroy and Douglas 1995), although we note a need for such evaluations to be revisited, particularly given the contemporary timescale of hybridization as documented herein. This is likely the result of secondary admixture, rather than a prolonged (i.e., primary) divergence that lead to only weakly-differentiated species. Pliocene fossils demonstrate that morphological divergence of *G. robusta* and *G. cypha* predates major geomorphic and tectonic events that could have triggered secondary contact. For example, the Upper Colorado River was segregated from the contemporary lower basin prior to the mid-Pliocene, (McKee et al. 1967), with the uplifting of the Colorado Plateau diverting its flow into one or more Colorado Plateau lakes (Spencer et al. 2008). Flows were subsequently diverted by headwater erosion though the Grand Canyon, forming the modern course of the river. Fossil evidence implies that ecological divergence occurred during, or prior to this time, and was sufficient in strength to generate both morphological forms (Uyeno and Miller 1965). This suggests the existence of ecological conditions that reflect those to which the species are now adapted. Additionally, numerous perturbations [i.e., tectonism, extreme drought (Meko et al.

2007)] also occurred during the interval between divergence and present, yet both species not only persisted but did so with some semblance of morphological continuity. Given this, one must again assume that an extant blurring of species-boundaries is, at least in part, a contemporary occurrence. To test this hypothesis, we considered the ecological dimensions underlying adaptive differentiation in these species.

Reproductive barriers in Gila

Phenotypic and ecological specializations of each species provide potential insights into the mechanisms promoting assortative mating. *Gila cypha* displays phenotypic characteristics interpreted as adaptations to the torrential flows of canyon-bound reaches (McElroy and Douglas 1995; Miller 1946; Valdez and Clemmer 1982). These include a prominent nuchal hump, dorsoventrally flattened head, embedded scales, terete body shape, and a very narrow caudal peduncle that terminates in a caudal fin with a high aspect ratio, indicative of a hydrodynamic shape and powerful propulsion. Its current distribution also reflects association with this type of habitat.

In contrast, *G. robusta* has a comparatively more generalized phenotype, characterized by a deeper and less streamlined body with non-imbedded scales and larger, more falcate fins (Miller 1946). It is found in the upper tributaries of larger rivers (Vanicek and Kramer 1969) with moderate flows. It fails to maintain position within the current when subjected to the extreme flows associated with *G. cypha*, and instead becomes benthic so as to avoid being swept away (Moran et al. 2018). This suggests a natural history diametrically opposed to that of *G. cypha,* where dynamic flow regimes clearly predominate. Accordingly, radiotelemetric studies verified habitat preferences for each species, with *G. cypha* seldom straying from the deep eddies and turbulent flows of canyon-bound reaches (Douglas and Marsh 1996; Gerig et al. 2014; Kaeding et al. 1990). These observations underscore the role that functional morphology plays with regards to species boundaries, in that intermediate morphologies would be maladaptive in either habitat.

However, barriers that sustain reproductive isolation are unclear, in that both species are broadcast-spawners (Johnston and Page 1992), with a temporal overlap in spawning period (Kaeding et al. 1990). The latter is likely a consequence of shared environmental cues triggering reproduction, namely seasonal changes in flow rate and temperature, with spatial segregation driven by subsequent alterations in microhabitat and substrate preference (Douglas and Douglas 2000; Minckley 1996). Widespread movements by *G. robusta* during the spawning season contrast with the relative localized focus found in *G. cypha* (Kaeding et al. 1990; Tyus et al. 1982), and again reinforce the restricted habitat requirements of the latter. Additionally, there is a stronger 'homing' component in the microhabitat preferences of *G. cypha* (Valdez and Clemmer 1982). These ecological differences, combined with overall higher abundance of *G. robusta* in most areas (e.g. Francis et al. 2015) likely contribute to the observed asymmetric introgression between the two species (Edelaar et al. 2008). Intraspecific recognition as a mate-choice mechanism is also an observed behavior that promotes reproductive isolation. Despite congruent reproductive condition and the presence of suitable substrate in a brood stock tank, natural spawning did not occur between *G. robusta* x *G. elegans* and *G. elegans* x *G. cypha* (Hamman 1981).

Thus, we contend that reproductive isolation in *G. robusta* and *G. cypha* is driven by extrinsic factors, with pre-mating isolation primarily in the form of microhabitat selection and post-mating isolation driven by functional morphological differences. Our data point to selection against hybrids, which may be reflective of either their relatively poor performance in the environment, or to a diminished success in mating. However, we noted a possible breakdown of this expectation at some localities when genomic clines were fitted to within-site patterns. The San Rafael River (RSSR), for example, is one such exception. Fortney (2015) quantified anthropogenic changes in this river over the last 100 years, with the channel being extensively canalized and diverted, and flows diminished by 83% due to water withdrawals. These manipulations yielded a narrower, relatively deeper channel that stands in sharp contrast to an historically wider and slower river whose flow regime was governed by geomorphology and dominated by flooding. Anthropogenic alterations apparently provided an opportunity for adaptive hybridization (e.g. Taylor et al. 2005), an hypothesis consistent with the exclusive presence of late-generation hybrids in the RSSR population (Fig. 4). Under this scenario, selective advantage would similarly drive outlier loci and the reduced-fit seen in our clinal models (Fig. 5). The origin of *G. cypha* alleles in this population is unclear, although they may possible be derived from a remnant population in the upper reaches (e.g., Black Box Gorge; P. Badame, pers. comm).

An examination of the degree to which anthropogenic pressures drive basin-wide hybridization point to a role for consumptive water use in driving a decline in overall genetic purity. Consumptive water usage, and the impact of the associated infrastructures (such as diversions and reservoirs), are often implicated as detrimental to freshwater fish diversity (e.g. Xenopoulos *et al*. 2005). Insofar as river discharge is one dimension of ecological heterogeneity, and given the trend of decreasing species richness as flow declines (Oberdorff et al. 1995), we posit that a coincidental relationship is rather extreme in the Colorado River when anthropogenic manipulations and extensive hybridization are contrasted. Yet, a test of this hypothesis is difficult without further experimental work (i.e. leveraging hatchery-produced interspecific hybrids to test for viability in varying habitats, represented within a series of mesocosms). While increased sampling is also necessary, it would be difficult given that we have already sampled 4 of the 5 extant *G. cypha* populations.

Modified environments and genetic swamping

Grabenstein and Taylor (2018) defined mechanisms that drive anthropogenically-mediated hybridization in coexisting species: 1) Interspecific contact promoted by habitat homogenization or altered phenology; 2) Disruption of mate selection/ choice; and 3) Habitat alteration, such that hybrid genotypes are favored (Anderson 1948). All three are plausible for *Gila*, with the 'hybrid swarm' of RSSR an extreme case. Asymmetric hybridization was implicated in all extant sympatric *G. cypha* populations, save the Cataract Canyon aggregate not evaluated in this study. The latter reflects a more '*robusta*-like' morphology (McElroy et al. 1997), with low population numbers and a slower growth rate relative to other extant populations (Badame 2008). Taken together, these suggest an elevated risk for genetic or demographic swamping in Cataract Canyon *G. cypha* (Todesco et al. 2016), and lend urgency to their inclusion in future genetic surveys.

Such a scenario may also be invoked for *G. cypha* in the Yampa River (HYAM), recognized even prior to our sampling as being of reduced and declining numbers (Tyus 1998). The ubiquity of highly-admixed genomes in our sampling (from 1999-2001), coupled with the absence of genetically pure individuals in more recent surveys (USFWS 2017), suggest the potential for local extirpation. Given the prevalence of asymmetric hybridization in other sympatric *G. cypha*, it is possible that genetic swamping may have also played a role in the decline of the HYAM population (although we cannot test that hypothesis). Of note, a more recent genetic survey of HYAM found a further breakdown of genetic purity, with most individuals being composed of either pure *G. robusta* or *robusta*-backrosses (Bohn et al. 2019). Similarly, recent surveys have also documented diminished catch ratios for *G. cypha* at other sympatric localities (Fig. 4; Francis et al. 2016; USFWS 2017). Thus, an elevated risk of genetic swamping appears as a strong potential for all *G. cypha* populations sympatric with *G. robusta*.

Genetic swamping and Allee effects

The capacity for populations to track changing conditions is constrained not only by standing genetic variation but also complex demographic processes that feed back to reproductive fitness (Kokko *et al*. 2018). As the effective population size (N_e) of a population decreases, so also do beneficial variants, primarily due to reduced efficacy of selection relative to genetic drift and associated inbreeding depression (i.e., Allee effects; Kramer et al. 2009). This in turn can induce a negative feedback that drives local extirpation (Polechová and Barton 2015). Using a similar logic, we posit that maladaptive introgression within diminishing populations could also synergistically trigger a "runaway" process of genetic swamping (Fig. S7).

In this conceptual model, demographically-driven Allee effects weakens purifying selection against maladaptive introgressed alleles, whereas their continued influx further reduces fitness via outbreeding depression. In this way, maladaptive gene flow can continually depreciate N_e and effectively promote an "extinction vortex" (Gilpin and Soulé 1986), and we posit this mechanism may contribute to the decline of those *G. cypha* populations sympatric with *G. robusta*. Although some signal of selection against heterospecific alleles was apparent, another manifestation of shrinking N_e is the expansion of genomic linkage disequilibrium (Nachman

2002). As a result, purifying selection can actually be counterproductive, wherein beneficial genetic variation is lost via selection against linked regions (Nachman and Payseur 2012).

Under this paradigm, the risk of swamping in *G. cypha* is elevated by the numerous factors that increase the relative impact of genetic drift. These are: Reduced population sizes in extant populations (Douglas and Marsh 1996; Tyus 1998); A fragmented distribution (Fagan 2002); and a "slow" life history (i.e., long generation time and extended lifespans; Olden et al. 2008), and higher vulnerability to regulated and reduced flows given its habitat preference of turbulent rivers (as above). The hybrid swarm in the San Rafael (RSRR), and the suspected genetic swamping of *G. cypha* in the Yampa River (HYAM) are potential harbingers of this erosion. Genetic integrity may be preserved in the short term by cultivating "pure" progeny via hatchery production, so as to potentially extend existing pure populations, although a propogation program risks further reducing N_e (Allendorf et al. 2001). The development of pure stock for *G. robusta* should be relatively easy, whereas upper basin *G. cypha* are more problematic in that they display various levels of hybridization (i.e., Figs. 3-4). In this regard, we echo the “producer’s gambit” philosophy (McElroy et al. 1997) where hybrid populations fall under an expanded conservation paradigm when genetic purity cannot otherwise be maintained (Lind-Riehl et al. 2016). Given apparent ecological non-equivalency of hybrids, we suggest that habitat restoration is the only long-term means to resurrect genetic purity in these populations (Wayne and Shaffer 2016). Here, restoration re-establishes adaptive gradients favoring specialist phenotypes, even so far as to drive their reemergence from hybridized swarms (Gilman and Behm 2011), as has been seen in European whitefish following eutrophication-driven hybridization (Jacobs et al. 2019). Thus, restoration efforts may be more effectively targeted to areas of already reduced purity, with continued genetic monitoring as a necessary assessment tool (e.g. Bohn et al. 2019).

Conclusion

A reduced-representation assay of nuclear genomes in *G. robusta* and *G. cypha* provided evidence of asymmetrical hybridization that is range-wide and spatially heterogeneous (Fig. 3, 4). We interpreted this as reflecting secondary contact, particularly given the pervasive selection we found with regard to genomic clines operating against interspecific heterozygotes (Fig. 5), although we do not exclude the potential for historically limited introgression. Although we lacked appropriate sampling to adequately test for temporal changes in hybridization rates, we did observe the expansion of a hybrid swarm in the San Rafael River (RSSR) over an eight-year period, as well as high levels of asymmetric hybridization in all sympatric populations of *G. cypha* (Table 2). This underscores the potential for genetic/demographic swamping by *G. robusta*, as well as exacerbating the extirpation risk for extant populations of *G. cypha*. We argue that conservation plans for *G. cypha* must consider this possibility. We also suspect the species-boundary for *G. cypha* is largely maintained by extrinsic factors (i.e., lower fitness of hybrid phenotypes and differential microhabitat preferences). As such, further habitat degradation and homogenization may lead to complete genetic erosion, either by contravening habitat selection for pure individuals, or by promoting modified anthropogenic riverscapes that serve as habitat for novel hybrid lineages/swarms. The scenario playing out in *Gila* emphasizes a philosophical dilemma that conservation policy must confront: Is hybridization antagonistic to the conservation of biodiversity or is it instead a natural adaptive mechanism employed routinely by species in their evolutionary struggle to persist.

References

Alexander D.H., Novembre J. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Allendorf F.W., Leary R.F., Spruell P., Wenburg J.K. 2001. The problems with hybrids: Setting conservation guidelines. Trends Ecol. Evol. 16:613–622.

Anderson E. 1948. Hybridization of the habitat. Evolution. 2:1–9.

Anderson E.C., Thompson E. a. 2002. A model-based method for identifying species hybrids using multilocus data. Genetics. 160:1217–1229.

Arnold M.L., Martin N.H. 2010. Hybrid fitness across time and habitats. Trends Ecol. Evol. 25:530–536.

Badame P. V. 2008. Population estimates for humpback chub (*Gila cypha*) in Cataract Canyon, Colorado River, Utah, 2003 – 2005. Final Rep. to Utah Divid. Wildl. Resour. Up. Color. River Endanger. Fish Recover. Progr. Proj. #22L.

Bangs M.R., Douglas M.R.M.E., Mussmann S.M., Douglas M.R.M.E. 2018. Unraveling historical introgression and resolving phylogenetic discord within *Catostomus* (Osteichthys : Catostomidae). BMC Evol. Biol. 18:86.

Behm J.E., Ives A.R., Boughman J.W. 2010. Breakdown in postmating isolation and the collapse of a species pair through hybridization. Am. Nat. 175:11–26.

Berec L., Angulo E., Courchamp F. 2007. Multiple Allee effects and population management. Trends Ecol. Evol. 22:185–191.

Buerkle C.A., Wolf D.E., Loren H. 2003. The origin and extinction of species through hybridization. In: Brigham C.A., Schwartz M.W., editors. Population Viability in Plants: Conservation, Management, and Modelling of Rare Plants. p. 117–141.

Cayan D.R., Das T., Pierce D.W., Barnett T.P., Tyree M., Gershunov A. 2010. Future dryness in the southwest US and the hydrology of the early 21st century drought. Proc. Natl. Acad. Sci. 107:21271–21276.

Chan W.Y., Peplow L.M., Menéndez P., Hoffmann A.A., van Oppen M.J.H. 2018. Interspecific Hybridization May Provide Novel Opportunities for Coral Reef Restoration. Front. Mar. Sci. 5:1–15.

DeMarais B.D., Dowling T.E., Douglas M.E., Minckley W.L., Marsh P.C. 1992. Origin of Gila seminuda (Teleostei: Cyprinidae) through introgressive hybridization: implications for evolution and conservation. Proc. Natl. Acad. Sci. U. S. A. 89:2747–2751.

Dittrich-Reed D.R., Fitzpatrick B.M. 2013. Transgressive hybrids as hopeful monsters. Evol. Biol. 40:310–315.

Douglas M., Marsh P. 1996. Population estimates/population movements of *Gila cypha*, an endangered cyprinid fish in the Grand Canyon region of Arizona. Copeia. 1:15–28.

Douglas M., Minckley W.L., DeMarais B.D. 1999. Did vicariance mold phenotypes of western North American fishes? Evidence from Gila River cyprinids. Evolution. 53:238–246.

Douglas M.E., Douglas M.R., Lynch J.M., McElroy D.M. 2001. Use of geometric morphometrics to differentiate Gila (Cyprinidae) within the Upper Colorado River Basin. Copeia. 2001:389–400.

Douglas M.E., Miller R.R., Minckley W.L. 1998. Multivariate discrimination of Colorado plateau *Gila* spp.: The "Art of seeing well" revisited. Trans. Am. Fish. Soc. 127:163–173.

Douglas M.E., Minckley W.L., Tyus H.M. 1989. Qualitative characters, identification of Colorado River chubs (Cyprinidae Genus *Gila*) and the "Art of Seeing Well." Copeia. 1989:653–662.

Douglas M.R., Douglas M.E. 2000. Genetic structure of humpback chub Gila cypha and roundtail chub G. robusta in the Colorado River ecosystem. Rep. to Gd. Canyon Monit. Reserach Center, U.S. Geol. Surv.

Douglas M.R., Douglas M.E. 2007. Genetic structure of humpback chub Gila cypha and roundtail chub G. robusta in the Colorado River ecosystem. Rep. to Gd. Canyon Monit. Reserach Center, U.S. Geol. Surv.:99.

Dowling T.E., DeMarais B.D. 1993. Evolutionary significance of introgressive hybridization in cyprinid fishes. Nature. 362:444–446.

Dowling T.E., Markle D.F., Tranah G.J., Carson E.W., Wagman D.W., May B.P. 2016. Introgressive hybridization and the evolution of lake-adapted catostomid fishes. PLoS One. 11:e0149884.

Dowling T.E., Secor C.L. 1997. The role of hybridization and introgression in the diversification of animals. Annu. Rev. Ecol. Syst. 28:593–619.

Eaton D.A.R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 30:1844–1849.

Edelaar P., Siepielski A.M., Clobert J. 2008. Matching habitat choice causes directed gene flow: A neglected dimension in evolution and ecology. Evolution. 62:2462–2472.

Evanno G., Regnaut S., Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol. Ecol. 14:2611–2620.

Fagan W.F. 2002. Connectivity, fragmentation, and extinction risk in dendritic metapopulations. Ecology. 83:3243–3249.

Fitzpatrick B.M. 2013. Alternative forms for genomic clines. Ecol. Evol. 3:1951–1966.

Fitzpatrick B.M., Ryan M.E., Johnson J.R., Corush J., Carter E.T. 2015. Hybridization and the species problem in conservation. Curr. Zool. 61:206–216.

Fortney S.T. 2015. A century of geomorphic change of the San Rafael River and implications for river rehabilitation. Thesis.

Francis T.A., Bestgen K.R., White G.C. 2016. Population status of humpback chub, *Gila cypha*, and catch indices and population structure of sympatric roundtail chub, *Gila robusta*, in Black Rocks, Colorado River, Colorado, 1998-2012. Larval Fish Lab. Contrib. 199. Final Rep. from U.S. Fish Wildl. Serv. to Up. Color. River Endanger. Fish Recover. Program, Proj. Number 131. Gd. Junction, Color.

Gerber A.S., Tibbets C.A., Dowling T.E. 2001. The role of introgressive hybridization in the evolution of the *Gila robusta* complex (Teleostei: Cyprinidae). Evolution. 55:2028–2039.

Gerig B., Dodrill M.J., Pine W.E. 2014. Habitat selection and movement of adult humpback chub in the Colorado River in Grand Canyon, Arizona, during an experimental steady flow release. North Am. J. Fish. Manag. 34:39–48.

Gilbert K.J., Whitlock M.C. 2017. The genetics of adaptation to discrete heterogeneous environments: frequent mutation or large-effect alleles can allow range expansion. J. Evol. Biol. 30:591–602.

Gompert Z., Alex Buerkle C. 2010. Introgress: A software package for mapping components of isolation in hybrids. Mol. Ecol. Resour. 10:378–384.

Gompert Z., Buerkle C.A. 2009. A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. Mol. Ecol. 18:1207–1224.

Grabenstein K.C., Taylor S.A. 2018. Breaking barriers: Causes, consequences, and experimental utility of human-mediated hybridization. Trends Ecol. Evol. 33:198–212.

Gregorius H.R., Roberds J.H. 1986. Measurement of genetical differentiation among subpopulations. Theor. Appl. Genet. 71:826–834.

Hamilton J.A., Miller J.M. 2016. Adaptive introgression as a resource for management and genetic conservation in a changing climate. Conserv. Biol. 30:33–41.

Hamman R.L. 1981. Hybridization of three species of chub in a hatchery. Progress. Fish-Culturist. 43:131–134.

Hoffmann A.A., Sgrò C., M. 2011. Climate change and evolutionary adaptation. Nature. 470:479–485.

Holden P.B., Stalnaker C.B. 1970. Systematic studies of the genus *Gila*, in the Upper Colorado River Basin. Copeia. 1970:409–420.

Hopken M.W., Douglas M.R., Douglas M.E. 2013. Stream hierarchy defines riverscape genetics of a North American desert fish. Mol. Ecol. 22:956–971.

Johnston C.E., Page L.M. 1992. The evolution of complex reproductive strategies in North American minnows (Cyprinidae). In: Mayden R.L., editor. Systematics, Historical Ecology, and North American Freshwater Fishes. Stanford University Press. p. 601–621.

Kaeding L.R., Burdick B.D., Schrader P.A., McAda C.W. 1990. Temporal and spatial relations between the spawning of humpback chub and roundtail chub in the upper Colorado River. Trans. Am. Fish. Soc. 119:134–144.

Kaeding L.R., Zimmerman M.A. 1983. Life history and ecology of the humpback chub in the Little Colorado and Colorado rivers of the Grand Canyon. Trans. Am. Fish. Soc. 112:577–594.

Karp C.A., Tyus H.M. 1990. Humpback chub (*Gila cypha*) in the Yampa and Green rivers, Dinosaur National Monument, with observations on roundtail chub (*G. robusta*) and other sympatric fishes. Gt. Basin Nat. 50:257–264.

Kearns A.M., Restani M., Szabo I., Schrøder-Nielsen A., Kim J.A., Richardson H.M., Marzluff J.M., Fleischer R.C., Johnsen A., Omland K.E. 2018. Genomic evidence of speciation reversal in ravens. Nat. Commun. 9:906.

Klein E.K., Lagache-Navarro L., Petit R.J. 2017. Demographic and spatial determinants of hybridization rate. J. Ecol. 105:29–38.

Klonner G., Dullinger I., Wessely J., et al. 2017. Will climate change increase hybridization risk between potential plant invaders and their congeners in Europe? Divers. Distrib. 23:934–943.

Kopelman N.M., Mayzel J., Jakobsson M., Rosenberg N.A., Mayrose I. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. 15:1179–1191.

Kramer A.M., Dennis B., Liebhold A.M., Drake J.M. 2009. The evidence for Allee effects. Popul. Ecol. 51:341–354.

Mallet J., Besansky N., Hahn M.W. 2016. How reticulated are species? BioEssays. 38:140–149.

Mandeville E.G., Parchman T.L., Thompson K.G., Compton R.I., Gelwicks K.R., Song S.J., Buerkle C.A. 2017. Inconsistent reproductive isolation revealed by interactions between *Catostomus* fish species. Evol. Lett.:1–14.

McElroy D.M., Douglas M.E. 1995. Patterns of morphological variation among endangered populations of *Gila robusta* and *Gila cypha* (Teleostei : Cyprinidae) in the upper Colorado River basin. Copeia. 1995:636–649.

McElroy D.M., Shoemaker J.A., Douglas M.E. 1997. Discriminating *Gila robusta* and *Gila cypha*: Risk assessment and the Endangered Species Act. Ecol. Appl. 7:958–967.

Mckee E.D. 1972. Pliocene uplift of the Grand Canyon region- Time of drainage adjustment. Geol. Soc. Am. Bull. 83:1923–1932.

Meier J.I., Marques D.A., Mwaiko S., Wagner C.E., Excoffier L., Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat. Commun. 8:1–11.

Meko D.M., Woodhouse C.A., Baisan C.A., Knight T., Lucas J.J., Hughes M.K., Salzer M.W. 2007. Medieval drought in the upper Colorado River Basin. Geophys. Res. Lett. 34:1–5.

Miller R.R. 1946. *Gila cypha*, a remarkable new species of cyprinid fish from the Colorado River in Grand Canyon, Arizona. J. Washingt. Acad. Sci. 36:409–415.

Minckley C.O. 1996. Observations on the biology of the humpback chub in the Colorado River Basin 1980-1990. Dissertation.

Minckley W.L. 1986. Geography of western North American freshwater fishes: description and relationships to intracontinental tectonism. Zoogeography North Am. Freshw. Fishes.:519–613.

Moran C.J., Gerry S.P., Neill M.W.O., Rzucidlo C.L., Gibb A.C. 2018. Behavioral and physiological adaptations to high-flow velocities in chubs (*Gila* spp.) native to Southwestern USA. J. Exp. Biol. 221.

Nachman M.W. 2002. Variation in recombination rate across the genome: Evidence and implications. Curr. Opin. Genet. Dev. 12:657–663.

Olden J.D., Poff N.L., Bestgen K.R. 2008. Trait synergisms and the rarity, extirpation, and extinction risk of desert fishes. Ecology. 89:847–856.

Orr H.A., Unckless R.L. 2014. The population genetics of evolutionary rescue. PLoS Genet. 10:1–9.

Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14:1–24.

Pecl G.T., Araújo M.B., Bell J.D., et al. 2017. Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. Science. 355.

Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. 2012. Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS One. 7:e37135.

Polechová J., Barton N.H. 2015. Limits to adaptation along environmental gradients. Proc. Natl. Acad. Sci. 112:6401–6406.

Portz D.E., Tyus H.M. 2004. Fish humps in two Colorado River fishes: A morphological response to cyprinid predation? Environ. Biol. Fishes. 71:233–245.

Pritchard J.K., Stephens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Schumer M., Rosenthal G.G., Andolfatto P. 2018. What do we mean when we talk about hybrid speciation? Heredity. 120:379–382.

Seehausen O., Takimoto G., Roy D., Jokela J. 2008. Speciation reversal and biodiversity dynamics with hybridization in changing environments. Mol. Ecol. 17:30–44.

Smith G.R., Miller R.R., Sable W.D. 1979. Species relationships among fishes of the genus *Gila* in the upper Colorado River drainage. Proc. First Conf. Sci. Res. Natl. Park. Vol. II.:613–623.

Spencer J.E., Smith G.R., Dowling T.E. 2008. Middle to late Cenozoic geology, hydrography, and fish evolution in the American Southwest. Geol. Soc. Am. 80301.

Stanley R.R.E., Jeffery N.W., Wringe B.F., DiBacco C., Bradbury I.R. 2017. genepopedit: a simple and flexible tool for manipulating multilocus molecular data in R. Mol. Ecol. Resour.

Thomaz A.T., Christie M.R., Knowles L.L. 2016. The architecture of river networks can drive the evolutionary dynamics of aquatic populations. Evolution. 70:731–739.

Todesco M., Pascual M.A., Owens G.L., Ostevik K.L., Moyers B.T., Hübner S., Heredia S.M., Hahn M.A., Caseys C., Bock D.G., Rieseberg L.H. 2016. Hybridization and extinction. Evol. Appl. 9:892–908.

Tyus H.M. 1998. Early records of the endangered fish *Gila cypha* Miller from the Yampa River of Colorado with notes on its decline. Copeia. 1998:190–193.

Tyus H.M., Burdick B.D., Valdez R.A., Haynes C.M., Lytle T.A., Berry C.R. 1982. Fishes of the upper Colorado River basin: Distribution, abundance, and status. Fishes Up. Color. River Syst. Present Futur.

Unmack P.J., Dowling T.E., Laitinen N.J., Secor C.L., Mayden R.L., Shiozawa D.K., Smith G.R. 2014. Influence of introgression and geological processes on phylogenetic relationships of Western North American mountain suckers (Pantosteus, Catostomidae). PLoS One. 9.

USFWS. 2011. 5-year Review: Humpback chub (*Gila cypha*) summary and evaluation. Up. Color. River Endanger. Fish Recover. Progr. Denver, Color.

USFWS. 2017. Species status assessment for the humpback chub (*Gila cypha*). U. S. Fish Wildl. Serv. Mt. Reg. (6), Denver, CO.

Uyeno T., Miller R.R. 1963. Summary of late Cenozoic freshwater fish records for North America. Occas. Pap. Museum Zool. Univ. Michigan. 631:1–34.

Uyeno T., Miller R.R. 1965. Middle Pliocene cyprinid fishes from the Bidahochi Formation. Copeia.:28–41.

Valdez R.A., Clemmer G.H. 1982. Life history and prospects for recovery of the humpback and bonytail chub. Fishes of the Upper Colorado River System: Present and Future. p. 109–119.

Valdez R.A., Holden P.B., Hardy T.B. 1990. Habitat suitability index curves for humpback chub of the upper Colorado River basin. Rivers. 1:31–42.

Vanicek C.D., Kramer R.H. 1969. Life history of the Colorado Squawfish, *Ptychocheilus lucius*, and the Colorado Chub, *Gila robusta*, in the Green River in Dinosaur National Monument, 1964-1966. Trans. Am. Fish. Soc. 98:193–208.

vonHoldt B.M., Brzeski K.E., Wilcove D.S., Rutledge L.Y. 2018. Redefining the role of admixture and genomics in species conservation. Conserv. Lett. 11:1–6.

Wayne R.K., Shaffer H.B. 2016. Hybridization and endangered species protection in the molecular era. Mol. Ecol. 25:2680–2689.

Wringe B.F., Stanley R.R.E., Jeffery N.W., Anderson E.C., Bradbury I.R. 2017. hybriddetective: A workflow and package to facilitate the detection of hybridization using genomic data in R. Mol. Ecol. Resour. 17:e275–e284.

Appendix

Table 1: Sampling locations for *Gila robusta*, *G. cypha* and *G. elegans*. Site=abbreviated locality identifier for each species, Major Drainage=River, Location=geographic site, County, State=per sampling site, and N=Number of samples excluding those that sequenced with sufficient coverage. (*) denotes sympatric localities

Site	Major Drainage	Location	County, State	*N*
Gila robusta				
*RBKR	Colorado	Black Rocks Canyon	Mesa, CO	11
*RDES	Green	Desolation Canyon	Uintah, UT	23
RC15	Colorado	15-mile reach	Mesa, CO	10
RECC	Little Colorado	East Clear Creek	Coconino, AZ	16
RMCO	San Juan	Mancos River	Montezuma, CO	10
RMGR	Green	Middle Green R. tributaries	Sweetwater, WY	24
RNJA	San Juan	Navajo R.	Rio Arriba, NM	10
RLSR	Yampa	Little Snake R. tributaries	Carbon, WY	31
RLYC	Yampa	Little Yampa Canyon	Moffat, CO	11
RSRR	Green	San Rafael R.	Emery, UT	23
RUGR	Green	Upper Green R. tributaries	Sublette, WY	16
RWRW	White	White R. mainstem near Weaver Cn.	Uintah, UT	15
*RWWC	Colorado	Colorado mainstem near Westwater Cn.	Grand, UT	11
*RYAM	Yampa	Yampa R. mainstem	Moffat, CO	15
Gila cypha				
*HBKR	Colorado	Black Rocks Canyon	Mesa, CO	18
*HDES	Green	Desolation Canyon	Uintah, UT	24
HGCN	Colorado	Grand Canyon	Coconino, AZ	37
HLCR	Little Colorado	Atomizer Falls and Colorado confluence	Coconino, AZ	22
*HWWC	Colorado	Westwater Canyon	Grand, UT	22
*HYAM	Yampa	Yampa Canyon	Moffat, CO	8
Gila elegans	Hatchery stock	USFWS Hatchery at Dexter, NM	Chaves, NM	11

Table 2: Proportions of *Gila robusta* and *G. cypha* assigned genealogically at each sample site. Site=abbreviated locality identifier for each species; P0=Pure *robusta*; P1=Pure *cypha*; F1=First filial hybrid; F2=second filial hybrid; B0=*G. robusta*-backcrossed hybrid; B1=*G. cypha*-backcrossed hybrid; FN=late-generation or uncertain hybrid. Samples were assigned to a genealogical class per posterior probability ≥0.80, as assessed using 250,000 post burn-in MCMC generations in NEWHYBRIDS.

Site	**P_0**	**P_1**	**F_1**	**F_2**	**B_0**	**B_1**	**F_N**
HLCR	-	1.000	-	-	-	-	-
HGCN	-	1.000	-	-	-	-	-
HDES	-	0.083	-	0.083	-	0.542	0.292
HWWC	-	0.150	-	0.200	-	0.550	0.100
HBKR	-	0.615	-	0.231	-	0.077	0.077
HYAM	-	-	-	0.667	-	0.333	-
RDES	0.046	0.046	-	0.136	-	0.364	0.409
RSRR ’17	-	-	-	0.083	0.667	-	0.250
RSRR ’09	-	-	-	0.455	0.455	-	0.091
RNJA	1.000	-	-	-	-	-	-
RMCO	0.800	-	-	-	0.100	-	0.100
RWWC	0.875	-	-	-	0.125	-	0.375
RC15	1.000	-	-	-	-	-	-
RBKR	0.857	-	-	-	-	-	0.143
RECC	0.938	-	-	-	0.063	-	-
RWRW	0.875	-	-	-	0.125	-	-
RUGR	1.000	-	-	-	-	-	-
RMGR	0.955	-	-	-	-	-	0.045
RYAM	1.000	-	-	-	-	-	-
RLYC	1.000	-	-	-	-	-	-
RLSR	1.000	-	-	-	-	-	-

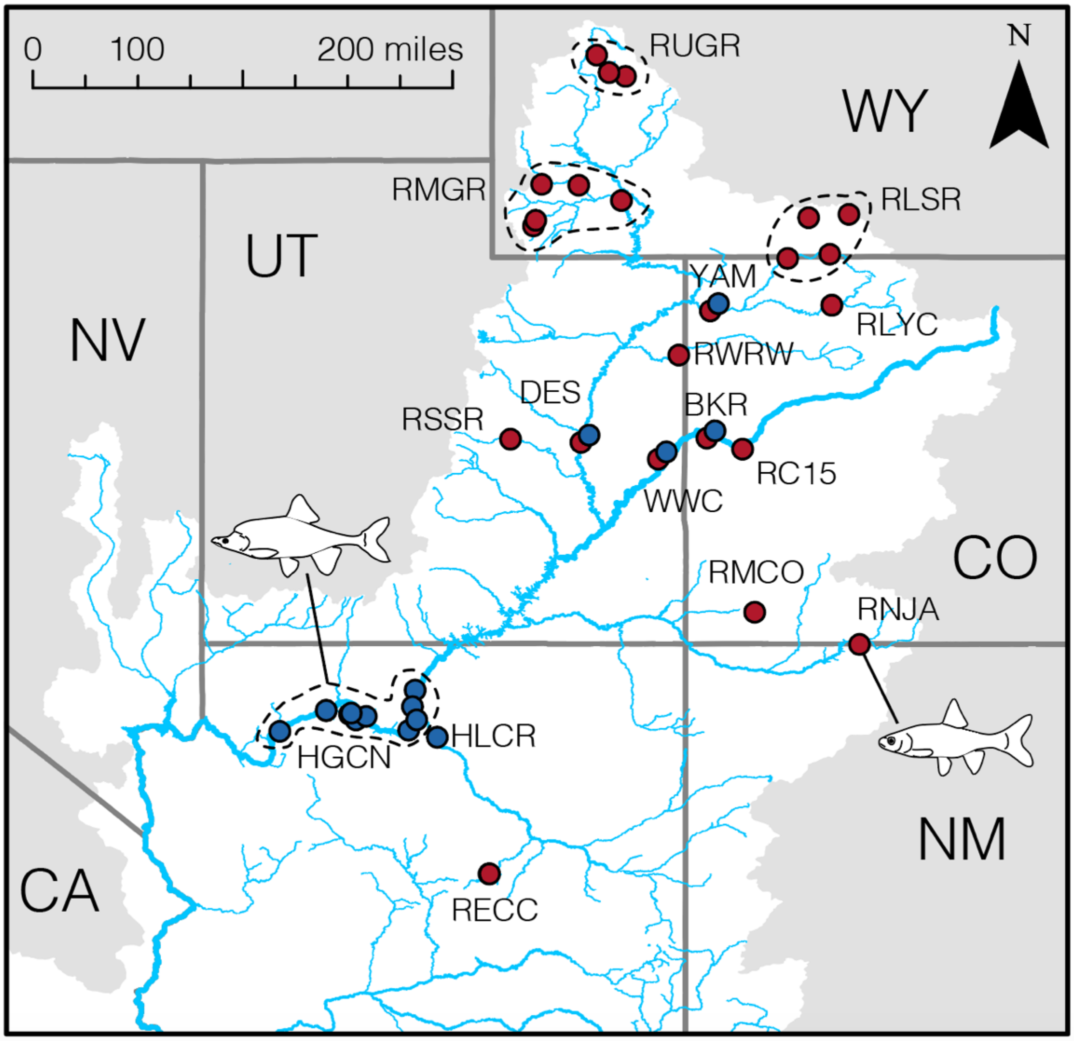

Figure 1: Sampling localities for *Gila cypha* (blue) and *G. robusta* (red) within the Colorado River Basin, western North America. Locality codes are defined in Table 2. Sympatric locations (BKR, DES, WWC, YAM) are slightly offset for visibility purposes. Inset cartoons the respective morphologies of each species

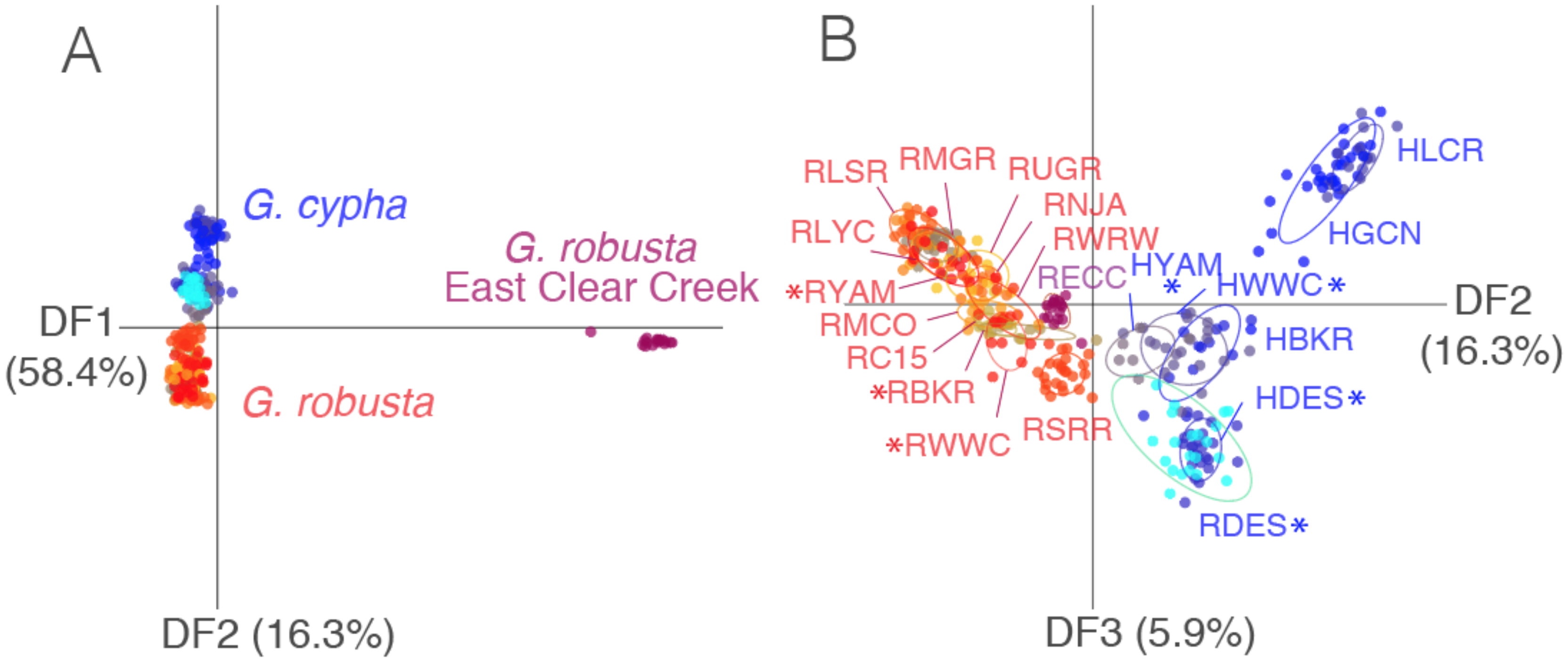

Figure 2: Results of a Discriminate Analysis of Principal Components (DAPC) analysis depicting *Gila robusta* (red), *G. cypha* (blue), and their respective populations (as colored). (A) discriminant function axes 1 and 2 (=DF1xDF2) showing discrimination among both species; (B) axes 2 and 3 (=DF2xDF3) reflecting the manner by which populations of each species (grouped within ellipses) are distributed in discriminant space. The relative percent variance captured by each discriminant function is presented in parentheses. Sample localities are defined in Table 1. (*) denotes sympatric localities

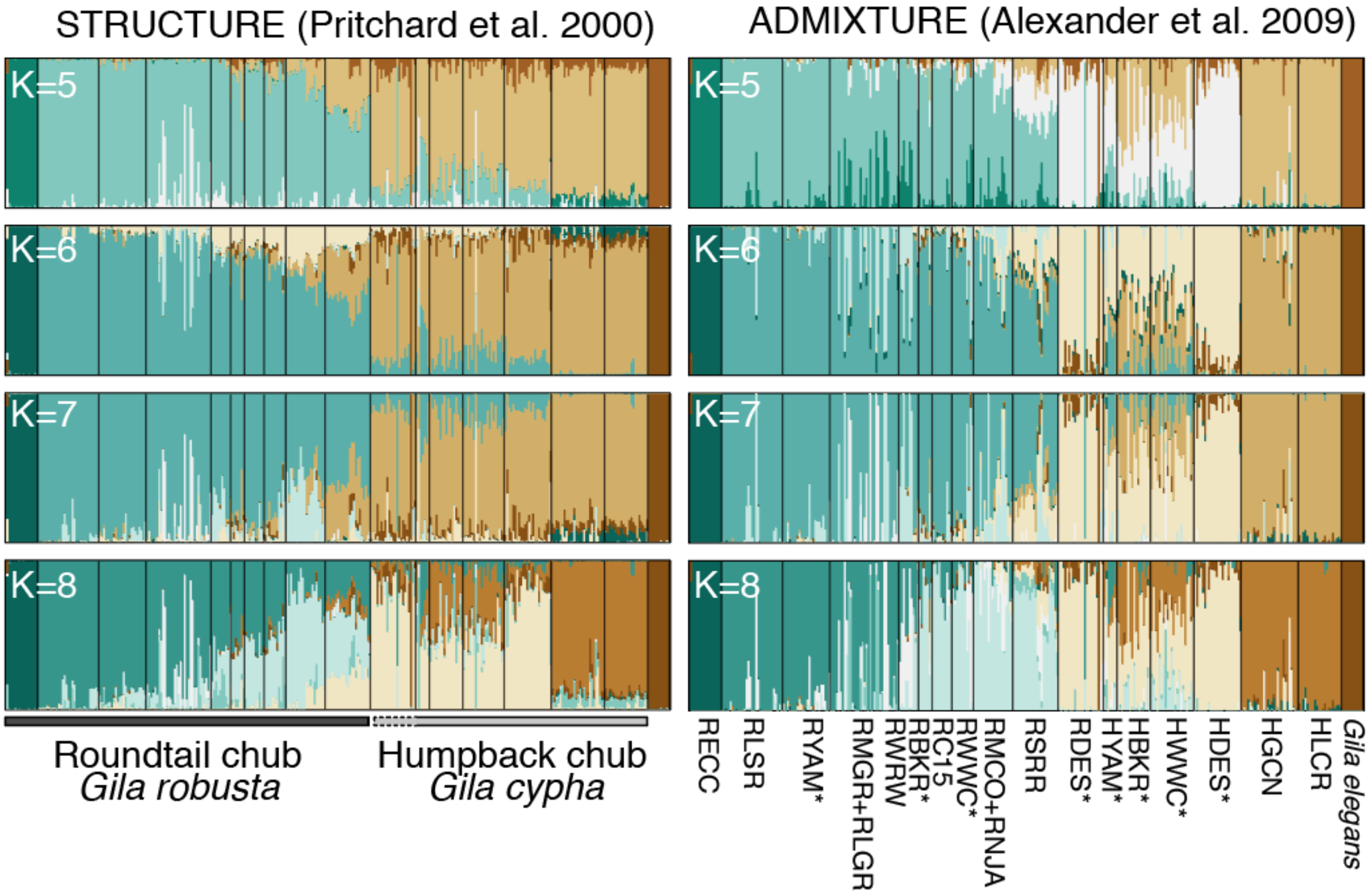

Figure 3: Assignment results for ADMIXTURE and STRUCTURE analyses involving *Gila robusta*, *G. cypha*, and *G. elegans*. K-values range from STRUCTURE optimum K=5 (see Fig. S2) to ADMIXTURE optimum K=8 (see Fig. S3). Locality abbreviations are as defined in Table 1. (*) denotes sympatric localities

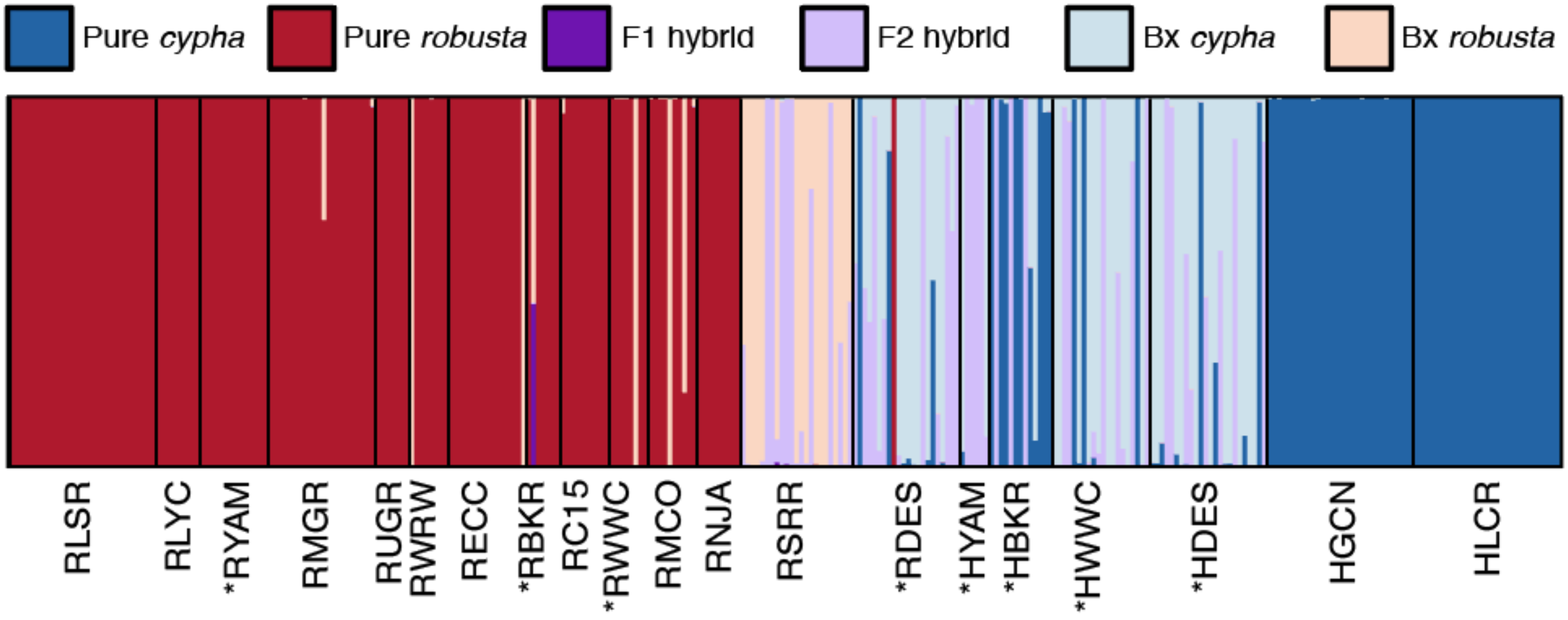

Figure 4: Genealogical assignment for individual *Gila robusta* and *G. cypha*, as compiled from NEWHYBRIDS analysis. Individuals are represented by colored bars, with proportion of color indicating posterior probability of assignment per genealogical class. Prior 'parental' allele frequencies for *G. cypha* were derived from the Little Colorado River (HLCR) and from the Little Snake River (RLSR) for *G. robusta* (alternative prior assignments had no significant affect; see Fig. S4 and S5). Colors are as follows: Red=pure *G. robusta*; Blue=pure *G. cypha*; Purple=F1 hybrid; Light purple=F2 hybrid; Light blue=cypha-backcrossed hybrid; Light red=robusta-backcrossed hybrid. (*) denotes sympatric localities

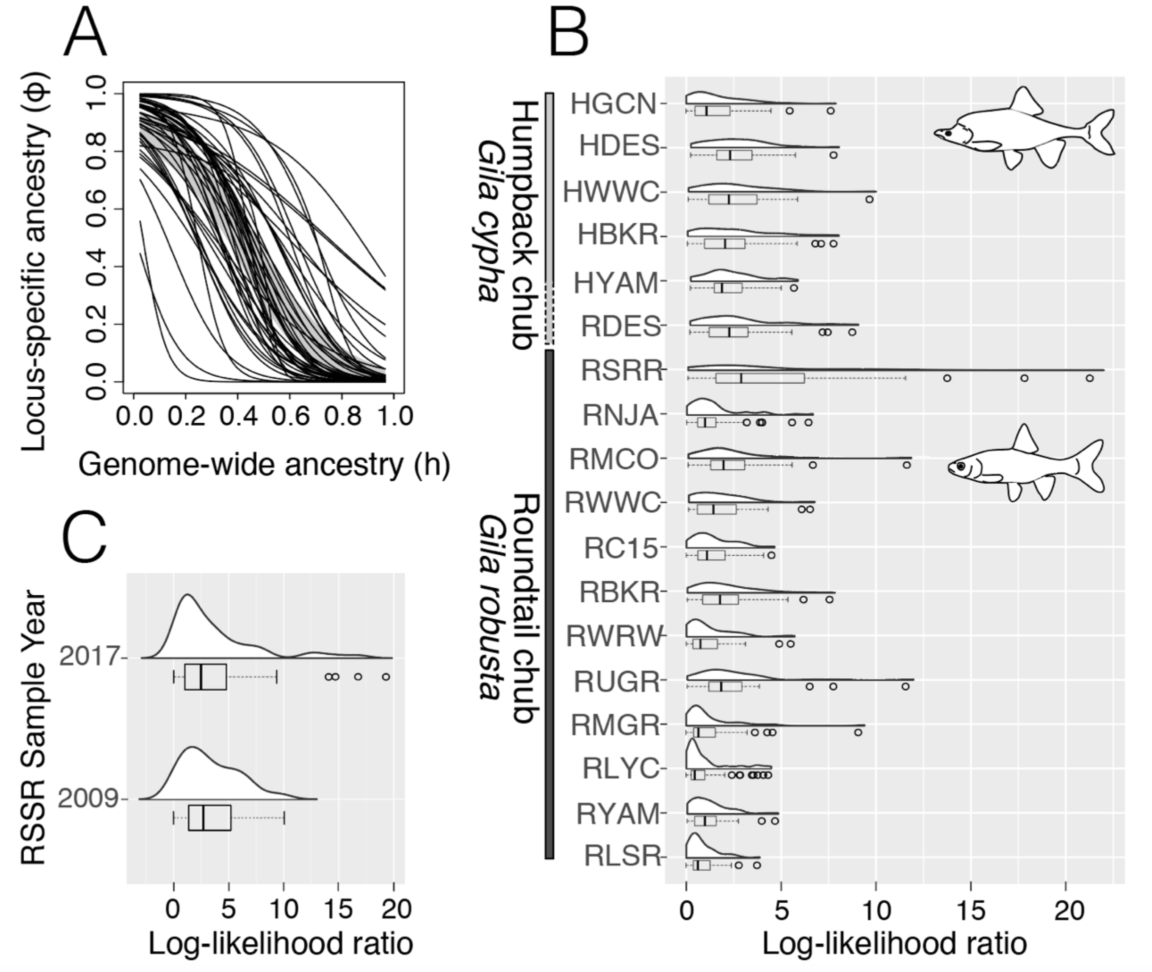

Figure 5: Genomic cline analyses for populations of *Gila robusta* and *G. cypha*, presented as: (A) Per-locus clinal relationships for 50 SNPs with δ >0.8 (all significantly non-neutral at α=0.001) compared to the neutral expectation (shaded gray region); (B) Log-likelihood ratio distribution of site-wise per-locus clines compared to the global pattern, where higher log-likelihood ratio indicates greater discordance; (C) Per-locus incongruence in genomic clines in the *Gila robusta* samples from the San Rafael River, partitioned by year (2009 versus 2017). Locality codes for populations of each species are defined in Table 1.

CHAPTER II

Taxonomic uncertainty and the anomaly zone: Phylogenomics resolve rapid radiation and hybrid origin in a contentious species complex

Chafin TK, Douglas MR, Bangs MR, Mussmann SM, Douglas ME

Abstract

Species are an indisputable unit for biodiversity conservation, yet their delimitation is fraught with both conceptual and methodological difficulties. A classic example is the taxonomic controversy surrounding the *Gila robusta* complex in the lower Colorado River of southwestern North America. Nominal species designations were originally defined according to weakly diagnostic morphological differences that conflicted with traditional genetic analyses. Consequently, the complex was re-defined as a single polytypic unit, with the proposed 'threatened' status of two being withdrawn at the federal level. Here, we utilized dense spatial and genomic sampling (N=387 and >22k loci) to re-evaluate the status of the complex, based on SNP-based coalescent and polymorphism-aware phylogenetic models. In doing so, all three species were supported as evolutionarily independent lineages, despite widespread phylogenetic discordance. To understand this discrepancy with past studies, we categorized evolutionary mechanisms driving discordance. We tested (and subsequently rejected) prior hypotheses suggesting that phylogenetic discord in the complex was hybridization-driven. Instead, we found the *G. robusta* complex to have diverged within the 'anomaly zone' of tree space and, as such, have accumulated inconsistent patterns of diversity which have confounded prior studies. After extending these analyses with phylogeographic modeling, we propose that this is reflective of a rapid radiation promoted by Plio-Pleistocene tectonism. Our results not only support resurrection

of the three species as distinct entities, but also offer an empirical example of how phylogenetic discordance can be categorized in other recalcitrant taxa.

Introduction

Complex evolutionary histories remain consistently difficult to disentangle, despite a recent paradigm shift towards the development of increasingly comprehensive datasets (e.g. Edwards 2009; Giarla and Esselstyn 2015). Regardless of these efforts, phylogenetic uncertainty is still prevalent, and with wide-ranging consequences on the study of macroevolutionary patterns (Stadler et al. 2016; Pereira and Schrago 2018), trait evolution (Hahn and Nakhleh 2016; Mendes et al. 2016; Wu et al. 2018), and ecological and biogeographic processes (Rangel et al. 2015; McVay et al. 2017).

Importantly, phylogenetic uncertainty also translates to taxonomic uncertainty. This is because modern systematic taxonomy fundamentally describes homology [i.e. Darwin's (1859) ‘propinquity of descent’ (Simpson 1961)], which, by definition, requires a phylogenetic context. Phylogenetic uncertainty in this sense can manifest itself as a soft polytomy (= ‘honest’ uncertainty), the erroneous promotion of non-monophyletic clades, or controversial ‘splitting’ *versus* ‘lumping.’ Incomplete or biased sampling is often a driver of this disparity (Ahrens et al. 2016; Reddy et al. 2017). Here, narrow taxon sampling may introduce substantial ascertainment bias (=systematic deviations due to sampling). On the other hand, a broader yet sparse sampling regime often fails to sample cryptic lineages (Heath et al. 2008) — with subsequent impacts on both the delimitation of species (Pante et al. 2015; Linck et al. 2019) and study of their traits (Beaulieu and O’Meara 2018).

These sources of uncertainty culminate in topologies that often fluctuate with regard to sampling designs or methodologies, and this results in taxonomic uncertainty [e.g. Ctenophora versus Porifera as sister to all other animals (Pisani et al. 2015; Whelan et al. 2015; Simion et al. 2017)]. Access to genome-scale data has alleviated some of these issues by offering a level of precision not possible with single-gene phylogenies (Philippe et al. 2005). However, their inherent complexity and heterogeneity introduces new problems, and consequently, additional sources of phylogenetic uncertainty.

Gene tree heterogeneity is a ubiquitous source of discordance in genomic data, and "noise" as a source of this variance must consequently be partitioned from "signal" (where "noise" is broadly categorized as systematic or stochastic error). Large genomic datasets can reduce stochastic error (Kumar et al. 2012), yet it still remains a prevalent issue when individual genes are examined (Springer and Gatesy 2016). On the other hand, systematic error in phylogenomics may represent a probabilistic bias towards incongruence that is inherent to the evolutionary process itself (Maddison 1997). This, in turn, exemplifies the complications introduced by genomic data: As genomic resolution increases, so also does the probability of sampling unmodeled processes (Rannala and Yang 2008; Lemmon and Lemmon 2013). This potential (i.e., simultaneously decreasing stochastic error as systematic error increases) produces the very real possibility of building a highly supported tree that is ultimately incorrect.

Certain demographic histories are more predisposed to systematic error than others. For instance, when effective population sizes are large and speciation events exceptionally rapid, time between divergence events may be insufficient to sort ancestral variation, such that the most probable gene topology will conflict with the underlying species branching pattern. This results in what has been coined an "anomaly zone" of tree space (i.e., dominated by anomalous gene trees (AGTs); Degnan and Rosenberg 2006). Inferring species trees is demonstrably difficult in this region (Liu and Edwards 2009), and exceedingly so if additional sources of phylogenetic discordance, such as hybridization, are also apparent (Bangs et al. 2018).

In clades with such complex histories, it is often unclear where the source of poor support and/or topological conflict resides. Yet, to analytically account for gene tree conflict, it is necessary to categorize these sources and select approaches accordingly. Failure to do so promotes a false confidence in an erroneous topology, as driven by model misspecification (Philippe et al. 2011). The overwhelmingly parametric nature of modern phylogenetics insures that imperative issues will revolve around the processes being modeled, and what they actually allow us to ask from our data (Sullivan and Joyce 2005). However, the selection of methods that model processes of interest requires an *a priori* hypothesis that delimits which processes are involved. Yet, diagnosing prominent processes is difficult in that a phylogenetic context is required from which to build hypotheses. Fortunately, a wealth of information can be parsed from otherwise “non-phylogenetic” signal (*sensu* Philippe et al. 2005). For example, many statistical tests diagnose hybridization via its characteristic signature on the distribution of discordant topologies (e.g. Pease and Hahn 2015). Theoretical predictions regarding AGTs and the parameters under which they are generated are also well characterized (Degnan and Salter 2005; Degnan and Rosenberg 2009). Thus, by applying appropriate analytical approaches that sample many independently segregating regions of the genome, empiricists can still derive biologically meaningful phylogenies, despite the presence of complicated species-histories (McCormack et al. 2009; Kumar et al. 2012).

Here, we demonstrate an empirical approach that infers species-histories and sources of subtree discordance when conflict originates not only from anomaly zone divergences but also hybridization. To do so, we used SNP-based coalescent and polymorphism-aware phylogenetic methods (Chifman and Kubatko 2014; Leache et al. 2014; De Maio et al. 2015) that bypass the necessity of fully-resolved gene trees. We combine coalescent predictions, phylogenetic network inference (Solís-Lemus and Ané 2016), and novel coalescent phylogeographic methods (Oaks 2018) to diagnose the sources of phylogenetic discordance and, by so doing, resolve a seemingly convoluted complex of study-species (the *Gila robusta* complex of the lower Colorado River). We then contextualize our results to demonstrate the downstream implications of 'problematic' tree-space for threatened and endangered taxa, as represented by our study complex.

Gila

Few freshwater taxa have proven as problematic in recent years as the *Gila robusta* complex (Cyprinoidea: Leuciscidae) endemic to the Gila River basin of southwestern North America (Fig. 1). The taxonomic debate surrounding this complex exemplifies an inherent conflict between the traditional rigidity of systematic taxonomy *versus* the urgency of decision-making for conservation and management (Forest et al. 2015). Our study system is the Gila River, a primary tributary of the lower basin Colorado River that drains the majority of Arizona and ~11% of New Mexico. The critical shortage of water in this region (Sabo et al. 2010) is a major geopolitical driver for the taxonomic controversy surrounding the study species. As an example, the lower Colorado basin is responsible for approximately half of the total municipal and agricultural water requirements of the state of Arizona, and nearly two-thirds of its total gross state product (GSP) (Bureau of Reclamation 2012; James et al. 2014). This disproportionate regional reliance creates tension between the governance of a resource and its usage (e.g. Huckleberry and Potts 2019) which in turn magnifies the stakes involved in conservation policy (Minckley 1979; Carlson and Muth 1989; Minckley et al. 2006).

We focused on three species (Roundtail chub, *G. robusta*; Gila chub, *G. intermedia*; and Headwater chub, *G. nigra*) that comprise a substantial proportion of the endemic ichthyfauna of the Gila Basin [=20% of 15 extant native species (excluding extirpated *G. elegans* and *Xyrauchen texanus*); Minckley and Marsh 2009]. Historically, the focal taxa have been subjected to numerous taxonomic rearrangements (Fig. 1). Until recently, the consensus was defined by Minckley and DeMarais (2000) on the basis of morphometric and meristic characters. These have since proven of limited diagnostic capacity in the field, thus provoking numerous attempts to re-define morphological delimitations (Brandenburg et al. 2015; Moran et al. 2017; Carter et al. 2018). Genetic evaluations have to date been unproductive (Schwemm 2006; Copus et al. 2018), leading to a recent taxonomic recommendation that subsequently collapsed the complex into a single polytypic species (Page et al. 2016, 2017).

Methods

Taxonomic Sampling

A representative panel of *N*=386 individuals (Table S1; Fig. 2) was chosen from existing collections (Douglas et al. 2001; Douglas and Douglas 2007), to include broad geographic sampling of the complex as well as congeners. For the sake of clarity, we employ herein the nomenclature of Minckley and DeMarais (2000) and retained species-level nomenclature for all members of the *Gila robusta* complex. Additionally, we discriminate between *G. robusta* from the upper and lower basins of the Colorado River ecosystem (Chafin et al. 2019)

No self-sustaining populations of wild *Gila elegans* exist, thus samples were provided by the Southwestern Native Aquatic Resources and Recovery Center (Dexter, NM). The genus *Ptychocheilus* served to root the *Gila* clade within the broader context of western leuciscids (Schönhuth et al. 2012, 2014, 2018).

Reduced-Representation Sequencing

Genomic DNA was extracted using either PureGene® or DNeasy® kits (Qiagen Inc.) and quantified via fluorometer (Qubit™; Thermo-Fisher Scientific). Library preparations followed the published ddRAD protocol (Peterson et al. 2012). Restriction enzyme and size-selection ranges were first screened using an *in silico* procedure (Chafin et al. 2018), with the target fragment sizes further optimized by quantifying digests for 15 representative samples on an Agilent 2200 TapeStation. Final library preparations were double-digested using a high-fidelity *PstI* (5'-CTGCAG-3') and *MspI* (5'-CCGG-3') following manufacturer's protocols (New England Biosciences). Digests were purified using bead purification (Ampure XP; Beckman-Coulter Inc.), and standardized at 100 ng per sample. Samples were then ligated with customized adapters containing unique in-line barcodes, pooled in sets of 48, and size-selected at 250-350bp (not including adapter length), using a Pippin Prep automated gel extraction instrument (Sage Sciences). Adapters were then extended in a 12-cycle PCR using Phusion high-fidelity DNA polymerase (New England Biosciences Inc.), completing adapters for Illumina sequencing and adding an i7 index. Libraries were pooled to N=96 samples per lane (i.e., 2 sets of 48) for 100bp single-end sequencing on an Illumina HiSeq 2500 at the University of Wisconsin Biotechnology Center (Madison, WI).

Data Processing and Assembly

Raw Illumina reads were demultiplexed and filtered using the PYRAD pipeline (Eaton 2014). We removed reads containing >1 mismatch in the barcode sequence, or >5 low-quality base-calls (Phred Q<20). Homologs assembly was then performed using *de novo* clustering in VSEARCH (Rognes et al. 2016) using an 80% mismatch threshold. Loci were excluded according to following criteria: >5 ambiguous nucleotides; >10 heterozygous sites in the alignment; >2 haplotypes per individual; <20X and >500X sequencing depth per individual; >70% heterozygosity per-site among individuals.

Our ddRAD approach generated 22,768 loci containing a total of 173,719 variable sites, of which 21,717 were sampled (=1/ locus). Mean per-individual depth of coverage across all retained loci was 79X. All relevant scripts for post-assembly filtering and data conversion are available as open-source (github.com/tkchafin/scripts).

Phylogenetic Inference

We formulated two simple hypotheses with regards to independent evolutionary sub-units. If populations represented a single polytypic species, then phylogenetic clustering should reflect intraspecific processes (e.g. structured according to stream heirarchy; Meffe and Vrijenhoek 1988). However, if *a priori* taxon assignments are evolutionarily independent, then they should be recapitulated in the phylogeny. Given well-known issues associated with application of supermatrix/ concatenation approaches (Degnan and Rosenberg 2006; Edwards et al. 2016) and pervasive gene-tree uncertainty associated with short loci (Leaché and Oaks 2017), we also employed SNP-based methods that bypassed the derivation of gene trees (Leaché and Oaks 2017).

We first explored population trees in SVDQUARTETS (Chifman and Kubatko 2014, 2015; as implemented in PAUP*, Swofford 2002) across 12 variably filtered datasets using four differing occupancy thresholds per SNP locus (i.e., 10, 25, 50, and 75%), along with three differing thresholds per individual (10, 25, and 50%). These filtered datasets ranged from 7357–21007 SNPs, with 8.48–43.65% missing data and 256–347 individuals. SVDQUARTETS eases computation by inferring coalescent trees from randomly sampled quartets of species (i.e. optimizing among 3 possible unrooted topologies). It then generates a population tree with conflicts among quartet trees minimized via implementation of a quartet-assembly algorithm (Snir and Rao 2012). Given run-time constraints (the longest was 180 days on 44 cores), all runs sampled $\binom{N_{tips}}{4}/2$ quartets and were evaluated across 100 bootstrap pseudo-replicates.

We also used a polymorphism-aware method (POMO; Schrempf et al. 2016) in IQ-TREE (Nguyen et al. 2014). POMO considers allele frequencies rather than single nucleotides, thus allowing evaluation of change due to both substitution and drift. To provide POMO with empirical estimates of polymorphism, we used the entire alignment, to include non-variable sequences. We filtered liberally using individual occupancy thresholds of 10% per-locus so as to maximize individual retention and per-population sample sizes. We then deleted populations that contained <2 individuals, and loci with >=90% missing data per-population. This yielded a dataset of 281,613 nucleotides and 40 tips. Non-focal outgroups were excluded due to their disproportionate effect on missing data.

We also calculated concordance factors (CFs) using a Bayesian concordance analysis in BUCKY (Larget et al. 2010), parallelized across all quartets via an adaptation of the TICR pipeline (Stenz et al. 2015). To prepare these data, we sampled all non-monomorphic full gene alignments for which at least 1 diploid genotype could be sampled per population. We excluded outgroups and non-focal *Gila* so as to maximize number of loci retained. This yielded 3,449 genes across 31 sampled tips. Gene-tree priors were generated using MRBAYES v.3.2.6 (Ronquist et al. 2012) with 4 independent chains, each of which was sampled every 10,000 iterations, with a total chain length of 100,000,000 iterations and 50% discarded as burn-in. BUCKY was then run in parallel to generate quartet CFs across 31,465 quartets, using a chain length of 10,000,000, again with 50% burn-in. Quartet topologies were used to generate a population tree using QUARTETMAXCUT (Snir and Rao 2012), using the *get-pop-tree.pl* script from TICR (Stenz et al. 2015; https://github.com/nstenz/TICR).

Comparing Phylogenies and Estimating Site-wise Conflict

To evaluate the performance of SVDQUARTETS, TICR, and PoMo, we first computed site-wise log-likelihood scores (*SLS*) for each topology by performing a constrained ML search in IQ-TREE. For comparison, we also generated an unconstrained concatenated tree. All ML analyses employed a GTR model with empirical base frequencies and gamma-distributed rates, and were assessed across 1,000 bootstrap pseudoreplicates. Analyses were also reduced to a subset of tips common across all variably filtered datasets. We quantified the phylogenetic signal supporting each resolution as the difference in site-wise log-likelihood scores (Δ*SLS*) between each population tree and the concatenation tree (Shen et al. 2017). We then calculated site-wise concordance factors (s*CF*) as an additional support metric (Minh et al. 2018).

Tests of Hybridization and Deep-Time Reticulation

D-statistics (Green et al. 2010; Eaton and Ree 2013) were calculated using COMP-D (Mussmann et al. 2019). To further test hypotheses of reticulation, we used quartet CFs as input for phylogenetic network inference using the SNAQ algorithm in PHYLONETWORKS (Solís-Lemus and Ané 2016; Solís-Lemus et al. 2017). The network was estimated under models of 0–5 hybrid nodes (*h*). Models were evaluated using 48 independent replicates, with the best-fit model being that which maximized change in pseudolikelihood. Given the computational constraints of network inference, we reduced the dataset to N=2 populations per focal species (=12 total tips).

Anomaly Zone Detection

Coalescent theory characterizes the boundaries of the anomaly zone in terms of branch lengths in coalescent units (Degnan and Rosenberg 2006). To test if contentious relationships in our tree fell within the anomaly zone, we first transformed branch lengths using quartet CFs (Stenz et al. 2015, equation 1), then tested if internode branch lengths fell within the theoretical boundary for the anomaly zone (Linkem et al. 2016, equation 1). Code for these calculations are modified from Linkem et al. (2016) and are available as open-source (github.com/tkchafin/anomaly_zone).

Tests of Co-divergence

The contemporary course of the Colorado River resulted from the Pliocene erosion of the Grand Canyon and subsequent connection of the modern-day upper and lower basins, to include stream capture of the Gila River (McKee et al. 1967; Minckley et al. 1986). *Gila* in the lower Colorado River basin then differentiated following one or more colonization events (e.g. Rinne 1976). Subsequent work (Douglas et al. 1999) supported this conclusion by examining contemporary phenotypic variation among all three species as a function of historical drainage connectivity, with the conclusion that body shape was most readily explained by Pliocene hydrography.

We tested if divergences were best explained by a model of *in situ* diversification following a single colonization event, or instead by multiple, successive colonizations. We compared divergence models using a Bayesian approach (program ECOEVOLITY, Oaks 2018) that used a coalescent model (Bryant et al. 2012) to update a prior expectation for the number of evolutionary events across independent comparisons. Four independent MCMC chains were run with recommended settings and a burn-in that maximized effective sample sizes. Event models followed a Dirichlet process, with the concentration parameter exploring four alternative gamma distributed priors (i.e. α=2.0, β=5.70; α=0.5, β=8.7; α=1.0, β=0.45; and α=2.0, β=2.18).

We randomly sampled 2,000 full-locus alignments, then examined potential co-divergences in the lower-basin complex by selecting a series of pairwise comparisons: *Gila elegans* x *G. robusta* (lower); *G. seminuda* x *G. robusta* (lower); *G. jordani* x *G. robusta* (lower); *G. intermedia* x *G. robusta* (lower); and *G. intermedia* x *G. nigra* (lower). These targeted nodes represent H, G, F, I, and N in the SVDQUARTETS topology (Fig. 3A).

Results

Phylogenetic Conflict in Gila

Tree reconstructions across all three population methods were relatively congruent (SVDQUARTETS = Fig. 3; TICR, and POMO = Fig. 4). The concatenated supermatrix tree (Fig. S1) was also largely congruent with the population trees, but with two major disparities (discussed below). Bootstrap support was variable and declined with decreasing nodal depth in the SVDQUARTETS analysis (Fig. 3), whereas the vast majority of nodes in POMO were supported at 100% (Fig. 4A).

All analyses consistently supported the monophyly of a clade consisting of *G. intermedia*, *G. nigra*, and lower basin *G. robusta* (hereafter the 'lower basin complex'). This clade had high bootstrap support in both SVDQUARTETS and POMO, and was universally placed as sister to *G. jordani*. *Gila robusta* was unequivocally polyphyletic in all analyses, forming two distinct groups geographically demarcated by the Grand Canyon. The lower basin *G. robusta* clade was monophyletic in all cases, save the concatenated tree, where it was paraphyletic (Fig. S8). It was also consistently recovered as sister to a monophyletic *G. nigra* + *G. intermedia*, with the exclusion of a single sample site (Aravaipa Creek) that nested within *G. intermedia* in the POMO tree. Of note, this population had been previously diagnosed as trending towards *G. intermedia* in terms of morphology (Rinne 1976; DeMarais 1986), although hybridization was not supported by *D*-statistics (Table 1).

Topology within the *G. nigra* + *G. intermedia* clade was less consistent. Both were reciprocally monophyletic in the SVDQUARTETS tree (albeit with low support; Fig. 3), whereas POMO yielded a monophyletic *G. intermedia*, with but one population (Spring Creek) contained within *G. nigra* (Fig. 4A). The POMO tree also conflicted with the other methods in its paraphyletic placement of upper basin *G. robusta*. We suspect this represents an artefact of well-known hybridization with sympatric *G. cypha* (Dowling and DeMarais 1993; Gerber et al. 2001; Douglas and Douglas 2007; Chafin et al. 2019).

Discriminating Among Sources of Phylogenetic Conflict

Phylogenetic conflict was variably attributable to either hybridization or rapid divergence. We found support for a single reticulation event connecting *G. seminuda* and *G. elegans*, an hypothesis consistent with prior interpretations (DeMarais et al. 1992). This particular model (i.e., *h*=1) was selected as the one that maximized both the first [L'(*h*) = L(*h*) – L(*h*-1)] and second order [L''(*h*) = L'(*h*+1) – L'(*h*)] rate of change in pseudolikelihood (Fig. S9; following Evanno et al. 2005). Of note, introgression between *G. elegans* and *G. seminuda* was supported by elevated values of *h,* and by *D*-statistics ($\overline{D}$ = 0.302 across 86,400 tests; Table 1). Introgression between upper basin *G. robusta* and *G. cypha* was also supported ($\overline{D}$ = -0.236 across 45,056 tests). No other introgressions were noted, thus rejecting the hypothesized hybrid origins for both *G. jordani* (Dowling and DeMarais 1993; Dowling and Secor 1997) and *G. nigra* (Demarais 1986; Minckley and DeMarais 2000).

Multiple internode pairs were observed in the anomaly zone (Fig. 5). In all cases, internode branches separating *G. nigra* and *G. intermedia,* and those separating their constituent lineages, reflected coalescent lengths that would yield anomalous gene trees. Not surprisingly, the internode separating *G. jordani* from the lower basin complex, and that of *G. robusta* from *G. intermedia*/ *G. nigra* (Fig. 5C; tan branches) also fell within the anomaly zone, per TICR and concatenated topology results.

Relative Performance of Species-Tree Methods

Change in site-likelihoods among constrained and unconstrained IQ-TREE searches in all cases suggested that our recovered species-trees were supported by a minority of sites (Fig. S10), an observation consistent with tree regions being in the anomaly zone. Several discrepancies also reflected idiosyncrasies among the different approaches. For example, the POMO topology has a paraphyletic upper basin *G. robusta* within which *G. elegans*, *G. cypha*, *G. seminuda*, *G. jordani*, and the lower basin complex were subsumed (Fig. 4A). However, only ~10% of SNPs supported this resolution (Fig. S11), a value far below the theoretical minimum s*CF* derived from completely random data (Minh et al. 2018). Of note, hybridization is a well-known artefact when a bifurcating tree is inferred from reticulated species (Sosef 1997; Schmidt-Lebuhn 2012), with concatenation or binning approaches using genomic data being demonstrably vulnerable (Bangs et al. 2018). Thus, we tentatively attribute the observed paraphyly as an artefact of documented hybridization between *G. cypha* and *G. robusta* (Chafin et al. 2019), and the inability of POMO to model hybridization. Hybridization also potentially drives the lack of monophyly in *G. seminuda,* per TICR and the concatenation tree (Fig. S8).

We also explored the impact of matrix occupancy filters on SVDQUARTETS, and bootstrap support and overall topological consistency declining with increasingly stringent filters (Fig. 3b). This corroborates prior evaluations with regard to the impacts of over-filtering RADseq data (Eaton et al. 2017). In all cases, site-wide concordance was significantly predicted by subtending branch lengths, but not by node depths (Fig. S12). This suggests that site-wise concordance was unbiased in our analyses at either shallower or deeper timescales but was affected instead by the extent of time separating divergences. Some bioinformatic biases such as ortholog misidentification or lineage-specific locus dropout will disproportionally affect deeper nodes (Eaton 2017). However, we interpret the lack of correlation between node depth and site-wise concordance as an indication that these processes lack substantial bias.

Biogeographic Hypotheses and Co-divergence

ECOEVOLITY model selection was not found to be vulnerable to alternative event priors (Fig. S13). The best-fitting model across all priors consistently demonstrated co-divergence of *G. jordani* with the lower basin complex (*G. robusta* x *G. intermedia* and *G. intermedia* x *G. nigra*; Fig. 6). The divergence of *G. elegans* and *G. seminuda* from a theoretical lower basin ancestor pre-dates this putatively rapid radiation, although it is unclear if these estimates were impacted by the aforementioned introgression between *G. seminuda* and *G. elegans*.

Posterior effective population size (N_e) estimates were large (e.g. >20,000) and consistent with previous estimates (Garrigan et al. 2002). *Gila jordani* was an exception, with a mean posterior N_e=6,062. This discrepancy is not surprising, given the extremely narrow endemism of this species (Tuttle and Scoppettone 1990), and its recent bottleneck (Hardy 1982), although this is still a rather large estimate given the latter. Posterior divergence time estimates suggested a late-Miocene/ early-Pliocene origin of *G. elegans*. Results for *G. seminuda* and the lower basin radiation indicated Pliocene and early Pleistocene divergences, respectively. These results are supported in the fossil record (Uyeno 1960; Uyeno and Miller 1963), although we note paleontological evaluations of *Gila* have been sparse. Thus, we hesitate to interpret these as absolute dates, given our fixed mutation rate for these analyses and an uncertainty regarding the capacity of RADseq methods to yield an unbiased sampling of genome-wide mutation rate variation (e.g. Cariou et al. 2016).

Discussion

The goal of our study was to determine if extensive geographic and genomic sampling could resolve the taxonomically recalcitrant *G. robusta* complex. We applied diverse phylogenetic models and tests of hybridization and predictions of parameter space within the anomaly zone to diagnose sources of subtree discordance. In so doing, we also tested multiple hypothesized hybrid speciation events. We detected a single reticulation (*G. seminuda*), although other events with a lower component of genomic introgression may have also occurred. We documented rapid co-divergence of lower basin taxa within the anomaly zone and were able to resolve these despite the prevalence of incomplete lineage sorting. This scenario (as outlined below) is consistent with the geomorphology of the region and seemingly represents an adaptive radiation by our study complex, as facilitated by drainage evolution.

Methodological Artefacts and Conflicting Phylogenetic Hypotheses for Gila

Increased geographic and genomic sampling revealed the presence of diagnosable lineages within the *G. robusta* complex, with both rapid and reticulate divergences influencing inter-locus conflict. Phylogenetic hypotheses for our focal group had previously been generated using allozymes (Dowling and DeMarais 1993), Sanger sequencing (Schwemm 2006; Schönhuth et al. 2014), microsatellites (Dowling et al. 2015), and more recently RADseq (Copus et al. 2018). None could resolve relationships within the lower basin complex. To explain these contrasts, we argue that prior studies suffered from systematic artefacts and ascertainment biases that were overcome, at least in part, by our approach.

Incomplete or biased sampling is a familiar problem for biologists (e.g. Hillis 1998; Schwartz and McKelvey 2009; Ahrens et al. 2016), and we suggest it served as a major stumbling block for delineating the evolutionary history of *Gila*. Although insufficient sampling is common in studies of threatened and endangered species, its repercussions are severe with regard to phylogenetic inference (Hillis 1998). This fact is substantiated by the many examples in which increasingly comprehensive geographic sampling spurred a revision of phylogenetic hypotheses (e.g. Oakey et al. 2004; Linck et al. 2019). Likewise, incomplete sampling of genome-wide topological variation (e.g. Maddison 1997; Degnan and Rosenberg 2009) is an additional source of bias, especially when a very small number of markers are sampled. These issues alone may explain the variation among prior studies. For example, Schwemm (2006) sampled extensively, including nearly all of the sites included in this study, but was only able to examine a handful of genes. Because anomalous gene trees are most probable under a scenario of rapid radiation (as documented herein), the reduced number of loci used by Schwemm (2006) could not recover a consistent species tree. Copus et al. (2016, 2018) examined a dataset containing 6,658 genomic SNP loci (across 1,292 RAD contigs), but only did so across a sparse sample of 19 individuals. A bioinformatic acquisition bias also likely impacted this study, in the form of strict filtering that disproportionately excluded loci with higher mutation rates (Huang and Knowles 2016).

A necessary consideration when validating phylogenetic hypotheses across methods (and datasets) is to gauge compatibility between the underlying evolutionary processes and those actually being modeled. In this sense, the consideration of statistical support metrics alone can be not only misleading, but also promote false conclusions. For example, bootstrapping is by far the most prevalent method of evaluating support in phylogenetic datasets (Felsenstein 1985). While bootstrap concordances may be appropriate for moderately-sized sequence alignments (e.g. Efron et al. 1996), they can be meaningless when applied to sufficiently large datasets (Gadagkar et al. 2005; Kumar et al. 2012). This is apparent in the high bootstrap support displayed for anomalous relationships in our own analysis (Fig. S8). Phylogenetic signal also varies among loci, such that in many instances, relatively few loci drive contentious relationships (Shen et al. 2017). Likewise, not all methods are equal with respect to their simplifying assumptions. Given this, we deem it imperative to consider the biases and imperfections in both our data, and the models we apply.

Complex Evolution and Biogeography of the Colorado River

The taxonomic instability in *Gila* is not uncommon for fishes of western North America, where confusing patterns of diversity were generated by tectonism and vulcanism (Minckley et al. 1986; Spencer et al. 2008). This issue is particularly emphasized when viewed through the lens of modern drainage connections (Douglas et al. 1999). Historic patterns of drainage isolation and intermittent fluvial connectivity not only support our genomic conclusions but also summarize the paleohistory of the Colorado River over temporal and spatial scales.

The earliest record of fossil *Gila* from the ancestral Colorado River is mid-Miocene (Uyeno and Miller 1963), with subsequent Pliocene fossils representing typical 'big river' morphologies now associated with *G. elegans*, *G. cypha*, and *G. robusta* (Uyeno and Miller 1965). The modern Grand Canyon region lacked any fluvial connection at the Miocene-Pliocene transition, due largely to regional tectonic uplifts that subsequently diverted the Colorado River (Spencer et al. 2001; House et al. 2005). Flows initiated in early Pliocene (c.a. 4.9 mya; Sarna-Wojcicki et al. 2011), and subsequently formed a chain of downstream lakes associated with the Bouse Formation (Lucchitta 1972; Spencer and Patchett 2002). Evidence suggests 'spillover' by a successive string of Bouse Basin paleolakes was episodic, and culminated in mid-Pliocene (House et al. 2008), with an eventual marine connection via the Salton Trough to the Gulf of California (Dorsey et al. 2007). Prior to this, the Gila River also drained into the Gulf (Eberly and Stanley 1978), and sedimentary evidence indicated that it was isolated from the Colorado until at least mid-Pliocene by a northward extension of the Gulf (Helenes and Carreno 2014). This geomorphology is reflected in a broader phylogeographic pattern that underscores marked differences between resident fish communities in the upper and lower basins (Hubbs and Miller 1948).

Intra-basin diversification also occurred as an addendum to hydrologic evolution. Although the course of the pluvial White River is now generally dry, it may have been a Pliocene-early Pleistocene tributary of a paleolake system when the proto-Colorado first extended into the modern-day lower basin (Dickinson 2013). This may represent an initial colonization opportunity for upper basin fishes, an hypothesis that coincidentally aligns well with our rudimentary age estimate for Virgin River chub, *G. seminuda* (Fig. 6). This early isolation, as well as the continued contrast between the spring-fed habitats therein, and the high flows of the ancestral Colorado River, provide an explanation for the unique assemblage of *Gila* and other fishes therein (Hubbs and Miller 1948).

Phylogenetic signatures of the anomaly zone (Fig. 5) coupled with co-divergence modeling (Fig. 6) suggest the diversification of lower basin *Gila* occurred rapidly post-colonization. Late Pliocene integration of the two basins provided an opportunity for dispersal into the lower basin tributaries. The Plio-Pleistocene climate of the region was quite different, with a relatively mesic Pliocene as precursor to a protracted monsoonal period extending through early Pleistocene (Thompson 1991; Smith et al. 1993). The latter, in turn, may have resulted in relatively unstable drainage connections (Huckleberry 1996). The potential for climate-driven instability, and the complex history of intra-drainage integration of Gila tributaries during the Plio-Pleistocene (Dickinson 2015), lends support to the 'cyclical-vicariance' model proposed by Douglas et al. (1999). These periods of isolation may have promoted an accumulation of ecological divergences that persisted post-contact, and were sufficient to maintain species boundaries despite contemporary sympatric distributions and weak morphological differentiation. This hypothesis is also supported by the non-random mating found among *G. robusta* and *G. nigra,* despite anthropogenically-induced contact (Marsh et al. 2017).

Management Implications

A request by the Arizona Game and Fish Department to review the taxonomy of the *Gila robusta* complex prompted the American Fisheries Society (AFS) and the American Society of Ichthyology and Herpetology (ASIH) to recommend the synonymization of *G. intermedia* and *G. nigra* with *G. robusta*, owing in part to their morphological ambiguity and an imprecise taxonomic key (Carter et al. 2018). Given this, a proposal to extend protection to lower basin *G. robusta* and *G. nigra* at the federal level was subsequently withdrawn (USFWS 2017; Fig. 1). As was the case prior to this withdrawal, *G. intermedia* alone is classified as endangered (USFWS 2005) under the Endangered Species Act (ESA 1973; 16 U.S.C. § 1531 et seq).

This study provides a much needed resolution to this debate by defining several aspects: First, our study reinforced the recognition of *G. robusta* as demonstrably polyphyletic, with two discrete, allopatric clades corresponding to the upper and lower basins of the Colorado River (Dowling and DeMarais 1993; Schönhuth et al. 2014). These data, together with the geomorphic history of the region that promoted endemic fish diversification (as above), clearly reject '*G. robusta*' as a descriptor of contemporary diversity. This underscores a major discrepancy in the taxonomic recommendations for the lower basin complex (Page et al. 2016). Given that the type locality of *G. robusta* is in the upper basin (i.e., the Little Colorado River), we note a pressing need either to determine taxonomic precedence for the lower basin '*G. robusta,*' or to provide a novel designation. The potential resurrection of a synonym is a possibility, necessitating a detailed examinations of the type specimens prior to a formal recommendation. This may be appropriately adjudicated by the AFS-ASIH Names of Fishes Committee, as a follow-up to their earlier involvement.

The situation with *G. intermedia* and *G. nigra* is slightly more ambiguous. The short internodes and anomaly zone divergences identified herein explain previous patterns found in population-level studies, with elevated among-population divergence but scant signal uniting species (Dowling et al. 2015). We also unequivocally rejected the previous hypothesis of hybrid speciation for *G. nigra* (Minckley and DeMarais 2000; Dowling et al. 2015).

Rather, intermediacy in the body shape of *G. nigra* reflects differences accumulated during historic isolation (Douglas et al. 1999) and/ or the retention of an adaptive ecomorphology (Douglas and Matthews 1992). These hypotheses warrant further exploration, with provisional results employed in future management decisions (Forest et al. 2015). With regards to taxonomy, we confidently recommend that *G. intermedia* be resurrected, and that additional studies be implemented to dissect the potential distinctiveness of *G. nigra*. For management purposes, we echo a conservative, population-centric approach (previously argued for by Dowling et al. 2015; Marsh et al. 2017).

Three primary components of a 'Darwinian shortfall' in biodiversity conservation are recognized (Diniz-Filho et al. 2013): (i) The lack of comprehensive phylogenies; (ii) Uncertain branch lengths and divergence times; and (iii) insufficient models linking phylogenies with ecological and life-history traits. Taxonomic uncertainty in *Gila* is severely impacted by the first two of these, with taxonomic resolution prevented by the comingling of sparse phylogenetic coverage with temporal uncertainty. We must now address the relationships between ecology, life history, and phylogeny in *Gila,* so as to understand the manner by which phylogenetic groupings (identified herein) are appropriate as a surrogate for adaptive/ functional diversity. For example: To what degree are *Gila* in the lower basin ecologically non-exchangeable? How do they vary in their respective life histories? Is reproductive segregation maintained in sympatry (as in Marsh et al. 2018), and if so, by what mechanism?

Conclusion

The intractable phylogenetic relationships in *Gila* were resolved herein through improved spatial and genomic sampling. Our data, coupled with polymorphism-aware methods and contemporary approaches that infer trees, yielded a revised taxonomic hypothesis for *Gila* in the lower Colorado basin. The geomorphic history of the Colorado River explains many anomalous patterns seen in this and previous studies, wherein opportunities for contact and colonization were driven by the tectonism characteristic for the region. The signal of rapid diversification is quite clear in our data, as interpreted from patterns inherent to phylogenetic discord. We emphasize that discordance in this sense does not necessarily represent measurement error or uncertainty, but rather an intrinsic component of phylogenetic variance that is not only expected within genomes (Maddison 1997), but also a necessary component from which to build hypotheses regarding the underlying evolutionary process (Hahn and Nakhleh 2016). Ignoring this variance in pursuit of a 'resolved phylogeny' can lead to incorrect inferences driven by systematic error. Similarly, insufficient spatial or genomic sampling may also promote a false confidence in anomalous relationships, particularly when character sampling is particularly dense, whereas taxon sampling is sparse.

We reiterate that phylogenetic hypotheses, by their very nature, cannot exhaustively capture the underlying evolutionary process. One approach is to categorize phylogenetic (and "non-phylogenetic") signals in those regions of the tree that are refractive to certain models (as done herein). We also acknowledge that attempting to reconstruct the past using contemporary observations is a battle against uncertainty and bias, with the revisions of phylogenetic/ taxonomic hypotheses expected as additional data are accrued. As such, we urge empiricists that engage in taxonomic controversies (such as this one) to interrogate their results for transparency. The task of sorting through conflicting recommendations invariably falls to natural resource managers, with unreported biases (be they methodological or geopolitical) only confounding those efforts.

References

Ahrens D., Fujisawa T., Krammer H.J., Eberle J., Fabrizi S., Vogler A.P. 2016. Rarity and incomplete sampling in DNA-based species delimitation. Syst. Biol. 65:478–494.

Bangs M.R., Douglas M.R., Mussmann S.M., Douglas M.E. 2018. Unraveling historical introgression and resolving phylogenetic discord within *Catostomus* (Osteichthys: Catostomidae). BMC Evol. Biol. 18:86.

Beaulieu J.M., O'Meara B.C. 2018. Can we build it? Yes we can, but should we use it? Assessing the quality and value of a very large phylogeny of campanulid angiosperms. Am. J. Bot. 105:417–432.

Brandenburg W.H., Kennedy J.L., Farrington M.A. 2015. Determining the historical distribution of the Gila robusta complex (Gila chub, *Gila intermedia*, Headwater Chub, *Gila nigra*, and Roundtail Chub, *Gila robusta*) in the Gila River Basin, New Mexico, using morphological analysis. Final Rep. to New Mex. Dep. Game Fish.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Bureau of Reclamation. 2012. Colorado River Basin water supply and demand study. Executive Summary.

Cariou M., Duret L., Charlat S. 2016. How and how much does RAD-seq bias genetic diversity estimates? BMC Evol. Biol. 16:1–8.

Carlson C.A., Muth R.T. 1989. The Colorado River: Lifeline of the American Southwest. Can. Spec. Publ. Fish. Aquat. Sci. 106:220–239.

Carter J.M., Clement M.J., Makinster A.S., Crowder C.D., Hickerson B.T. 2018. Classification success of species within the *Gila robusta* complex using morphometric and meristic characters—A re-examination. Copeia. 106:279–291.

Chafin T.K., Douglas M.R., Martin B.T., Douglas M.E. 2019. Hybridization drives genetic erosion in sympatric desert fishes of western North America. Heredity. 123:759-773.

Chafin T.K., Martin B.T., Mussmann S.M., Douglas M.R., Douglas M.E. 2018. FRAGMATIC: *in silico* locus prediction and its utility in optimizing ddRADseq projects. Conserv. Genet. Resour. 10:325–328.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics. 30:3317–3324.

Chifman J., Kubatko L. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes. J. Theor. Biol. 374:35–47.

Copus J.M., Montgomery W.L., Forsman Z.H., Bowen B.W., Toonen R.J. 2018. Geopolitical species revisited: genomic and morphological data indicate that the roundtail chub *Gila robusta* species complex (Teleostei, Cyprinidae) is a single species. PeerJ. 6:e5605.

Darwin C. 1859. On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life. London: John Murray.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Degnan J.H., Salter L.A. 2005. Gene tree distributions under the coalescent process. Evolution. 59:24–37.

De Maio N., Schrempf D., Kosiol C. 2015. PoMo: An allele frequency-based approach for species tree estimation. Syst. Biol. 64:1018–1031.

DeMarais B.D. 1986. Morphological variation in *Gila* (Pisces: Cyprinidae) and geological history: lower Colorado River basin [thesis]. Tempe: Arizona State University. 85p.

DeMarais B.D., Dowling T.E., Douglas M.E., Minckley W.L., Marsh P.C. 1992. Origin of *Gila seminuda* (Teleostei: Cyprinidae) through introgressive hybridization: implications for evolution and conservation. Proc. Natl. Acad. Sci. U. S. A. 89:2747–2751.

Dickinson W.R. 2013. Rejection of the lake spillover model for initial incision of the Grand Canyon, and discussion of alternatives. Geosphere. 9:1–20.

Dickinson W.R. 2015. Integration of the Gila River drainage system through the Basin and Range province of southern Arizona and southwestern New Mexico (USA). Geomorphology. 236:1–24.

Diniz-Filho J.A.F., Loyola R.D., Raia P., Mooers A.O., Bini L.M. 2013. Darwinian shortfalls in biodiversity conservation. Trends Ecol. Evol. 28:689–695.

Dorsey R.J., Fluette A., McDougall K., Housen B.A., Janecke S.U., Axen G.J., Shirvell C.R. 2007. Chronology of Miocene-Pliocene deposits at Split Mountain Gorge, Southern California: A record of regional tectonics and Colorado River evolution. Geology. 35:57–60.

Douglas M.E., Douglas M.R., Lynch L.M., McElroy D.M. 2001. Use of geometric morphometrics to differentiate *Gila* (Cyprinidae) within the upper Colorado River basin. Copeia. 2001(2): 389–400.

Douglas M.E., Minckley W.L., DeMarais B.D. 1999. Did vicariance mold phenotypes of western North American fishes? Evidence from Gila River cyprinids. Evolution. 53:238–246.

Douglas M.E., Matthews W.J. 1992. Does morphology predict ecology? Hypothesis testing within a freshwater stream fish assemblage. Oikos. 65:213.

Douglas M.R., Douglas M.E. 2007. Genetic structure of humpback chub *Gila cypha* and roundtail chub *G. robusta* in the Colorado River ecosystem. Rep. to Gd. Canyon Monit. Reserach Center, U.S. Geol. Surv. 99pp.

Dowling T.E., Anderson C.D., Marsh P.C., Rosenberg M.S. 2015. Population structure in the Roundtail chub (*Gila robusta* complex) of the Gila River Basin as determined by microsatellites: Evolutionary and conservation implications. PLoS One. 10:e0139832.

Dowling T.E., DeMarais B.D. 1993. Evolutionary significance of introgressive hybridization in cyprinid fishes. Nature. 362:444–446.

Dowling T.E., Secor C.L. 1997. The role of hybridization and introgression in the diversification of animals. Annu. Rev. Ecol. Syst. 28:593–619.

Eaton D.A.R. 2014. PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. Bioinformatics. 30:1844–1849.

Eaton D.A.R., Ree R.H. 2013. Inferring phylogeny and introgression using RADseq data: an example from glowering plants (Pedicularis: Orobanchaceae). Syst. Biol. 62:689–706.

Eaton D.A.R., Spriggs E.L., Park B., Donoghue M.J. 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. Syst. Biol. 66:399–412.

Eberly L.D., Stanley T.B. 1978. Cenozoic stratigraphy and geologic history of southwestern Arizona. Geol. Soc. Am. Bull. 89:921–940.

Edwards S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

Edwards S. V., Xi Z., Janke A., Faircloth B.C., et al. 2016. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. 94:447–462.

Efron B., Halloran E., Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. 93:13429–13429.

Evanno G., Regnaut S., Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol. Ecol. 14:2611–2620.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution. 39:783–791.

Forest F., Crandal K.A., Chase M.W., Faith D.P. 2015. Phylogeny, extinction and conservation: Embracing uncertainties in a time of urgency. Philos. Trans. R. Soc. B Biol. Sci. 370:1–8.

Gadagkar S.R., Rosenberg M.S., Kumar S. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. J. Exp. Zool. Part B Mol. Dev. Evol. 304:64–74.

Garrigan D., Marsh P.C., Dowling T.E. 2002. Long-term effective population size of three endangered Colorado River fishes. Anim. Conserv. 5:95–102.

Gerber A.S., Tibbets C.A., Dowling T.E. 2001. The role of introgressive hybridization in the evolution of the *Gila robusta* complex (Teleostei: Cyprinidae). Evolution. 55:2028–2039.

Giarla T.C., Esselstyn J.A. 2015. The challenges of resolving a rapid, recent radiation: Empirical and simulated phylogenomics of Philippine shrews. Syst. Biol. 64:727–740.

Green R.E., Krause J., Briggs A.W., et al. 2010. A draft sequence of the Neandertal genome. Science. 328:710–722.

Hahn M.W., Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution. 70:7–17.

Hardy T.B. 1982. Ecological interactions of the introduced and native fishes in the outflow of Ash Springs, Lincoln County, Nevada [thesis]. Las Vegas: University of Nevada. 99p.

Heath T.A., Hedtke S.M., Hillis D.M. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46:239–257.

Helenes J., Carreno A.L. 2014. Neogene sedimentary record of the Gulf of California: towards a highly biodiverse scenario. In: Wehncke E. V., Lara-lAra J.R., Alvarez-Borrego S., Ezcurra E., editors. Conservation Science in Mexico's Northwest: Ecosystem Status and Trends in the Gulf of Calfiornia.

Hillis D.M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:3–8.

House P.K., Pearthree P.A., Howard K.A., Bell J.W., Perkins M.E., Faulds J.E., Brock A.L. 2005. Birth of the lower Colorado River—Stratigraphic and geomorphic evidence for its inception near the conjunction of Nevada, Arizona, and California. GSA Field Guide

House P.K., Pearthree P.A., Perkins M.E. 2008. Stratigraphic evidence for the role of lake spillover in the inception of the lower Colorado River in southern Nevada and western Arizona. Spec. Pap. 439 Late Cenozoic Drain. Hist. Southwest. Gt. Basin Low. Color. River Reg. Geol. Biot. Perspect. 2439:335–353.

Huang H., Knowles L.L. 2016. Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. Syst. Biol. 65:357–65.

Hubbs C.L., Miller R.R. 1948. The zoological evidence: correlation between fish distribution and hydrogeographic history in the desert basin of western North America. Bull. Univ. Utah. 10:17–166.

Huckleberry G. 1996. Historical geomorphology of the Gila River. Rep. to Arizona Geol. Surv.

Huckleberry J.K., Potts M.D. 2019. Constraints to implementing the food-energy-water nexus concept: Governance in the Lower Colorado River Basin. Environ. Sci. Policy. 92:289–298.

James T., Evans A., Madly E., Kelly C. 2014. The economic importance of the Colorado River to the Basin region. L William Seidman Res. Institute, W. P. Carey Sch. Business, Arizona State Univ.

Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L., Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–472.

Larget B.R., Kotha S.K., Dewey C.N., Ané C. 2010. BUCKy: Gene tree reconciliation with concordance. Bioinformatics. 26:2910–2911.

Leaché A.D., Fujita M.K., Minin V.N., Bouckaert R.R. 2014. Species delimitation using genome-wide SNP data. Syst. Biol. 63:534–542.

Leaché A.D., Oaks J.R. 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 48:69–84.

Lemmon E.M., Lemmon A.R. 2013. High-throughput genomic data in systematics and phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44:99–121.

Linck E., Epperly K., van Els P., Spellman G.M., Bryson Jr R.W., McCormack J.E., Canales-del-Castillo R., Klicka J. 2019. Dense geographic and genomic sampling reveals paraphyly and a cryptic lineage in a classic sibling species complex. Syst. Biol. Early Acce:syz027.

Linkem C.W.C., Minin V.N.V., Leache A., Leaché A.D. 2016. Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (Squamata: Scincidae). Syst. Biol. 65:465–477.

Liu L., Edwards S. V. 2009. Phylogenetic analysis in the anomaly zone. Syst. Biol. 58:452–460.

Lucchitta I. 1972. Early history of the Colorado River in the Basin and Range province. Geol. Soceity Am. Bulletin. 83:1933–1948.

Maddison W. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Marsh P.C., Clarkson R.W., Dowling T.E. 2017. Molecular genetics informs spatial segregation of two desert stream *Gila* species. Trans. Am. Fish. Soc. 146:47–59.

McCormack J.E., Huang H., Knowles L.L. 2009. Maximum likelihood estimates of species trees: How accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst. Biol. 58:501–508.

McKee E.D., Wislon R.F., Breed W.J., Breed C.S. 1967. Evolution of the Colorado River in Arizona- An hypothesis developed at the Symposium on Cenozoic Geology of the Colorado Plateau in Arizona, August 1964. Museum of Northern Arizona Bulletin.

McVay J.D., Hipp A.L., Manos P.S. 2017. A genetic legacy of introgression confounds phylogeny and biogeography in oaks. Proc. R. Soc. B Biol. Sci. 284.

Meffe G.K., Vrijenhoek R.C. 1988. Conservation genetics in the management of desert fishes. Conserv. Biol. 2:157–169.

Mendes F.K., Hahn Y., Hahn M.W. 2016. Gene tree discordance can generate patterns of diminishing convergence over time. Mol. Biol. Evol. 33:3299–3307.

Minckley W.L. 1979. Aquatic habitats and fishes of the lower Colorado River, southwestern United States. Rep. to U.S. Bur. Reclam.

Minckley W.L., Henderson D.A., Bond C.E. 1986. Geography of western North American freshwater fishes: description and relationships to intracontinental tectonism. Zoogeography North Am. Freshw. Fishes.:519–613.

Minckley W.L., DeMarais B.D. 2000. Taxonomy of chubs (Telostei, Cyprinidae, genus *Gila*) in the American Southwest with comments on conservation. Copeia. 2000:251–256.

Minckley W.L., Marsh P.C. 2009. Inland fishes of the greater Southwest: chronicle of a vanishing biota. Tuscon (AZ): University of Arizona Press.

Minckley W.L., Marsh P.C., Deacon J.E., Dowling T.E., Hedrick P.W., Matthews W.J., Mueller G. 2006. A conservation plan for native fishes of the lower Colorado River. Bioscience. 53:219.

Minh B.Q., Hahn M., Lanfear R. 2018. New methods to calculate concordance factors for phylogenomic datasets. bioRxiv. 10.1101/487801

Moran C.J., O'Neill M.W., Armbruster J.W., Gibb A.C. 2017. Can members of the south-western *Gila robusta* species complex be distinguished by morphological features? J. Fish Biol. 91:302–316.

Mussmann S.M., Douglas M.R., Bangs M.R., Douglas M.E. 2019. Comp-D: a program for comprehensive computation of D-statistics and population summaries of reticulated evolution. Conserv. Genet. Resour. 0:0.

Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2014. IQ-TREE: A fast and effective stochastic algorithm for estimating Maximum-Likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Oakey D.D., Douglas M.E., Douglas M.R. 2004. Small fish in a large landscape: Diversification of *Rhinichthys osculus* (Cyprinidae) in Western North America. Copeia. 2004:207–221.

Oaks J.R. 2018. Full Bayesian comparative phylogeography from genomic data. Syst. Biol. 0:1–25.

Page L.M., Baldwin C.C., Espinosa-Pérez H., Findley L.T., Gilbert C.R., Hartel K.E., Lea R.N., Mandrak N.E., Schmitter-Soto J.J., Walker H.J. 2017. Taxonomy of *Gila* in the Lower Colorado River Basin of Arizona and New Mexico. Fisheries. 42:456–460.

Page L.M., Baldwin C.C., Espinosa-Pérez H., Gilbert C.R., Hartel K.E., Lea R.N., Mandrak N.E., Schmitter-Soto J.J., Walker H.J. 2016. Final report of the AFS/ASIH Joint Committee on the Names of Fishes on the taxonomy of *Gila* in the Lower Colorado River basin of Arizona and New Mexico. Rep. to Arizona Game Fish Dep.

Pante E., Puillandre N., Viricel A., Arnaud-Haond S., Aurelle D., Castelin M., Chenuil A., Destombe C., Forcioli D., Valero M., Viard F., Samadi S. 2015. Species are hypotheses: Avoid connectivity assessments based on pillars of sand. Mol. Ecol. 24:525–544.

Pease J.B., Hahn M.W. 2015. Detection and polarization of introgression in a five-taxon phylogeny. Syst. Biol. 64:651–662.

Pereira A.G., Schrago C.G. 2018. Incomplete lineage sorting impacts the inference of macroevolutionary regimes from molecular phylogenies when concatenation is employed: An analysis based on Cetacea. Ecol. Evol. 8:6965–6971.

Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. 2012. Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS One. 7:e37135.

Philippe H., Brinkmann H., Lavrov D. V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol. 9:e1000602.

Philippe H., Delsuc F., Brinkmann H., Lartillot N. 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 36:541–562.

Pisani D., Pett W., Dohrmann M., Feuda R., Rota-Stabelli O., Philippe H., Lartillot N., Wörheide G. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. 112:15402–15407.

Rangel T.F., Colwell R.K., Graves G.R., Fučíková K., Rahbek C., Diniz-Filho J.A.F. 2015. Phylogenetic uncertainty revisited: Implications for ecological analyses. Evolution. 69:1301–1312.

Rannala B., Yang Z. 2008. Phylogenetic inference using whole genomes. Annu. Rev. Genomics Hum. Genet. 9:217–231.

Reddy S., Kimball R.T., Pandey A., et al. 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst. Biol. 66:857–879.

Rinne J.N. 1976. Cyprinid fishes of the genus *Gila* from the lower Colorado River basin. Wasmann J. Biol. 34(1):65-107.

Rognes T., Flouri T., Nichols B., Quince C., Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 4:e2584.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Hohna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Sabo J. L., Sinha T., Bowling L.C., et al. 2010. Reclaiming freshwater sustainability in the Cadillac Desert. Proc. Nat. Acad. Sci. 107(50):21256-21262.

Sarna-Wojcicki A.M., Deino A.L., Fleck R.J., McLaughlin R.J., Wagner D., Wan E., Wahl D., Hillhouse J.W., Perkins M. 2011. Age, composition, and areal distribution of the Pliocene Lawlor Tuff, and three younger Pliocene tuffs, California and Nevada. Geosphere. 7:599–628.

Schmidt-Lebuhn A.N. 2012. Fallacies and false premises-a critical assessment of the arguments for the recognition of paraphyletic taxa in botany. Cladistics. 28:174–187.

Schönhuth S., Perdices A., Lozano-Vilano L., García-de-León F.J., Espinosa H., Mayden R.L. 2014. Phylogenetic relationships of North American western chubs of the genus *Gila* (Cyprinidae, Teleostei), with emphasis on southern species. Mol. Phylogenet. Evol. 70:210–230.

Schönhuth S., Shiozawa D.K., Dowling T.E., Mayden R.L. 2012. Molecular systematics of western North American cyprinids (Cypriniformes: Cyprinidae). Zootaxa. 303:281–303.

Schönhuth S., Vukić J., Šanda R., Yang L., Mayden R.L. 2018. Phylogenetic relationships and classification of the Holarctic family Leuciscidae (Cypriniformes: Cyprinoidei). Mol. Phylogenet. Evol. 127:781–799.

Schrempf D., Minh B.Q., De Maio N., von Haeseler A., Kosiol C. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. J. Theor. Biol. 407:362–370.

Schwartz M.K., McKelvey K.S. 2009. Why sampling scheme matters: The effect of sampling scheme on landscape genetic results. Conserv. Genet. 10:441–452.

Schwemm M.R. 2006. Genetic variation in the *Gila robusta* complex (Teleostei: Cyprinidae) in the lower Colorado River [thesis]. Tempe: Arizona State University. 123p.

Shen X.X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:1–10.

Simion P., Philippe H., Baurain D., et al. 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. Curr. Biol. 27:958–967.

Simpson G.G. 1961. Principles of animal taxonomy. New York (NY): Columbia University Press.

Smith G.A., Yang Wang, Cerling T.E., Geissman J.W. 1993. Comparison of a paleosol-carbonate isotope record to other records of Pliocene-early Pleistocene climate in the western United States. Geology. 21:691–694.

Snir S., Rao S. 2012. Quartet MaxCut: A fast algorithm for amalgamating quartet trees. Mol. Phylogenet. Evol. 61:1–8.

Solís-Lemus C., Ané C. 2016. Inferring phylogenetic networks with Maximum Pseudolikelihood under incomplete lineage sorting. PLoS Genet. 12:1–21.

Solís-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: A package for phylogenetic networks. Mol. Biol. Evol. 34:3292–3298.

Sosef M.S.M. 1997. Hierarchical models, reticulate evolution and the inevitability of paraphyletic supraspecific taxa. Taxon. 46:75–85.

Spencer J.E., Patchett P.J. 2002. Sr isotope evidence for a lacustrine origin for the upper Miocene to Pliocene Bouse Formation, lower Colorado River trough, and implications for timing of Colorado Plateau uplift. Geol. Soc. Am. Bull. 109:767–778.

Spencer J.E., Peters L., McIntosh W.C., Patchett P.J., Young R.A., Spamer E.E. 2001. 40Ar/39Ar geochronology of the Hualapai Limestone and Bouse Formation and implications for the age of the lower Colorado River. Color. River Orig. Evol. Gd. Canyon, Arizona, Gd. Canyon Assoc.:89–91.

Spencer J.E., Smith G.R., Dowling T.E. 2008. Middle to late Cenozoic geology, hydrography, and fish evolution in the American Southwest. Geol. Soc. Am. 80301.

Springer M.S., Gatesy J. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94:1–33.

Stadler T., Degnan J.H., Rosenberg N.A. 2016. Does gene tree discordance explain the mismatch between macroevolutionary models and empirical patterns of tree shape and branching times? Syst. Biol. 65:628–639.

Stenz N.W.M., Larget B., Baum D.A., Ané C. 2015. Exploring tree-like and non-tree-like patterns using genome sequences: An example using the inbreeding plant species *Arabidopsis thaliana (L.) heynh*. Syst. Biol. 64:809–823.

Sullivan J., Joyce P. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36:445–466.

Thompson R.S. 1991. Pliocene environments and climates in the western United States. Quat. Sci. Rev. 10:115–132.

Tuttle P., Scoppettone G. 1990. Status and life history of Pahranagat River fishes. Rep. to Nevada Dep. Wildl.

USFWS. 2005. Endangered and Threatened Wildlife and Plants: Listing Gila chub as endangered with critical habitat. Fed. Regist. 70:66664–66721.

USFWS. 2017. Endangered and Threatened Wildlife and Plants: Threatened species status for the Headwater chub and Roundtail chub distinct population segment. Fed. Regist. 82:16981–16988.

Uyeno T. 1960. Osteology and phylogeny of the American cyprinid fishes allied to the genus *Gila* [dissertation]. Ann Arbor: University of Michigan. 173p.

Uyeno T., Miller R.R. 1963. Summary of late Cenozoic freshwater fish records for North America. Occas. Pap. Museum Zool. Univ. Michigan. 631:1–34.

Uyeno T., Miller R.R. 1965. Middle Pliocene cyprinid fishes from the Bidahochi Formation. 1965(1):28–41.

Whelan N. V., Kocot K.M., Moroz L.L., Halanych K.M. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. 112:5773–5778.

Wu M., Kostyun J.L., Hahn M.W., Moyle L.C. 2018. Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. Mol. Ecol. 27:3301–3316.

Appendix

Table 1: Four-taxon *D*-statistic Tests of Admixture. Tests were performed for quartets sampled from *N*=386 *Gila* individuals. Results are reported across *N* separate quartet samples per four-taxon test, randomly sampled without replacement, with site patterns calculated from 21,717 unlinked SNPs. Significance is reported as the proportion of tests at *p*<0.05 (nSig/*N*) using chi-squared (χ^2), *Z*-test[1], and *Z*-test with Bonferroni correction[2]. Positive and negative values of *D* suggest introgression of the P3 lineage with either P2 or P1, respectively. Results in bold were also supported by the phylogenetic network. See Table S1 for detailed locality information.

P3	P2	P1	Mean *D*	*N*	nSig/*N* (χ^2)	nSig/*N* (Z^1)	nSig/ *N* (Z^2)
cypha	*jordani*	lower basin	0.175	86400	0.033	0.072	0.001
cypha	*seminuda*	lower basin	0.099	86400	0.102	0.130	0.002
elegans	*jordani*	lower basin	-0.063	84800	0.029	0.050	0.000
elegans	*robusta* (lower)	*nigra/int.*	-0.026	413600	0.014	0.047	0.001
elegans	*robusta* (upper)	*cypha*	-0.236	45056	0.380	0.415	0.045
elegans	***seminuda***	lower basin	**0.302**	**86400**	**0.654**	**0.674**	**0.251**
jordani	*robusta* (lower)	*nigra/int.*	0.087	601600	0.042	0.072	0.001
jordani	*robusta* (lower)	*nigra/int.*	0.091	212800	0.041	0.067	0.005
nigra	*int.* (Salt)	*int.* (Verde)	0.086	126976	0.057	0.082	0.001
robusta (lower)	*intermedia*	*nigra*	0.041	793600	0.001	0.002	0.000
robusta (upper)	*jordani*	*robusta* (lower)	0.165	168000	0.050	0.081	0.001
robusta (upper)	*robusta* (lower)	*nigra/int.*	-0.009	601600	0.011	0.031	0.000
robusta (upper)	*seminuda*	lower basin	-0.017	180800	0.030	0.053	0.004
seminuda	*jordani*	lower basin	-0.204	81920	0.107	0.152	0.000
seminuda	*robusta* (lower)	*nigra/int.*	0.054	212800	0.011	0.031	0.001
atraria	*robusta* (upper)	*cypha*	0.082	57344	0.064	0.095	0.033
nigrescens	*robusta* (lower)	*nigra/int.*	-0.075	485472	0.023	0.079	0.002
nigrescens	*robusta* (upper)	*cypha*	-0.039	53248	0.040	0.066	0.005
pandora	*robusta* (lower)	*nigra/int.*	-0.123	225600	0.012	0.105	0.010
pandora	*robusta* (upper)	*cypha*	-0.047	24576	0.031	0.057	0.003

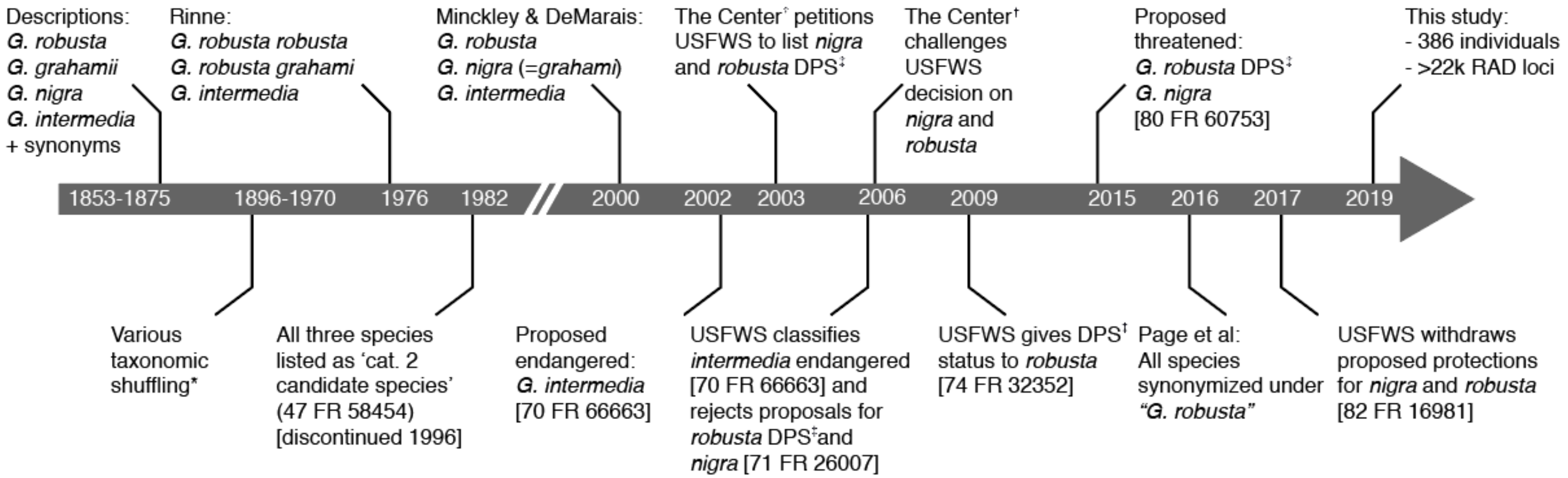
Descriptions:
G. robusta
G. grahamii
G. nigra
G. intermedia
+ synonyms
Rinne:
G. robusta robusta
G. robusta grahami
G. intermedia
Minckley & DeMarais:
G. robusta
G. nigra (=grahami)
G. intermedia
The Center† petitions USFWS to list nigra and robusta DPS‡
The Center† challenges USFWS decision on nigra and robusta
Proposed threatened:
G. robusta DPS‡
G. nigra
[80 FR 60753]
This study:
- 386 individuals
- >22k RAD loci
1853-1875
1896-1970
1976
1982
2000
2002
2003
2006
2009
2015
2016
2017
2019
Various taxonomic shuffling*
All three species listed as 'cat. 2 candidate species' (47 FR 58454) [discontinued 1996]
Proposed endangered:
G. intermedia
[70 FR 66663]
USFWS classifies intermedia endangered [70 FR 66663] and rejects proposals for robusta DPS‡and nigra [71 FR 26007]
USFWS gives DPS‡ status to robusta [74 FR 32352]
Page et al:
All species synonymized under "G. robusta"
USFWS withdraws proposed protections for nigra and robusta [82 FR 16981]

Figure 1: Timeline of the conservation status of *Gila* species endemic to the lower Colorado River basin [*See Copus et al (2018) for a detailed overview of taxonomic synonymies; †'The Center' refers to the Center for Biological Diversity (501c3), Tuscon, AZ; ‡'DPS' = Distinct Population Segment as referenced in the Endangered Species Act (ESA 1973; 16 U.S.C. § 1531 et seq), here referring specifically to a lower basin sub-unit of *G. robusta*]. Note that timeline is not to scale.

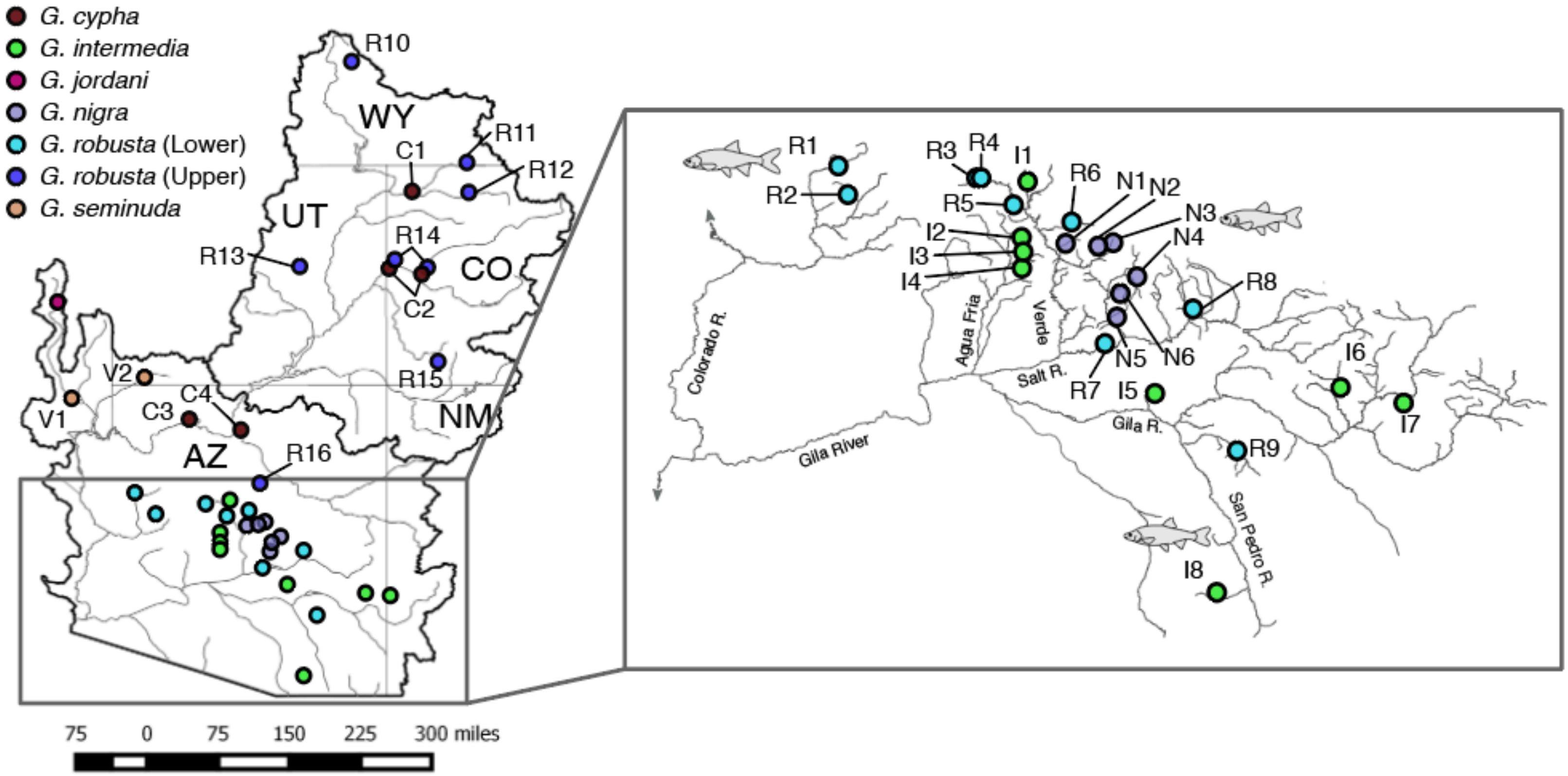

Figure 2: Sampling localities for *Gila* within the Colorado River Basin, southwestern North America. Locality codes are defined in Table S1. Sympatric locations (R14 and C2) are slightly offset for visibility purposes. Map insert increases the viewing scale for sampling sites within the lower basin 'complex' (Bill Williams and Gila rivers).

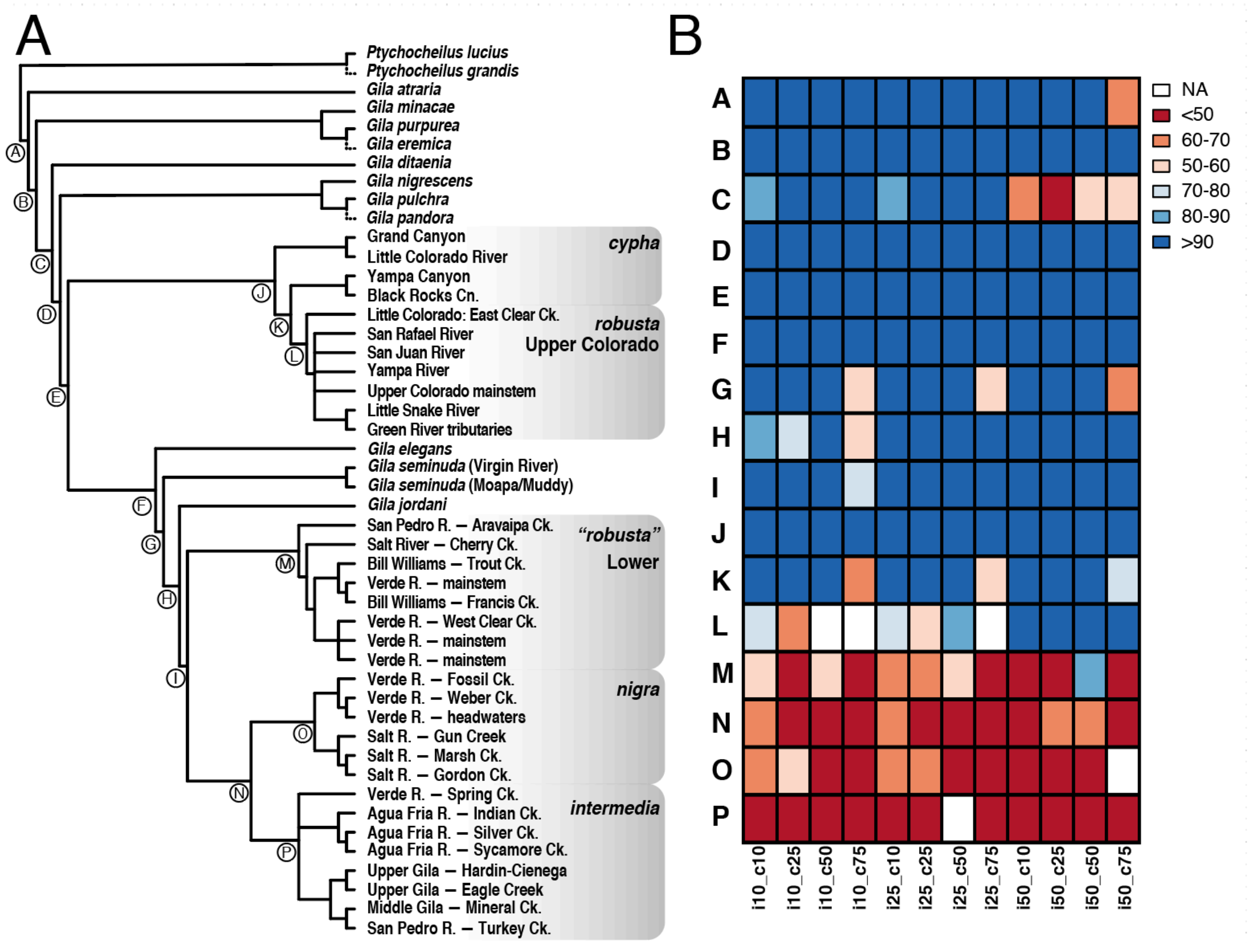

Figure 3: (A) Majority-rule consensus cladogram of SVDQUARTETS across 12 variably filtered SNP datasets varying from 7,357–21,007 SNPs and 256–347 individuals. (B) Binned bootstrap concordance values are reported for each dataset, coded by the matrix occupancy threshold per individual ("i") and per column ("c"; e.g. i50_c50 = 50% occupancy required per individual and per column). Dashed terminal branches indicate positions for taxa missing from >50% of datasets. For detailed locality information, refer to Table S1.

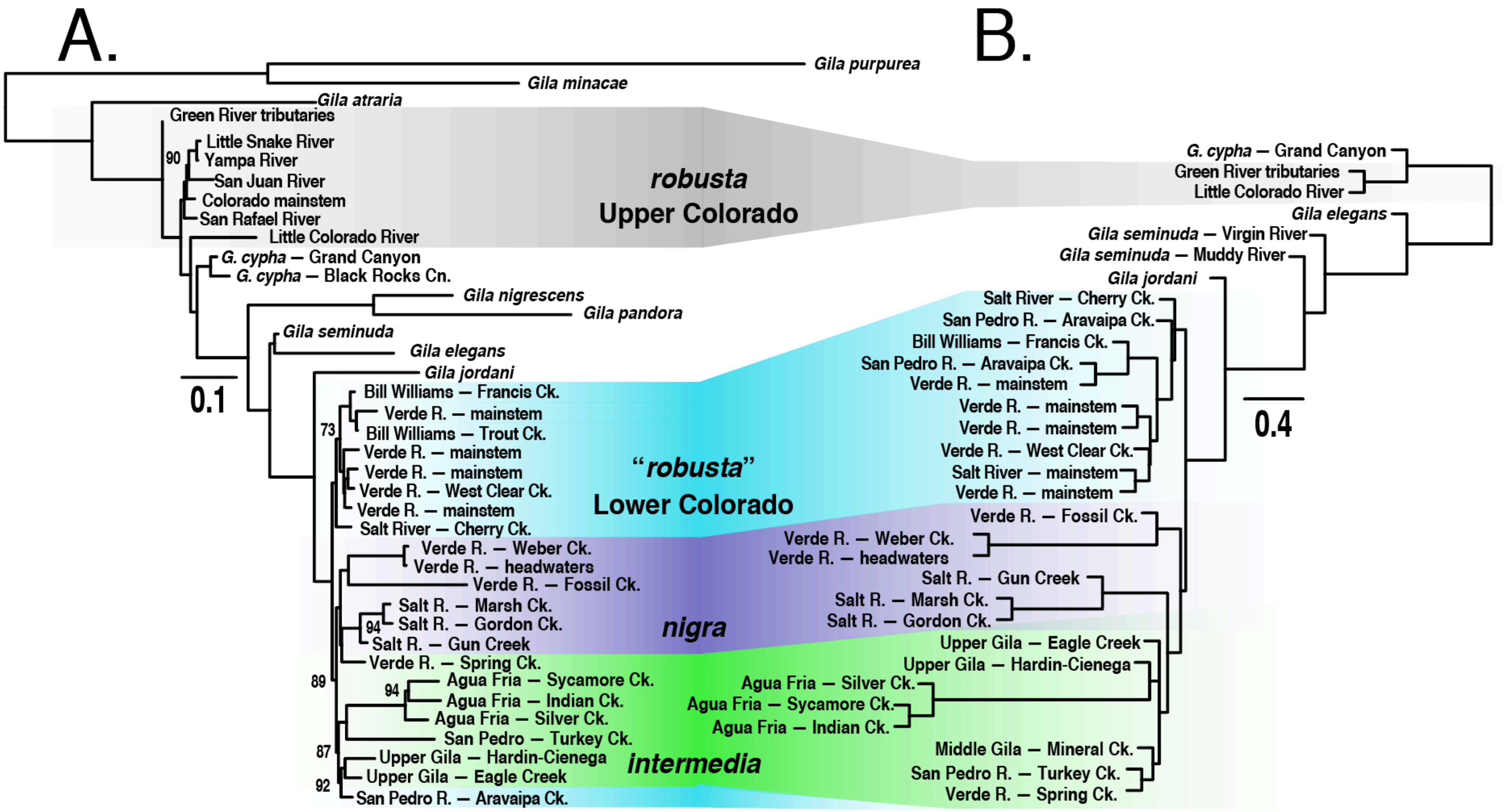

Figure 4: (A) PoMo phylogram with branch lengths as the number of substitutions *and* inferred number of drift events per site, with branch supports (as values <100%) representing concordance among 1,000 bootstrap replicates, inferred using a dataset consisting of 281,613 nucleotides and 40 tips; (B) TICR phylogram reporting branch lengths in coalescent units, calculated from 31,465 quartets evaluated across 3,449 full alignments of ddRAD loci. For detailed locality information, refer to Table S1.

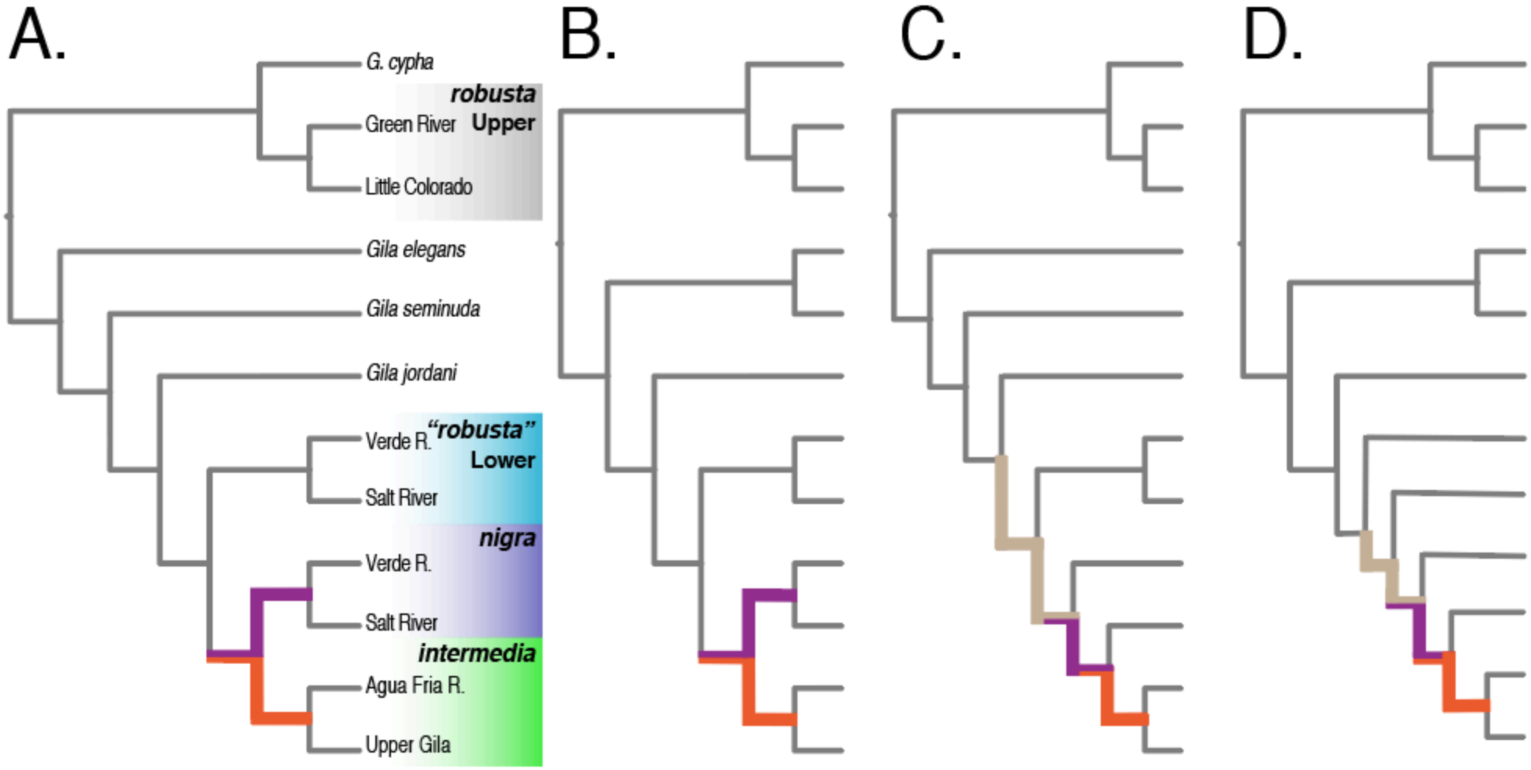

Figure 5: Internode pairs within the anomaly zone, as determined using coalescent-unit transformed branch lengths mapped onto the (A) SVDQUARTETS, (B) POMO, (C) TICR, and (D) concatenated trees (displayed here as cladograms). Paired internodes are color-coded, with those bicolored indicating multiple anomalous divergences. For more detailed representations, refer to Figs. 3 and 4.

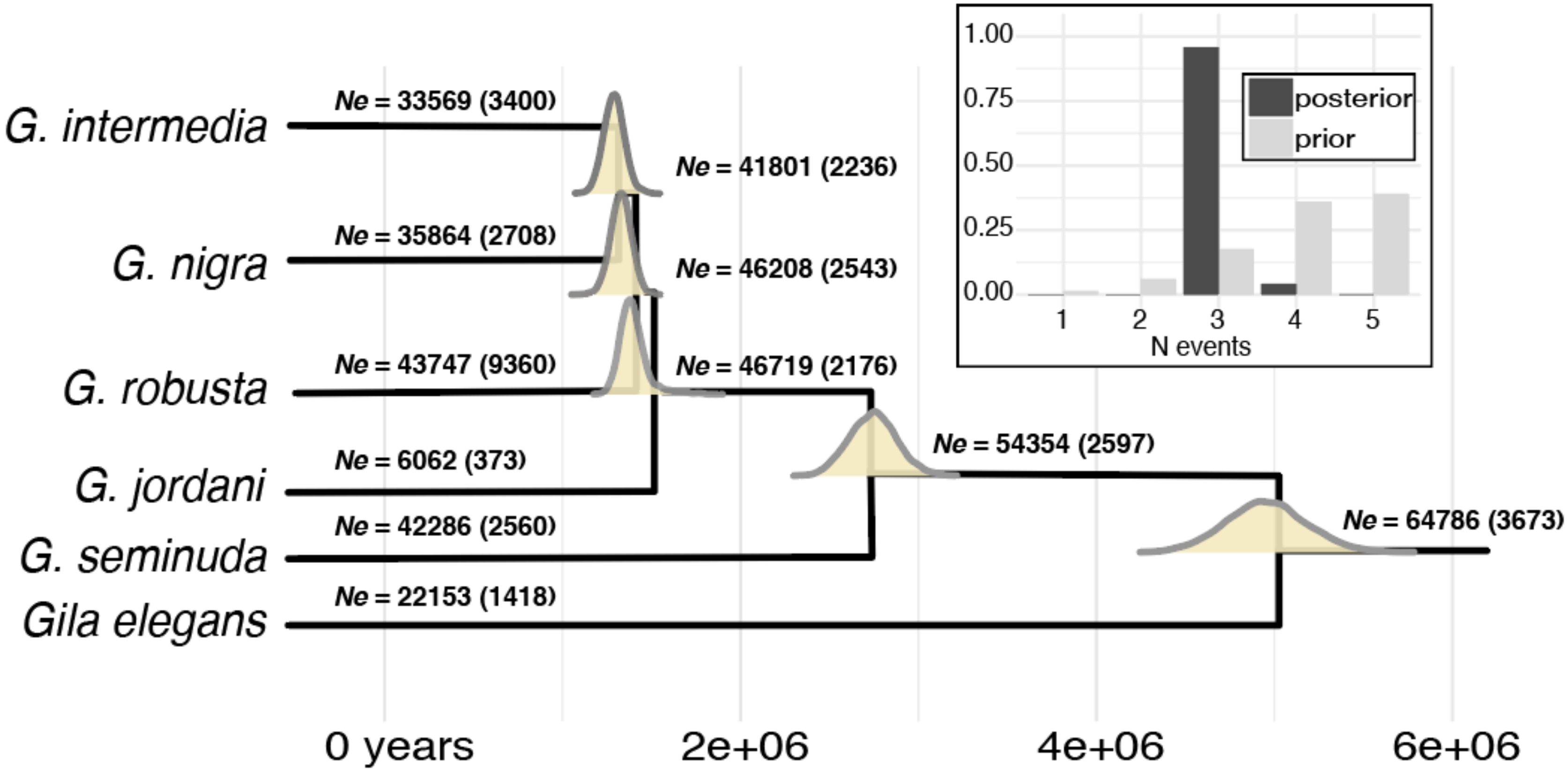

Figure 6: Posterior estimates for divergences times and effective populations sizes (N_e) derived from EcoEvolity and 2,000 randomly sampled full-length ddRAD locus alignments. Branches are annotated with mean (std. dev.) N_e and posterior probabilities for divergence times are plotted on corresponding nodes. Units are in years, using a static mutation rate of 1.2 e^{-08} substitutions per year. Posterior probabilities for divergence models (insert) suggest the co-divergence of *Gila jordani*, *G. robusta*, *G. nigra*, and *G. intermedia*

CHAPTER III

Genome-wide local ancestries discriminate homoploid hybrid speciation from secondary introgression in the red wolf (Canidae: *Canis rufus*)

Chafin TK, Douglas MR, Douglas ME

Abstract

Hybridization is well recognized as a driver of speciation, yet it often remains difficult to parse phylogenomically in that post-speciation gene flow frequently supersedes an ancestral signal. Here we examined how interactions between recombination and gene flow shaped the phylogenomic landscape of red wolf to create non-random retention of introgressed ancestry. Our re-analyses of genomic data recapitulate fossil evidence by demonstrating red wolf was indeed extant and isolated prior to more recent admixture with other North American canids. Its more ancient divergence, now sequestered within low-recombination regions on the X-chromosome (i.e., chromosomal 'refugia'), is effectively masked by multiple, successive waves of secondary introgression that now dominate its autosomal ancestry. These interpretations are congruent with more theoretical explanations that describe the manner by which introgression can be localized within the genome through recombination and selection. They also tacitly support the large-X effect, i.e., the manner by which loci that contribute to reproductive isolation can be enriched on the X-chromosome. By contrast, high recombination regions were enriched with very shallow gene trees reflecting compressed divergence estimates to 1/20th of that found in recombination 'cold spots', a likely product of post-speciation introgression. Our results effectively reconcile conflicting hypotheses regarding the impact of hybridization on evolution of North American canids and support an emerging framework within which the analysis of a

phylogenomic landscape structured by recombination can be used to successfully address the macroevolutionary implications of hybridization.

Introduction

Hybridization was once considered a rare event. However, its adaptive potential as a macroevolutionary process (i.e., an unfettered access to an extensive panoply of genetic variation; Grant and Grant 2019) has been enhanced by the widespread adoption of genomic approaches (Abbott et al. 2013; Twyford and Ennos 2012; Taylor and Larson 2019). As such, hybridization has now become one component of a more contemporary approach to species diversification. Prior to the onset of genomics, there were few examples of homoploid hybrid speciation in animals (Mavarez and Linares 2008), with notable exceptions being the Virgin River chub (*Gila seminuda*; DeMarais et al. 1992; Chafin et al. 2019) and the red wolf (*Canis rufus*; Wayne and Jenks 1991; Reich et al. 1999). Being that hybrid speciation is now becoming a common hypothesis (Yakimowski and Rieseberg 2014; Elgvin et al. 2017; Lamichhaney et al. 2018; Eberlein et al. 2019; Marques et al. 2019), we argue that a framework must now be developed so as to discriminate among its alternative outcomes [e.g. a more explicit definition of 'homoploid hybrid speciation;' (Schumer et al. 2014; Nieto Feliner et al. 2017; Schumer et al. 2018)].

Given the inherent difficulties associated with diagnosing hybrid speciation as the basis of reproductive isolation, it has often been defined on the basis of genomic mosaicism (Blanckaert and Bank 2018; Schumer et al. 2018a). However, doing so risks overlooking a more varied evolutionary role for hybridization. A contributory aspect is the recognized difficulty in detecting hybridization, for it is but one of several mechanisms driving phylogenetic discordance in the genome (Maddison 1997; Degnan and Rosenberg 2009). Hybridization–speciation dynamics are further complicated by the fact that evidence of archaic branching (i.e. those that precede introgression) can be depleted, and especially so in those lineages with a history of secondary introgression. However, the parsing of genealogical histories is dependent on the interactions between recombination, genetic drift, and selection (McGaugh et al. 2012; Schumer et al. 2018). As such, branching patterns are often retained non-randomly, with reduced permeability to gene flow found in those genomic areas with low recombination, where introgression of deleterious alleles is restricted by an increased efficacy of linked selection (Payseur and Rieseberg 2016; Runemark et al. 2018; Schumer et al. 2018).

The interaction between selection and recombination through time allows fundamental predictions to be made with regard to the stability of hybrids genomes, and this may promote the role that hybridization plays in a given lineage. In the generations following a hybridization event, recombination creates junction-points where ancestries transition from one parental genome to another (Fisher 1954). Their densities along the length of a chromosome can be used to find loci relating to hybrid fitness, because selection against incompatible loci will alter the breadth of correlated ancestry, depressing local recombination with proportionately larger distances between junctions (Sedghifar et al. 2016; Hvala et al. 2018).

We thus hypothesized if signatures of archaic introgression are indeed masked by secondary introgression, then the probability of observing the 'original' ancestry will increase as local recombination rates decrease [even when hybrid ancestries dominate, as is sometimes the case (Fontaine et al. 2015)]. Thus, our prediction is that patterns of coalescence will be multimodal, reflecting the times and manner by which populations have diverged and subsequently intermingled (Rosenberg and Feldman 2002). Here we explore how this distribution in the red wolf is shaped by genome structure and recombination rate heterogeneity. To do so, we test multiple opposing hypotheses regarding the role hybridization has played in the history of this species.

The red wolf as a case study

Our capacity to more precisely delineate hybridization has precipitated ancillary issues, such as the disparity that now exists between evolutionary complexity and species conservation (Ellstrand et al. 2010; Fitzpatrick et al. 2015; Supple and Shapiro 2018; vonHoldt et al. 2018). The U.S. Endangered Species Act (ESA 1973; 16 U.S.C. § 1531 et seq), as well as similar legislations globally, do not protect hybrids (Jackiw et al. 2015), despite scientific support (O'Brien and Mayr 1991; Allendorf et al. 2001; Haig and Allendorf 2006; Lind-Riehl et al. 2016). Few species have been as integral to this debate as red wolf (*Canis rufus*), fueled in part by the long-standing ambiguity surrounding its origins (Gittleman and Pimm 1991; Wayne and Jenks 1991; Dowling et al. 1992; Nowak 1992).

DNA evidence implicates hybridization, which some have attributed to recent coyote (*C. latrans*) and grey wolf (*C. lupus*) admixture (Wayne and Jenks 1991; Roy et al. 1996; Reich et al. 1999; vonHoldt et al. 2011; vonHoldt et al. 2016). Alternatively, others have instead argued that data point to an earlier red wolf origin, with introgression occurring as a subsequent phenomenon (Nowak 1979; Dowling et al. 1992; Nowak 1992; Wilson et al. 2000; Nowak 2002; Hohenlohe et al. 2017). Hypotheses regarding the status of red wolf are as follows (per Waples et al. 2018). It is: (1) An evolutionary distinct lineage derived from common ancestry with either *C. lupus* or *C. latrans* (=secondary introgression); (2) A transient product produced by contemporary hybridization (=hybrid swarm), or (3) An admixture subsidiary to a more ancient hybridization (=hybrid speciation). We discriminate among these scenarios by establishing predictions with regard to the respective footprint each would leave on the genomic landscape of red wolf, then testing each to ascertain which has the greatest probability of occurrence.

Methods

Read processing, quality filtering, and genotyping

We used previously published genomes for red wolf (=RW; *Canis rufus*), North American gray wolf (=GW; *Canis lupus*), and coyote (=COY; *Canis latrans*), with the red fox (=VUL; *Vulpes vulpes*) serving as an outgroup (vonHoldt et al. 2016a; Kukekova et al. 2018). Paired-end reads were downloaded from the NCBI SRA (SRR7107787; SRR7107783; SRR1518489; SRR5328101-115) and mapped to the domestic dog assembly (CanFam3.1) using BOWTIE2 (Langmead and Salzberg 2012) with sensitive settings, and excluding discordant pairs and unaligned reads. Further processing, sorting, and indexing was performed in SAMTOOLS (Li et al. 2009). PCR duplicates were filtered in PICARD (Broad Institute; broadinstitute.github.io/picard), followed by indel realignment and base quality recalibration in GATK (McKenna et al. 2010; Van der Auwera et al. 2013) as preparation for the HAPLOTYPECALLER pipeline using the 'Best Practices' workflow. Genotypes were then inferred jointly using GATK GENOTYPEGVCFS, followed by post-processing, quality filtering, and merging of variant and indel calls.

Genome-wide phylogenetic patterns

To examine topological and coalescent patterns, we first delimited ancestry blocks within full chromosomal pseudoalignments using a conservative phylogenetic approach, then built pseudoalignments from variant data using a custom Python code (github.com/tkchafin/vcf2msa.py). One issue with this approach is that one cannot assume the genomic reference state for a given nucleotide position will be consistent across the sampled genomes. Thus, within each genome, non-polymorphic bases were treated as un-callable (“N”) when local read depth was < 5. A single-pass algorithm was then used to examine variants (SNPs) for failure of the four-gamete condition (FGT; Hudson and Kaplan 1985). Given the resulting set of incompatible intervals, we then resolved a minimum set of ancestry breakpoints for which no FGT incompatibilities persisted (available as open-source; Chafin 2020).

Delimited blocks were then assigned ancestry using a phylogenetic method. Here, we computed a maximum likelihood estimate (MLE) in IQ-TREE (Nguyen et al. 2014) using integrated model selection and optimization of rate parameters. We discriminated weakly supported relationships by additionally calculating likelihoods under constrained topology searches for each possible quartet resolution, and testing for significant exclusion of alternatives from the MLEs by calculating a bootstrap proportion computed using the RELL approximation (Kishino et al. 1990). Sources of mixed support as a result of systematic errors were differentiated within a given block. For example, the incorrect spanning of recombination events (resulting in concatenated ancestry blocks) was separated from that due to unphased hybrid diplotypes by measuring interspecific heterozygosity. This was derived as the fraction of fixed nucleotide polymorphisms between coyote and gray wolves that were heterozygous for red wolf.

Testing for multiple-pulse and gradual admixture

Hybrid ancestries are expected to be arranged in large contiguous blocks following an admixture event, with the size of linkage blocks subsequently breaking down over time (Baird et al. 2003). The distribution of ancestry tracts lengths post-admixture can thus be used to understand the timings of genomic contributions (Gravel 2012; Liang and Nielsen 2014), as well as to discriminate multiple-pulse versus continuous admixture models (Zhou et al. 2017; Ni et al. 2018).

To explicitly test among these scenarios, we built a custom SNAKEMAKE pipeline (github.com/tkchafin/multiwaver_snakemake_workflow) for running MULTIWAVER_2.0 (Ni et al. 2018a). To do so, we converted from physical (bp) to genetic (cM) coordinates by utilizing the available comprehensive linkage map for the dog genome (Wong and Neff 2009; Wong et al. 2010). Because the linkage map was built for an earlier version of the assembly (CanFam2), we first converted them using a Python wrapper (available as open-source at github.com/tkchafin/scripts/liftoverCoords.py) for the UCSC LIFTOVER command-line utility (Hinrichs et al. 2006). We then used the LIFTOVER-converted linkage map to construct Marey maps (Siberchicot et al. 2017) for each chromosome, and convert junction positions using cubic interpolation. Ancestries were then assigned to each block based on the phylogenetic results, with blocks having interspecific heterozygosity >0.1 randomly haploidized. We generated 100 independent replicates for each chromosome so as to quantify stochastic variation caused by random 'pseudo-haploid' resolution.

Fitting full-genome admixture histories using coalHMMs

We inferred divergence time parameters using an MCMC (Markov Chain Monte Carlo) sampler for an admixture coalescent HMM (hidden Markov model). HMMs provide a means to probabilistically model transitions along serial or sequential datasets, and are employed widely in genomics and phylogenetics (gene prediction, Stanke and Waack 2003; nucleotide evolution, Yang 1995; Felsenstein and Churchill 1996; and patterns of phylogenetic and geographic diversification, Beaulieu and O'Meara 2016; Caetano et al. 2018). Coalescent HMMs

(=coalHMMs) construct a Markov model along a sequence alignment, with ‘hidden’ states as features to reconstruct (Dutheil et al. 2009; Li and Durbin 2011). Hidden states that represent genealogies or coalescent histories are themselves unobservable yet can be predicted from the observed states (=sequence data). Parameters involve processes controlling transitions among hidden states, such as recombination rates (r), effective population sizes (N_e), and speciation times (τ)(Dutheil et al. 2009). Often (as herein) the primary objective is to infer those demographic parameters from which transition rates are derived (Mailund et al. 2011). In the case of the admixture coalHMM (Cheng and Mailund 2015, 2020), HMMs implementing isolation-with-migration models (Mailund et al. 2012) are combined to generate a pseudolikelihood (or ‘composite’ likelihood) of more complex models that involve multiple lineages. Here we specified priors for the MCMC optimization using demographic estimates from vonHoldt et al. (2016).

Due to the computational complexity of the coalHMM approach (Cheng and Mailund 2015, 2020), the analysis was run separately in 1-million base blocks in two independent replicates per block. We then determined optimal burn-in values using an iterative approach (removing 5% of samples per iteration) using the Geweke diagnostic (Geweke 1992). We also computed effective sample sizes (ESS) for all parameters and assessed convergence of independent chains using the Gelman-Rubin convergence test (Gelman and Rubin 1992; Brooks and Gelman 1998) in the R package CODA (Plummer et al. 2006), removing any blocks for which any parameter-wise ESS fell below 100 or having a Gelman-Rubin statistic <1.01.

Coalescent demographic modeling

As in vonHoldt et al. (2016), we employed a protocol (Freedman et al. 2014) that targeted a reduced set of putatively neutral loci (1kb in length) for demographic modelling (via G-PHOCS; Gronau et al. 2011). We first excluded regions within a 10kb flanking distance of coding genes (Hoeppner et al. 2014), or conserved non-coding elements (CNEs). The latter were annotated using PHASTCONS scores (Siepel et al. 2005) provided for the *Euarchontoglires* clade, as mapped to the mouse genome (mm9) on UCSC (Freedman et al. 2014). CNEs were then defined as contiguous (over 50bp in length) PHASTCONS scores >0.7 (per Freedman et al. 2014). Interval coordinates for both CNEs and coding genes were converted to the CANFAM3.1 coordinate system (Hinrichs et al. 2006). Our filtered VCF, with reference genome and BED file defining excluded regions, were input to a generalized pipeline (Chafin et al. 2018) that allows for discovery of targeted sub-alignments in genomic datasets. Additional constraints targeted sub-alignments with a maximum proportion of 0.5 uncalled (N) or gap bases. We then subtracted regions from this which were identified as having heterozygous ancestry, and further sampled regions which were at least 100kb apart, truncating regions greater than 5kb in length. Resulting intervals were then extracted as full pseudo-alignments using custom Python code (github.com/tkchafin/vcf2msa.py), with an additional constraint that invariant bases for each species retain the reference base only where >5 reads present; lower-coverage bases were treated as un-callable (“N”). These were then divided into ‘sub-genomes’ by querying dominant phylogenetic ancestry assignments, removing alignments shorter than 500bp, resulting in N=6,100 and 6,255 for gray wolf and coyote sub-genomes (N=12225 loci in total). These served as input for demographic inference in G-PHOCS following the same protocol used in prior studies (Freedman et al. 2014; vonHoldt et al. 2016a).

Results

From prior publications (vonHoldt et al. 2016; Kukekova et al. 2018), we obtained whole-genome sequences for the red wolf (*Canis rufus*) and its putative progenitor species [the North American gray wolf (*C. lupus*) and coyote (*C. latrans*)], as well as an outgroup species (red fox; *Vulpes vulpes*). We aligned these data against the domestic dog genome (Kirkness et al. 2003; Lindblad-Toh et al. 2005; Hoeppner et al. 2014), then extracted full chromosome-length 'pseudoalignments' from all nucleotide positions having sufficient sequencing depth. This resulted in an average of 95.5% of the genome having called bases across species.

To identify sub-genomic ancestry blocks, we partitioned the 38 autosomes and the X chromosome into 913,849 non-overlapping windows by using an algorithm that defined a 'most parsimonious' set of hypothesized ancestry breakpoints, given a four-gamete assumption (Chafin 2020). This provided data with an average length of 2.2 kb (10.3 kb if merging consecutive ancestry blocks; see Fig. S14). We then analyzed each chromosome separately, and additionally partitioned regions by recombination rate, as inferred using an existing high-density linkage map (Fig. S15)(Wong and Neff 2009; Wong et al. 2010).

Our analyses are presented in two stages: The first examines the distribution of phylogenies across the genome. Here, we reasoned that sub-genomes from the putative parental lineages could be assigned via Maximum Likelihood (ML) estimates of the local branching order. We then identified heterozygous ancestry blocks by calculating the interspecific heterozygosity of red wolf sequence within each sub-alignment. We established relatively simple predictions for these early analyses: If indeed hybridization is recent (per Wayne and Jenks 1991; vonHoldt et al. 2016), then ancestry blocks will be large, given scant time for linkage blocks to be broken up by recombination (Falush et al. 2003; Pool and Nielsen 2009)]. Also, interspecific heterozygosity should be high (Rieseberg and Linder 1999; Anderson and Thompson 2002), whereas raw divergence will be low (vonHoldt et al. 2017). Likewise, we also expect an enrichment of introgressed histories in regions with higher recombination (Schumer et al. 2018; Li et al. 2019). We then used a model of linkage block decay to test several alternative models of hybridization, with gene flow either being gradual (e.g. declining through time from an initial event), continuous, or have occurred in multiple independent waves (Ni et al. 2018). However, when gene flow is high, signals of more ancient divergence could be 'masked.'

To untangle this, our second approach employed local phylogenetic signals as latent variables within a hidden Markov model [=coalHMMs (Dutheil et al. 2009; Spence et al. 2018)]. It allowed us to extract parameter estimates (e.g. divergence times) by integrating results from coalescent theory, despite the fact that the 'true' history at each nucleotide is masked (Hobolth et al. 2007; Dutheil et al. 2009). We then contrasted this approach with a second coalescent-based method [g-PhoCS; (Gronau et al. 2011)] that replicates the analyses of vonHoldt et al. (2016) with the exception that inputs were additionally partitioned by their respective sub-genomic histories.

Representation and divergence of parental genomes

Phylogenetic estimation and interspecific heterozygosity revealed that 26.8–36.5% of ancestral blocks were heterozygous (Fig. S16), with per-base gray wolf ancestry representing 23.2–41.7% (depending on measurement; Table S2). These results are congruent with previous studies that estimated 20–25% from SNP data (vonHoldt et al. 2011), and ~17–33% using microsatellite data (Roy et al. 1994; Bertorelle and Excoffier 1998). Of note, an anomalous sister-relationship of red wolf to red fox (as outgroup) was supported by 17.9% of the data, a likely result of direct introgression between coyote and gray wolf (Lehman et al. 1991; Gopalakrishnan et al. 2018; Pilot et al. 2019), and/ or inflated discordance due to bottlenecks in contemporary red wolf populations (Brzeski et al. 2014; Waples et al. 2018). Divergence from source genomes was remarkably low, with homozygous ancestry blocks across all chromosomes with D_{XY}=0 identified as 49.3% coyote and 54.5% gray wolf.

The distribution of ancestries was notably non-random (Fig. S17-S18), with a substantial enrichment of coyote ancestry on the X-chromosome (Fig. 1A). This was most pronounced in regions of low recombination (<0.5 cM/Mb). It thus comes as no surprise that a significantly higher mean recombination rate was found when gray wolf ancestry blocks were compared between autosomes and X-chromosome (where it is enriched at higher recombination rates; Table 1).

Testing the hybrid origin hypothesis

Given the observed distribution of recombination-structured ancestries, two questions emerge with regard to hybridization: (1) Did the temporal context of hybridization contribute to the non-random representation of ancestries? (2) Did hybridization occur as a 'homoploid hybrid speciation' event? or (3) Did admixture occur subsequent to a pre-existing isolation? To address these questions, we developed several predictions as a test mechanism in the context of a discriminative framework.

We first recognized the positive relationship between the efficacy of linked selection and the size of linkage blocks in the genome (Nachman and Payseur 2012). Given this, should there be signatures of isolation that pre-date admixture? If so, they would then be expected to occur with highest probability in those regions with low recombination. Likewise, introgressed ancestries are more probable within high-recombination regions where deleterious alleles can be more readily decoupled from neutral or beneficial surroundings (Schumer et al. 2018).

We found positive results in the non-random distribution of ancestries, where a more pronounced occurrence of enriched coyote ancestry was seen within low-recombination regions of the X-chromosome. To expand on this, we also predicted if introgression does indeed mask prior isolation, then those affected genomic regions would display a more shallow coalescence with respect to divergence events (Rosenberg and Feldman 2002; Leache et al. 2014). Juxtaposition of these predictions allowed us to test the hypothesis of hybrid origin versus secondary admixture: If older divergences predominate in areas of low recombination, then the 'original' branching pattern is retained (e.g. Fontaine et al. 2015). By partitioning divergence according to recombination rate, we can then unmask ancestral divergence previously obscured.

We fitted a coalescent hidden Markov model (coalHMM) implementing admixture (Cheng and Mailund 2020) to 1Mb blocks of the red wolf genome. We did so to obtain local estimates for red wolf divergence times with regard to coyote (τ_{COY}) and gray wolf (τ_{WOLF}) progenitors, as well as putative estimates of post-gene flow isolation (τ_{H}). In so doing, we uncovered a marked disparity in the range of these estimates between autosomes and the X-chromosome (Fig. 1B and S19–S21). The autosomal estimates were reasonably homogenous across recombination rate bins. This was not so on the X-chromosome: While τ_{WOLF} and τ_{H} were relatively consistent among recombination rate bins, τ_{COY} suggested 20-times older divergence in regions where cM/Mb < 0.5 than in regions where cM/Mb > 2.0 (μ=0.004 versus μ=0.0002). Thus, divergence was found to be substantially higher in low-recombination regions of the X-chromosome, with younger branching times instead dominating high-recombination regions and the autosomal genome.

We then sampled ~6000 putatively neutral regions from each parental red wolf sub-genome (following Freedman et al. 2014; vonHoldt et al. 2016) as a means of applying the same demographic modelling approach used in previous studies (i.e. g-PhoCS; Gronau et al. 2011). Results indicated much younger age estimates than those from the COALHMM approach (Fig. 2 and S22), with a mean posterior mutation-scaled divergence time τ_{COY}=3.8x10^{-5} and τ_{WOLF}=5.9x10^{-6} (Table S3). Assuming a generation time of three years and an average per-generation mutation rate of 4x10^{-9} (vonHoldt et al. 2016), these correspond to ~28,500 and ~4,425 years, respectively. These are congruent with COALHMM estimates taken from high recombination regions of the X chromosome. Interestingly, these results echo a known effect wherein the inclusion of introgressed gene histories promotes ‘tree compression,’ or an underestimation of divergence times (Leache et al. 2014). We also noted several long contiguous blocks showing complete loss of heterozygosity (LOH), in some cases stretching >25Mb (Fig. 3). However, there was no difference in branching time estimates among LOH and non-LOH segments (Fig. 3).

The observed multi-modality estimates in divergence time suggest multiple separate exchanges between red wolf and putative progenitors (Fig. 1B and S8). To assess this, we took advantage of another prediction: The expected decline in lengths of ancestry blocks over time, as a product of meiotic recombination (Gravel 2012; Ni et al. 2018). Here, the distribution of ancestry-tract lengths in each chromosome was best explained by either two- or three- pulse admixture models (Fig. S23-S25), with the exception of chr9, chr13, and chr37 which fit more appropriately with a gradual admixture model (e.g. with the rate of gene flow continually declining with time since an initial event). The timing of the most recent admixture among those displaying multiple waves (N=36/39) was estimated to be within the last few hundred generations. Older admixtures had a more diffuse distribution, ranging from ~250–2000 generations (Fig. 4). Mean admixture proportions for distinct pulses ranged from 0.286–0.533 for coyote, and 0.367–0.557 for gray wolf, although the estimated variance in older events was elevated (Fig. S24).

Discussion

Our findings suggest that extensive secondary introgression, as facilitated by increased permeability of autosomes relative to the X-chromosome, effectively obscured the pre-existing divergence of red wolf. In this sense, the autosomal genome is comparatively homogenous (Fig. 1A), with a low raw divergence and a systematic under-estimation of divergence times stemming from the predominance of introgressed ancestry (Fig. 1B). These results provide quantitative data in support of previous studies that found disproportionate retention of ancient branching patterns in low-recombining regions of sex chromosomes (Fontaine et al. 2015; Schumer et al. 2018; Edelman et al. 2019). This stems from a simultaneous reduction in the rate at which contiguous phylogenetic histories are degraded by linkage, as well as the bolstered efficacy of selection in purging deleterious introgressed elements (Nachman and Payseur 2012; Martin et al. 2019). Moreover, our replication of previous studies (Fig. 2) yielded substantially younger divergence estimates than those from previous studies partitioned by chromosome and recombination rate (Fig. 1B). Our observations agree with prior studies in underscoring the presence of 'tree compression,' or a branch lengths reduced/ distorted due to an inability to partition introgressed fron non-introgressed ancestries (Leache et al. 2014; Bangs et al. 2018).

The disparity between autosomes and the X-chromosome reiterates an established phenomenon found in those taxa exhibiting XY and ZW sex determination systems (Fontaine et al. 2015; Seixas et al. 2018; Martin et al. 2019). It is consistent with a 'large-X effect' that predicts loci contributing to reproductive isolation accumulate disproportionately on the X- (or Z-) chromosome (Coyne and Orr 1989; Van Belleghem et al. 2018; Presgraves 2018; Runemark et al. 2018). Our results also demonstrated an enrichment of coyote ancestry in low-recombination regions of the X-chromosome, whereas shallower divergence was found within high-recombination regions. A similar logic was presented in Fontaine et al. (2015), wherein a phylogenomic study of *Anopholes* mosquitos also revealed extensive conflict between autosomes and the X-chromosome in the locally dominant branching pattern. They reasoned that gene trees whose branching patterns reflect that of speciation rather than secondary introgression should exhibit deeper coalescence (Fontaine et al. 2015). Such a scenario would similarly explain the patterns of divergence observed in red wolf. Thus, we posit that while masked by secondary introgression in the majority of the genome, lower X-permeability acts as a barrier to exchange, effectively preserving those coalescent patterns established during a more ancient divergence of the red wolf with a coyote-like ancestor.

Reconciling conflict among genetic and morphological hypotheses

Previous analyses employing these data sparked considerable disagreement, primarily with regards to the timing of gene flow (vonHoldt et al. 2016; Hohenlohe et al. 2017; vonHoldt et al. 2017). The discrepancy between our results and prior studies stems from the predominance of shallow coalescence throughout most of the genome, with scant regions retaining signatures of prior ancestry. These results are instead most consistent with an older divergence between coyote and red wolf, an occurrence which has since been obscured by multiple pulses of contemporary admixture (Fig. 4 and S11).

We interpret our results as reconciling the conflict between inferences based on recent molecular work, and those stemming from analyses of modern and historical skeletal remains (e.g. Nowak, 1992). Indeed, the genome does indeed harbor signals of recent and ancient divergences, as established from multiple waves of admixture that successively degraded archaic branching patterns. The most recent admixture event is potentially associated with contemporary anthropogenic change. In this sense, morphological studies could not demonstrate hybridization until the early 1900s, when specimens began trending towards coyote morphologies (Nowak 1979; Nowak 1992; Nowak 2002).

One prevailing question is the status of red wolf prior to modern admixture. Our data suggests its earlier origin, although an absolute estimate is difficult to establish in that effective population sizes, mutation rates, and generation times are all indeterminate (Hohenlohe et al. 2017). Haplotype block lengths suggest admixture as old as ~1500–2000 generations (Fig. 4), which would place an upper bound extending into the early Holocene, depending on how generation time is defined. Fossil evidence suggests an ecological niche shift in coyote corresponding to megafaunal extinctions at the Pleistocene-Holocene boundary (Meachen and Samuels 2012). Individual body size during the transition period were intermediate between large Pleistocene individuals and more contemporary counterparts that were comparatively diminutive (Meachen et al. 2014). Response of canids to dietary shifts, demographic instability at the glacial-interglacial interface, and wide-spread shuffling of distributions (Pardi and Smith 2016; Loog et al. 2019) may have promoted interspecific contact. We suggest this scenario has plausibility, given the emerging adaptive role for hybridization now commonly evoked in diverse taxa (Lewontin and Birch 2006; Meier et al. 2017; Jones et al. 2018), to include canids (Kays et al. 2010; vonHoldt et al. 2016).

Conclusion

We employed a fine-scaled, partitioned analysis of genome-wide phylogenetic patterns in red wolf to show that branching patterns reflecting secondary introgression dominate. We also discovered that the presence of introgression varies throughout the genome, as a product of genomic context (e.g. local recombination rates). Because of this, evidence of older divergence was retained by only a fraction of the historically reduced recombination. These findings highlight the difficulties in studying the prevalence of hybridization within the broader Tree of Life, where a sufficiently large numbers of loci can presumably render a singular species history as transparent (Philippe et al. 2011; Hahn and Nakhleh 2016).

However, two biases hinder this approach: (1) The magnitude of signal among loci is clearly disproportionate (Arcila et al. 2017; Shen et al. 2017); and (2) signatures in the genome are deposited by different processes in a heterogeneous manner (as herein). Methods to correct for these biases must explicitly consider the rate of recombination that effectively drives this discrepancy (Payseur and Rieseberg 2016). A failure to do so with regard to red wolf yielded divergence estimates orders of magnitude less than those suggested by fossil evidence (Nowak 1992; Nowak 2002). We reconciled this discrepancy herein by employing estimates based solely on low recombinant regions of the X chromosome. Given this, a failure to partition distinct coalescent histories (e.g. Springer and Gatesy 2018) may result in some phylogenomic studies being interpreted as an artefact of substantial branch length distortion (Leache et al. 2014). The solution is to reconsider the non-random manner by which phylogenetic signal is retained in the genome. This was possible in our red wolf study due to the presence of substantial *a priori* data, to include chromosomal reference assemblies and high-density linkage maps. There are two stumbling blocks to the widespread application of this approach: Resource-limitations and methodological-deficiencies. Both are crucially important if we are to develop a more mature theory of hybridization as a macroevolution process (per Folk et al. 2018).

References

Abbott R., Albach D., Ansell S., Arntzen J.W., et al. 2013. Hybridization and speciation. J. Evol. Biol. 26:229–246.

Allendorf F.W., Leary R.F., Spruell P., Wenburg J.K. 2001. The problems with hybrids: Setting conservation guidelines. Trends Ecol. Evol. 16:613–622.

Anderson E.C., Thompson E. a. 2002. A model-based method for identifying species hybrids using multilocus data. Genetics. 160:1217–1229.

Arcila D., Ortí G., Vari R., Armbruster J.W., Stiassny M.L.J., Ko K.D., Sabaj M.H., Lundberg J., Revell L.J., Betancur-R. R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat. Ecol. Evol. 1:1–10.

Van der Auwera G.A., Carneiro M.O., et al. 2013. From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinforma. 43:11–10.

Baird S.J.E., Barton N.H., Etheridge A.M. 2003. The distribution of surviving blocks of an ancestral genome. Theor. Popul. Biol. 64:451–471.

Bangs M.R., Douglas M.R.M.E., Mussmann S.M., Douglas M.R.M.E. 2018. Unraveling historical introgression and resolving phylogenetic discord within *Catostomus* (Osteichthys : Catostomidae). BMC Evol. Biol. 18:86.

Beaulieu J.M., O'Meara B.C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst. Biol. 65:583–601.

Van Belleghem S.M., Baquero M., Papa R., Salazar C., McMillan W.O., Counterman B.A., Jiggins C.D., Martin S.H. 2018. Patterns of Z chromosome divergence among Heliconius species highlight the importance of historical demography. Mol. Ecol. 27:3852–3872.

Bertorelle G., Excoffier L. 1998. Inferring admixture proportions from molecular data. Mol. Biol. Evol. 15:1298–1311.

Blanckaert A., Bank C. 2018. In search of the Goldilocks zone for hybrid speciation. PLoS Genet. 14:1–23.

Brooks S.P., Gelman A. 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. 7:434–455.

Brzeski K.E., Rabon D.R., Chamberlain M.J., Waits L.P., Taylor S.S. 2014. Inbreeding and inbreeding depression in endangered red wolves (*Canis rufus*). Mol. Ecol. 23:4241–4255.

Caetano D.S., O'Meara B.C., Beaulieu J.M. 2018. Hidden state models improve state-dependent diversification approaches, including biogeographical models. Evolution. 72:2308–2324.

Chafin T.K. 2020. FGTpartitioner : A rapid method for parsimonious delimitation of ancestry breakpoints in large genome-wide SNP datasets. J. Open Source Softw. 5:2030.

Chafin T.K., Douglas M.R., Bangs M.R., Mussmann S.M., Douglas M.E. 2019. Taxonomic Uncertainty and Phylogenomics: Rescuing a Contentious Species Complex from the Anomaly Zone. bioRxiv. 692509.

Chafin T.K., Douglas M.R., Douglas M.E. 2018. MrBait: universal identification and design of targeted-enrichment capture probes. Bioinformatics. 34:4293–4296.

Cheng J.Y., Mailund T. 2015. Ancestral population genomics using coalescence hidden Markov models and heuristic optimisation algorithms. Comput. Biol. Chem. 57:80–92.

Cheng J.Y., Mailund T. 2020. Ancestral Population Genomics with Jocx, a Coalescent Hidden Markov Model. In: Dutheil J.Y., editor. Statistical Population Genomics, Methods in Molecular Biology vol. 2090. New York: Springer Protocols. p. 167–189.

Coyne J.A., Orr H.A. 1989. Patterns of speciation in Drosophila. Evolution. 43:362–381.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

DeMarais B.D., Dowling T.E., Douglas M.E., Minckley W.L., Marsh P.C. 1992. Origin of *Gila seminuda* (Teleostei: Cyprinidae) through introgressive hybridization: implications for evolution and conservation. Proc. Natl. Acad. Sci. U. S. A. 89:2747–2751.

Dowling T.E., DeMarais B.D., Minckley W.L., Douglas M.E., Marsh P.C. 1992. Use of genetic characters in conservation biology. Conserv. Biol. 6:7–8.

Dutheil J.Y., Ganapathy G., Hobolth A., Mailund T., Uyenoyama M.K., Schierup M.H. 2009. Ancestral population genomics: The coalescent hidden Markov model approach. Genetics. 183:259–274.

Eberlein C., Hénault M., Fijarczyk A., Charron G., Bouvier M., Kohn L.M., Anderson J.B., Landry C.R. 2019. Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. Nat. Commun. 10:923.

Edelman N.B., Frandsen P.B., Miyagi M., Clavijo B., Davey J., Dikow R.B., García-accinelli G., Belleghem S.M. Van, Patterson N. 2019. Genomic architecture of introgression shape a butterfly radiation. Science. 599:594–599.

Elgvin T.O., Trier C.N., Tørresen O.K., Hagen I.J., Lien S., Nederbragt A.J., Ravinet M., Jensen H., Sætre G.P. 2017. The genomic mosaicism of hybrid speciation. Sci. Adv. 3:1–16.

Ellstrand N.C., Biggs D., Kaus A., Lubinsky P., McDade L.A., Preston K., Prince L.M., Regan H.M., Rorive V., Ryder O.A., Schierenbeck K.A. 2010. Got Hybridization? A Multidisciplinary Approach for Informing Science Policy. Bioscience. 60:384–388.

Falush D., Stephens M., Pritchard J.K. 2003. Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. Genetics. 164:1567–1587.

Felsenstein J., Churchill G.A. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93–104.

Fisher R.A. 1954. A fuller theory of "junctions" in inbreeding. Heredity. 8:187–197.

Fitzpatrick B.M., Ryan M.E., Johnson J.R., Corush J., Carter E.T. 2015. Hybridization and the species problem in conservation. Curr. Zool. 61:206–216.

Folk R.A., Soltis P.S., Soltis D.E., Guralnick R. 2018. New prospects in the detection and comparative analysis of hybridization in the tree of life. Am. J. Bot. 105:364–375.

Fontaine M.C., Pease J.B., Steele A., et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. 347:1258524.

Freedman A.H., Gronau I., Schweizer R.M., et al. 2014. Genome sequencing highlights the dynamic early history of dogs. PLoS Genet. 10:e1004016.

Gelman A., Rubin D.B. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–472.

Geweke J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Fed. Reserv. Bank Minneapolis, Res. Dep.

Gittleman J.L., Pimm S.L. 1991. Crying wolf in North America. Nature. 351:524–525.

Gopalakrishnan S., Sinding M.H.S., Ramos-Madrigal J., et al. 2018. Interspecific gene flow shaped the evolution of the genus Canis. Curr. Biol. 28:3441–3449.

Grant P.R., Grant B.R. 2019. Hybridization increases population variation during adaptive radiation. Proc. Natl. Acad. Sci. U. S. A. 116:23216–23224.

Gravel S. 2012. Population genetics models of local ancestry. Genetics. 191:607–619.

Gronau I., Hubisz M.J., Gulko B., Danko C.G., Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. 43:1031–1035.

Hahn M.W., Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution. 70:7–17.

Haig S.M., Allendorf F.W. 2006. Hybrids and Policy. In: Scott J.M., Goble D.D., Davis F., editors. The Endangered Species Act at Thirty: Conserving Biodiversity in Human-Dominated Landscapes. Island Press. p. 150–163.

Hinrichs A.S., Karolchik D., Baertch R., et al. 2006. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 34:D590–D598.

Hobolth A., Christensen O.F., Mailund T., Schierup M.H. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 3:0294–0304.

Hoeppner M.P., Lundquist A., Pirun M., et al. 2014. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. PLoS One. 9:e91172.

Hohenlohe P.A., Rutledge L.Y., Waits L.P., et al. 2017. Comment on "Whole genome sequence analysis shows two endemic species of North American Wolf are admixtures of the coyote and gray Wolf." Sci. Adv. 3:1–4.

Hudson R.R., Kaplan N.L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 111:147–164.

Hvala J.A., Frayer M.E., Payseur B.A. 2018. Signatures of hybridization and speciation in genomic patterns of ancestry. Evolution.:1540–1552.

Jackiw R.N., Mandil G., Hager H.A. 2015. A framework to guide the conservation of species hybrids based on ethical and ecological considerations. Conserv. Biol. 29:1040–1051.

Jones M.R., Mills L.S., Alves P.C., Callahan C.M., Alves J.M., Lafferty D.J.R., Jiggins F.M., Jensen J.D., Melo-Ferreira J., Good J.M. 2018. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. Science. 360:1355–1358.

Kays R., Curtis A., Kirchman J.J. 2010. Rapid adaptive evolution of northeastern coyotes via hybridization with wolves. Biol. Lett. 6:89–93.

Kirkness E.F., Bafna V., Halpern A.L., Levy S., Remington K., Rusch D.B., Delcher A.L., Pop M., Wang W., Fraser C.M., Venter J.C. 2003. The dog genome: Survey sequencing and comparative analysis. Science. 301:1898–1903.

Kishino H., Miyata T., Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31:151–160.

Kukekova A. V, Johnson J.L., Xiang X., et al. 2018. Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours. Nat. Ecol. Evol. 2:1479–1491.

Lamichhaney S., Han F., Webster M.T., Andersson L., Grant B.R., Grant P.R. 2018. Rapid hybrid speciation in Darwin's finches. Science. 359:224–228.

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9:357–359.

Leache A.D., Harris R.B., Rannala B., Yang Z. 2014. The influence of gene flow on species tree estimation: A simulation study. Syst. Biol. 63:17–30.

Lehman N., Eisenhawer A., Hansen K., Mech L.D., Peterson R.O., Gogan P.J.P., Wayne R.K. 1991. Introgression of coyote mitochondrial DNA into sympatric North American gray wolf populations. Evolution. 45:104.

Lewontin R.C., Birch L.C. 2006. Hybridization as a source of variation for adaptation to new environments. Evolution. 20:315.

Li G., Figueiró H. V., Eizirik E., Murphy W.J., Yoder A. 2019. Recombination-aware phylogenomics reveals the structured genomic landscape of hybridizing cat species. Mol. Biol. Evol. 36:2111–2126.

Li H., Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25:2078–2079.

Liang M., Nielsen R. 2014. The lengths of admixture tracts. Genetics. 197:953–967.

Lind-Riehl J.F., Mayer A.L., Wellstead A.M., Gailing O. 2016. Hybridization, agency discretion, and implementation of the U.S. Endangered Species Act. Conserv. Biol. 30:1288–1296.

Lindblad-Toh K., Wade C.M., Mikkelsen T.S., et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 438:803–819.

Loog L., Thalmann O., Sinding M.H.S., Schuenemann V.J., et al. 2019. Ancient DNA suggests modern wolves trace their origin to a Late Pleistocene expansion from Beringia. Mol. Ecol.:1–15.

Maddison W. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Mailund T., Dutheil J.Y., Hobolth A., Lunter G., Schierup M.H. 2011. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. PLoS Genet. 7:1–15.

Mailund T., Halager A.E., Westergaard M., et al. 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet. 8:e1003125.

Marques D.A., Meier J.I., Seehausen O. 2019. A combinatorial view on speciation and adaptive radiation. Trends Ecol. Evol. 34:531–544.

Martin S.H., Davey J.W., Salazar C., Jiggins C.D. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. PLoS Biol. 17:1–28.

Mavarez J., Linares M. 2008. Homoploid hybrid speciation in animals. Mol. Ecol. 17:4181–4185.

McGaugh S.E., Heil C.S.S., Manzano-Winkler B., Loewe L., Goldstein S., Himmel T.L., Noor M.A.F. 2012. Recombination modulates how selection affects linked sites in *Drosophila*. PLoS Biol. 10:e1001422.

McKenna A., Hanna M., Banks E., et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Meachen J.A., Janowicz A.C., Avery J.E., Sadleir R.W. 2014. Ecological changes in coyotes (*Canis latrans*) in response to the ice age megafaunal extinctions. PLoS One. 9:1–15.

Meachen J.A., Samuels J.X. 2012. Evolution in coyotes (*Canis latrans*) in response to the megafaunal extinctions. Proc. Natl. Acad. Sci. U. S. A. 109:4191–4196.

Meier J.I., Marques D.A., Mwaiko S., Wagner C.E., Excoffier L., Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat. Commun. 8:14363.

Nachman M.W., Payseur B.A. 2012. Recombination rate variation and speciation: Theoretical predictions and empirical results from rabbits and mice. Philos. Trans. R. Soc. B Biol. Sci. 367:409–421.

Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2014. IQ-TREE: A fast and effective stochastic algorithm for estimating Maximum-Likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Ni X., Yuan K., Liu C., Feng Q., Tian L., Ma Z., Xu S. 2018a. MultiWaver 2.0: modeling discrete and continuous gene flow to reconstruct complex population admixtures. Eur. J. Hum. Genet. 27:133–139.

Ni X., Yuan K., Yang X., Feng Q., Guo W., Ma Z., Xu S. 2018b. Inference of multiple-wave admixtures by length distribution of ancestral tracks. Heredity. 121:52–63.

Nieto Feliner G., Álvarez I., Fuertes-Aguilar J., et al. 2017. Is homoploid hybrid speciation that rare? An empiricist's view. Heredity. 118:513–516.

Nowak R.M. 1979. North American Quaternary *Canis*. Monogr. Museum Nat. Hist. Univ. Kansas. 6:154pp.

Nowak R.M. 1992. The red wolf is not a hybrid. Conserv. Biol. 6:593–595.

Nowak R.M. 2002. The original status of wolves in eastern North America. Southeast. Nat. 1:95–130.

O'Brien S.J., Mayr E. 1991. Bureaucratic mischief: Recognizing endangered species and subspecies. Science. 251:1187–1188.

Pardi M.I., Smith F.A. 2016. Biotic responses of canids to the terminal Pleistocene megafauna extinction. Ecography. 39:141–151.

Payseur B.A., Rieseberg L.H. 2016. A genomic perspective on hybridization and speciation. Mol. Ecol. 25:2337–2360.

Philippe H., Brinkmann H., Lavrov D. V., et al. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol. 9:e1000602.

Pilot M., Moura A.E., Okhlopkov I.M., et al. 2019. Global phylogeographic and admixture patterns in grey wolves and genetic legacy of an ancient Siberian lineage. Sci. Rep. 9:1–13.

Plummer M., Best N., Cowles K., Vines K. 2006. CODA: Convergence diagnosis and output analysis for MCMC. R News. 6:7–11.

Pool J.E., Nielsen R. 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics. 181:711–719.

Presgraves D.C. 2018. Evaluating genomic signatures of "the large X-effect" during complex speciation. Mol. Ecol. 27:3822–3830.

Reich D.E., Wayne R.K., Goldstein D.B. 1999. Genetic evidence for a recent origin by hybridization of red wolves. Mol. Ecol. 8:139–144.

Rieseberg L.H., Linder C.R. 1999. Hybrid classification: Insights from genetic map-based studies of experimental hybrids. Ecology. 80:361–370.

Rosenberg N., Feldman M. 2002. The relationship between coalescence times and population divergence times. In: Slatkin M.L., Veuille M., editors. Devolopments in Theorical Populations Genetics. Oxford University Press. p. 130–159.

Roy M.S., Gefenj E., Wayne R.K., Smith D., Ostrander E.A. 1994. Patterns of differentiation and hybridization in North American wolflike canids, revealed by analysis of microsatellite loci. Mol. Biol. Evol. 11:553–570.

Roy M.S., Geffen E., Smith D., Wayne R.K. 1996. Molecular genetics of pre-1940 red wolves. Conserv. Biol. 10:1413–1424.

Runemark A., Fernández L.P., Eroukhmanoff F., Sætre G.P. 2018a. Genomic contingencies and the potential for local adaptation in a hybrid species. Am. Nat. 192:10–22.

Runemark A., Trier C.N., Eroukhmanoff F., Hermansen J.S., Matschiner M., Ravinet M., Elgvin T.O., Sætre G.P. 2018b. Variation and constraints in hybrid genome formation. Nat. Ecol. Evol. 2:549–556.

Schumer M., Rosenthal G.G., Andolfatto P. 2014. How common is homoploid hybrid speciation? Evolution. 68:1553–1560.

Schumer M., Rosenthal G.G., Andolfatto P. 2018a. What do we mean when we talk about hybrid speciation? Heredity. 120:379–382.

Schumer M., Xu C., Powell D.L., Durvasula A., Skov L., Holland C., Blazier J.C., Sankararaman S. 2018b. Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science. 660:656–660.

Sedghifar A., Brandvain Y., Ralph P. 2016. Beyond clines: lineages and haplotype blocks in hybrid zones. Mol. Ecol. 25:2559–2576.

Seixas F.A., Boursot P., Melo-Ferreira J. 2018. The genomic impact of historical hybridization with massive mitochondrial DNA introgression. Genome Biol. 19:1–20.

Shen X.X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:1–10.

Siberchicot A., Bessy A., Guéguen L., Marais G.A.B. 2017. MareyMap online: A user-friendly web application and database service for estimating recombination rates using physical and geneticmaps. Genome Biol. Evol. 9:2506–2509.

Siepel A., Bejerano G., Pedersen J.S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.

Spence J.P., Steinrücken M., Terhorst J., Song Y.S. 2018. Inference of population history using coalescent HMMs: review and outlook. Curr. Opin. Genet. Dev. 53:70–76.

Springer M.S., Gatesy J. 2018. Delimiting coalescence genes (C-genes) in phylogenomic data sets. Genes. 9:1–19.

Stanke M., Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 19:215–225.

Supple M.A., Shapiro B. 2018. Conservation of biodiversity in the genomics era. Genome Biol. 19:1–12.

Taylor S.A., Larson E.L. 2019. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. Nat. Ecol. Evol. 3:170–177.

Twyford A.D., Ennos R.A. 2012. Next-generation hybridization and introgression. Heredity. 108:179–189.

vonHoldt B.M., Brzeski K.E., Wilcove D.S., Rutledge L.Y. 2018. Redefining the role of admixture and genomics in species conservation. Conserv. Lett. 11:1–6.

vonHoldt B.M., Cahill J.A., Fan Z., Gronau I., Robinson J., Pollinger J.P., Shapiro B., Wall J., Wayne R.K. 2016a. Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. Sci. Adv. 2:e1501714–e1501714.

vonHoldt B.M., Cahill J.A., Gronau I., Shapiro B., Wall J., Wayne R.K. 2017. Response to Hohenlohe et al. Sci. Adv. 3:1–3.

vonHoldt B.M., Kays R., Pollinger J.P., Wayne R.K. 2016b. Admixture mapping identifies selectively introgressed genomic regions in North American canids. Mol. Ecol. 25:2443–2453.

vonHoldt B.M., Pollinger J.P., Earl D.A., et al. 2011. A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Res. 21:1294–1305.

Waples R.S., Kays R., Fredrickson R.J., Pacifici K., Mills L.S. 2018. Is the red wolf a listable unit under the US Endangered Species Act? J. Hered. 109:585–597.

Wayne R.K., Jenks S.M. 1991. Mitochondrial DNA analysis implying extensive hybridization of the endangered red wolf *Canis rufus*. Nature. 351:565–568.

Wilson P.J., Grewal S., Lawford I.D., et al. 2000. DNA profiles of the eastern Canadian wolf and the red wolf provide evidence for a common evolutionary history independent of the gray wolf. Can. J. Zool. 78:2156–2166.

Wong A.K., Neff M.W. 2009. DOGSET: Pre-designed primer sets for fine-scale mapping and DNA sequence interrogation in the dog. Anim. Genet. 40:569–571.

Wong A.K., Ruhe A.L., Dumont B.L., Robertson K.R., Guerrero G., Shull S.M., Ziegle J.S., Millon L. V., Broman K.W., Payseur B.A., Neff M.W. 2010. A comprehensive linkage map of the dog genome. Genetics. 184:595–605.

Yakimowski S.B., Rieseberg L.H. 2014. The role of homoploid hybridization in evolution: A century of studies synthesizing genetics and ecology. Am. J. Bot. 101:1247–1258.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics. 139:993–1005.

Zhou Y., Yuan K., Yu Y., Ni X., Xie P., Xing E.P., Xu S. 2017. Inference of multiple-wave population admixture by modeling decay of linkage disequilibrium with polynomial functions. Heredity. 118:503–510.

Appendix

Table 1: Recombination rate differences between homozygous and heterozygous ancestry blocks in autosomes versus the X-chromosome. Recombination rate is reported as the ratio of centimorgan (cM) per Mb and partitioned into homozygous coyote and gray wolf ancestry and heterozygous ancestry (=HET). Significance is reported for Mann-Whitney *U* test comparing X-chromosome and autosome cM/Mb within each partition.

	AUTOSOMES		X-ONLY		
	cM/Mb	*N*	cM/Mb	*N*	*P*-value
PHY subset					
Coyote	1.00	22996	0.97	2087	
Gray Wolf	1.01	16648	1.15	940	<0.0001
Het.	1.14	3859	1.03	43	<0.0001
All blocks					
Coyote	1.07	90978	1.05	4477	
Gray Wolf	1.07	101999	1.13	3158	<0.0001
Het.	1.06	106550	1.03	509	

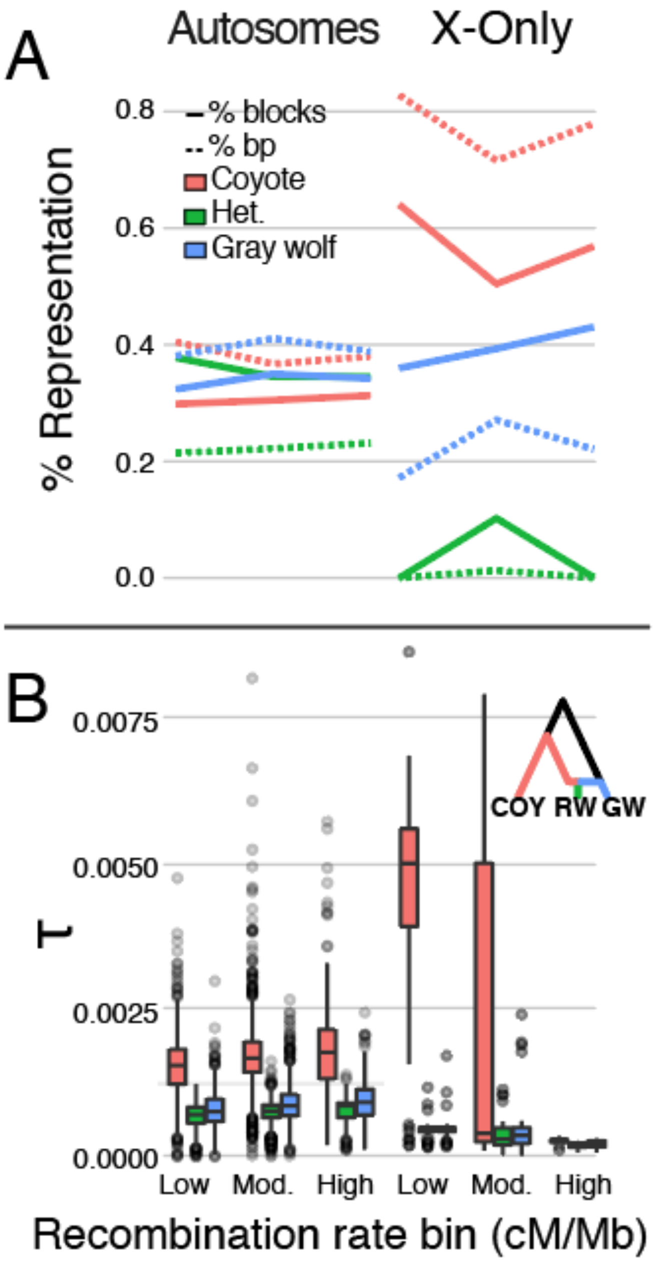

Figure 1: Effect of recombination rate on genomic ancestry proportions (A) and absolute divergence times (B) in the red wolf genome, partitioned among autosomes and X-chromosome. Percent ancestry is computed among genomic ancestry blocks which could be assigned as heterozygous (green), homozygous gray-wolf (blue), or homozygous coyote (red). Genomic representation is reported both as percentage of ancestry blocks (solid) and percentage of base-pairs (bp; dashed). Divergence times (τ) measured in expected substitutions were estimated using coalescent hidden Markov models (coalHMM) applied to 1Mb blocks. Results are further partitioned by local recombination rate (cM/Mb) binned as 'high' (>=2.0), 'moderate' (0.5–2.0), and 'low' (<=0.5), and show posterior estimates of: 1) red wolf–coyote divergence (τ_{COY}; red); 2) red wolf–gray wolf divergence (τ_{WOLF}; blue); and 3) the time of post-hybridization isolation (τ_H; green).

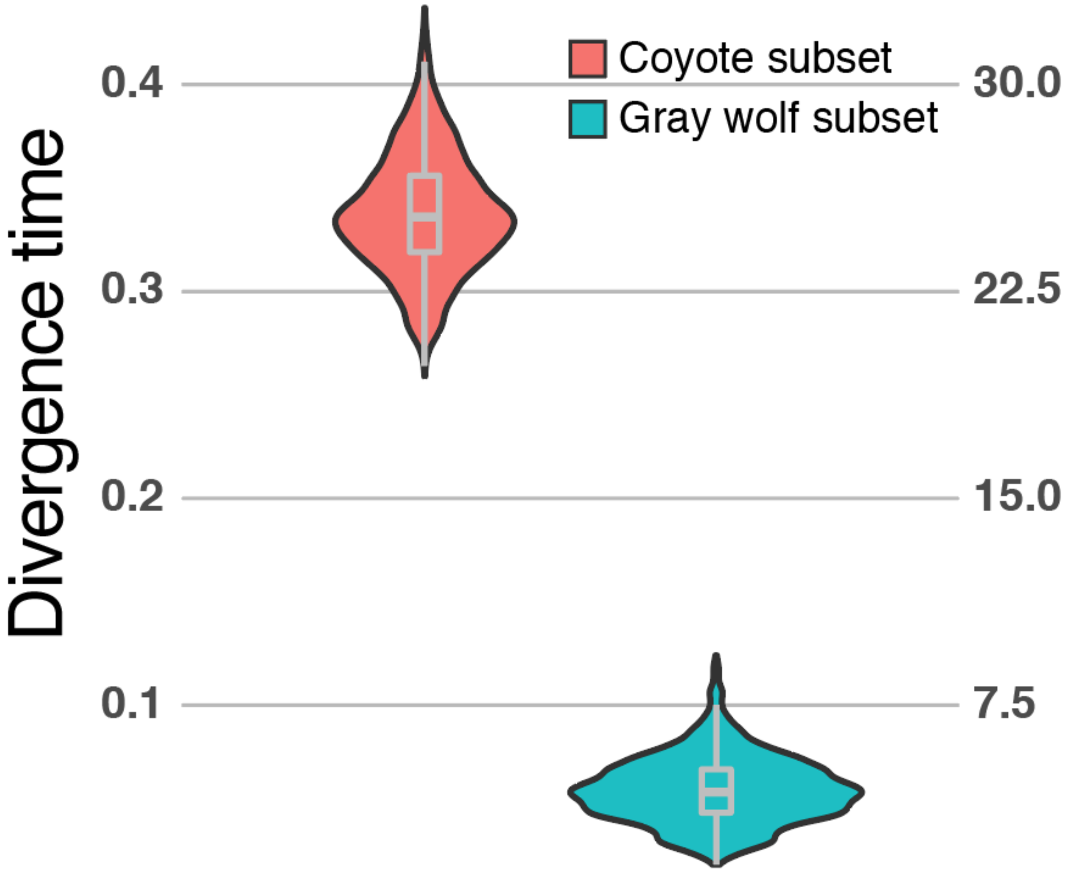

Figure 2: Divergence times estimated using g-PhoCS applied to parental sub-genomes of the red wolf. Divergence times (τ) measured in expected substitutions are shown for the red wolf–coyote divergence (τ_{COY}; red) and red wolf–gray wolf divergence (τ_{WOLF}; blue). Values are scaled up by a factor of 10,000 (left y-axis) and also provided in calibrated form (right y-axis) in thousands of years, assuming a generation time of three years and an average per-generation mutation rate of μ=4x10^{-9} / base pair.

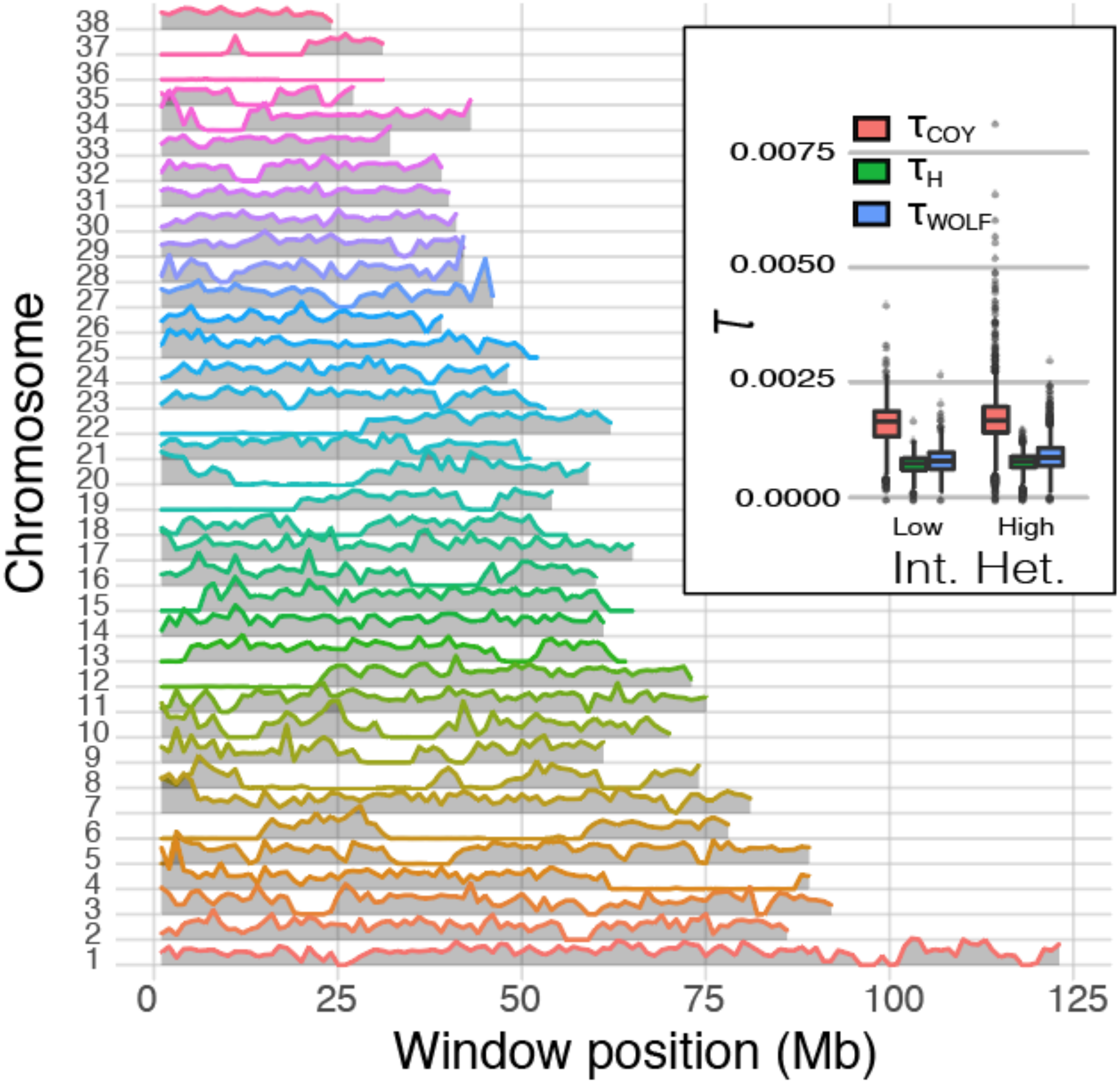

Figure 3: bution of interspecific heterozygosity in the red wolf genome, and inferred divergence ages (inset) within low (<0.1 mean interspecific heterozygosity) and high (>=0.1) regions of autosomes. Interspecific heterozygosity was computed as a weighted mean among delimited ancestry blocks encompassed by 1Mb non-overlapping sliding windows.

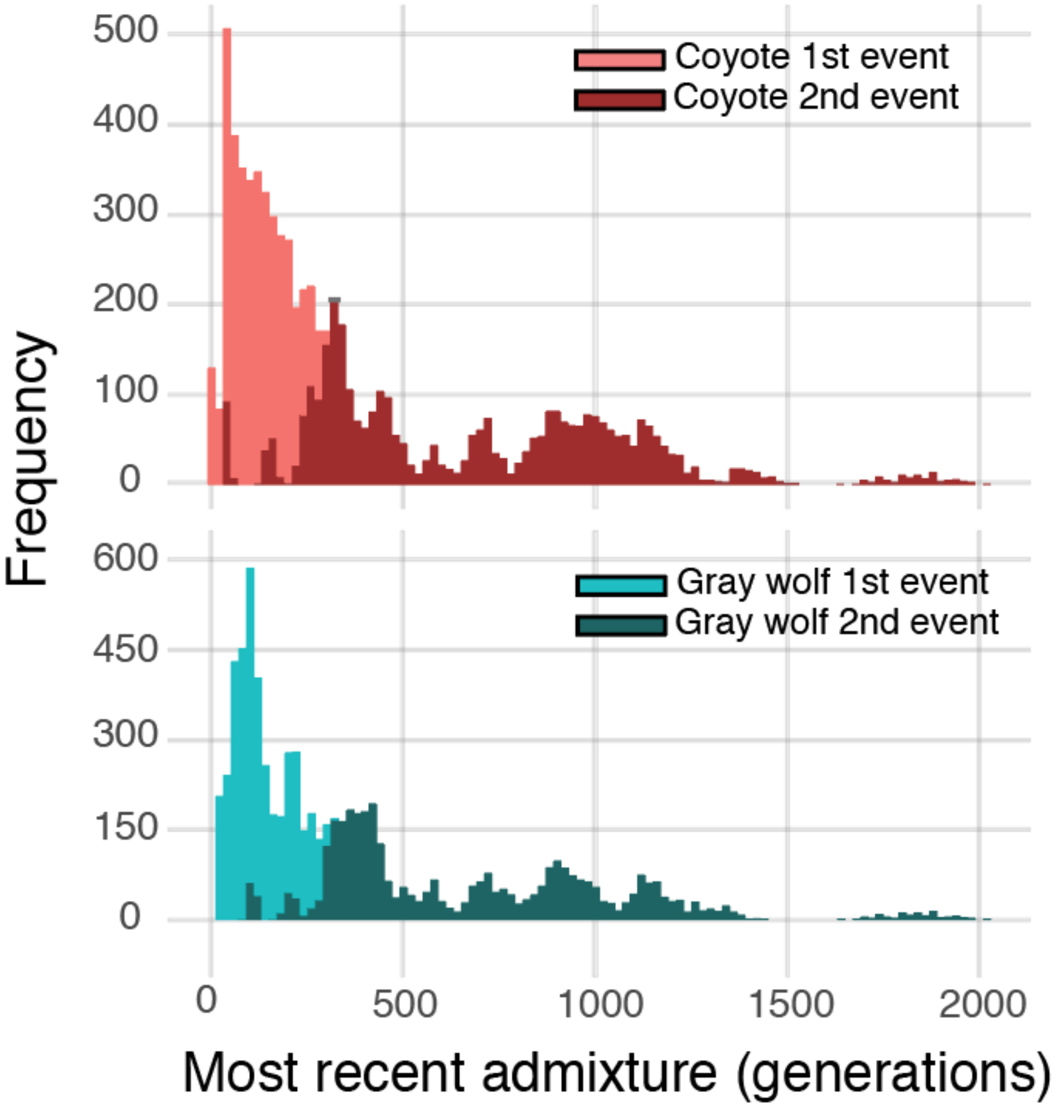

Figure 4: Times of inferred admixture events from coyote (above) and gray wolf (below) into the red wolf genome, measured in generations before the present. Results are shown aggregated from all chromosomes, excluding N=3 chromosomes for which a single-pulse gradual admixture model was selected.

CHAPTER IV

FRAGMATIC: *in silico* locus prediction and its utility in optimizing ddRADseq projects

Chafin TK, Martin BT, Mussmann SM, Douglas MR, Douglas ME

Abstract

Reduced-representation genomic methods are an invaluable data acquisition tool for conservation geneticists, yet *a priori* estimates of locus recovery are difficult for non-model organisms. We present a simple *in silico* approach (FRAGMATIC) that predicts locus recovery in ddRAD sequencing which utilizes genomic data for related organisms. Its applicability was tested by quantifying prediction accuracy versus genetic distances across five non-model organisms and reference genomes for related organisms of varying phylogenetic distance. We additionally examined sensitivity of the method using one organism (*Danio rerio*) with an available genome. FRAGMATIC supports population genomic projects in non-model species by providing *a priori* estimates of targeted ddRAD loci that, in turn, will curb wasted sequencing effort and optimize cost-efficiency. Validation shows that while predictive error is minimized when applied to a closely related reference genome, *in silico* estimates may also be robust to deeper (e.g. within-family) relationships, although weak correlation suggests that specific characteristics of genome architecture may be more predictive than genetic distance. This indicates that a more extensive exploration of genomes, including a broader taxonomic scope (e.g. beyond vertebrates), may be informative. All code is freely available at: https://github.com/tkchafin/fragmatic.

Introduction

Extending next-generation sequencing to non-model organisms often require reduction of genomic complexity that is replicable across related individuals (Edwards et al. 2015; Andrews et al. 2016). This can be accomplished by “targeted” fragmentation of DNA via restriction digest. Such reduced-representation genomic approaches that utilize restriction-site associated DNA (RAD) have been used widely (Baird et al. 2008), and promote the study of species for which few genomic resources exist.

A modification of RAD (i.e., ddRAD) employs a second restriction enzyme that provides greater uniformity in the fragments produced (Peterson et al. 2012). Subsequent size selection of these fragments, often via gel-excision, then reduces genomic complexity. Homologous and randomly interspersed sequences are recovered across closely related populations or species (Davey et al. 2011), under the assumption that shared restriction sites are distributed equitably across shallow phylogenetic scales. Adopting this strategy, a targeted number of loci can be obtained with great flexibility in economy and scale (Puritz et al. 2014).

A major consideration when designing ddRAD projects is to reduce sequencing of non-informative, over-redundant regions common in eukaryotic genomes (de Koning et al. 2011). Enrichment of repetitive elements limits cost efficiency and biases library composition, yet these can be minimized when enzyme selection is informed by genomic data (Heffelfinger et al. 2014). In the case of traditional RADseq, locus recovery may be predicted on a probabilistic basis, although Herrera et al. (2014) pointed out that these estimates are sensitive to *a priori* estimates of genome size and composition. Simulation has also been presented as a potential means to predict loci when more complicated RAD-protocols are involved [e.g. digestion with multiple enzymes (Lepais and Weir 2014)]. We developed a custom in silico utility for locus prediction with a primary focus on simplicity of use, termed FRAGMATIC, and evaluated its applicability for non-model organisms.

Methods

Locus Prediction

FRAGMATIC estimates all possible ddRAD fragments by partitioning input sequences at user-specified restriction sites. When a fully assembled reference genome is available, FRAGMATIC will predict a very accurate number of loci. It can also digest a contig-level assembly, although the already fragmentary input may result in a bias favoring small loci. Contig-level assemblies can be randomly concatenated as a potential solution, yet with spurious fragments produced that may traverse artificially manufactured contig boundaries. If one lacks a scaffold-level assembly containing estimated gaps, a resampling method could be implemented to iteratively rearrange contigs into replicates and to assess variance induced by arbitrary contig splicing. FRAGMATIC is extensible to any number of restriction sites, including those with degenerate bases. Fragments are categorized by flanking restriction sites, with output tabulated as fragment length frequencies or sequences. For example, an *in silico* digest with two enzymes, such as *PstI* (CTGCAG) and *MspI* (CCGG), will yield fragments flanked b: two *PstI* sites; two *MspI* sites; a *PstI* and *MspI* site (the target of *in vitro* sequencing in a ddRAD study); as well as fragments lacking a restriction site at one or either end ('Missing_sites' in the FRAGMATIC output). The latter is more frequent with incomplete assemblies, and represents fragments prematurely terminated by contig boundaries. A similar pattern would occur in vitro by digesting highly-degraded DNA. In practice, recovered loci can be skewed by "small fragment carryover" (DaCosta and Sorenson 2014) and non-canonical enzyme activity (Kamps-Hughes et al. 2013). Recovery is also affected by methylation, restriction site mutations [allelic drop-out (Gautier et al. 2013)], and variance in library preparation.

Validation

Given the above, FRAGMATIC might be expected to over predict loci. To test this prediction, we generated fragment distributions for 16 reference genomes from NCBI for comparison with *in vitro* digests of five related non-model organisms. To evaluate accuracy, we also compared *in vitro* digests of *Danio rerio* versus *in silico* FRAGMATIC results derived from its genome (Howe et al. 2013). Sample preparation followed a standard ddRAD protocol and samples were sequenced in a multiplexed lane with other projects (Peterson et al. 2012) with sequences processed using pyRAD (Eaton 2014).

For *Danio*, we compared locus recovery among *in vitro* and *in silico* digests using PstI and MspI, with a size selection of 250 – 350 bp (excluding adaptor sequences). FRAGMATIC predicted 57,688 loci, whereas clustering raw Illumina reads (N=888,260) using VSEARCH (Rognes et al. 2016) with a similarity threshold of 95% and minimum alignment length of 90% produced 49,813 presumptive loci excluding singletons (<10% of sequences). Of the recovered raw Illumina reads, 97.7% overall mapped to predicted PstI-MspI loci (using BBmap; Bushnell, 2014), however only 78.69% mapped to *in silico* loci in the targeted size range, presumably a result of non-specific fragment carryover in size selection. When sequencing depth is not considered, 79.86% of clustered loci align to valid PstI-MspI *in silico* loci, with only 58.96% mapping to loci expected within the target size range. However, when loci with low depth (<10 reads per locus) are removed, these numbers increase to 91.67% and 83.77% respectively, suggesting that inclusion of spurious fragments during library prep (e.g. those introduced via contamination) is a notable source of wasted sequencing effort. Additionally, when mapping raw sequences to the *Danio rerio* genome, we find disproportionate sequencing depth at some loci, likely reflecting inclusion of restriction sites within repetitive elements. Sequence effort must be adjusted to compensate and allow for exclusion of loci with insufficient (e.g., <10 reads/locus) or disproportionate read-depth. Insufficient coverage will reduce percentage of predicted loci actually being recovered *in vitro*.

To assess extensibility for non-model organisms, we also compared the number of *in vitro* sequenced loci to *in silico* predictions from reference genomes available at varying genetic divergences to 5 non-model organisms sequenced following a standard ddRAD protocol. Genetic distances were HKY corrected, and estimated using available mtDAN sequences from NCBI GenBank. As expected, a positive correlation between prediction error (i.e. difference in number of loci predicted per million bases) and genetic distance was the result (Fig 1), although correlation was notably weak (Pearson's $r = 0.325$; or 0.271 with removal of outlier). Factors other than genetic distance are likely more predictive of similarity in RAD locus recovery (e.g. genome architecture or assembly quality) and selection of optimal 'reference' genomes for *in silico* predictions may not always be reliable when considering genetic distance alone.

To quantify variance in locus recovery among individuals, we compared predicted locus yields for several ongoing studies with *in silico* estimates based on an appropriate within-family genome (if such a genome was available), corrected for artificially terminated fragments (e.g. resulting from a fragmentary assembly) by multiplying the number of these observed in the desired size range by the observed proportion of sequence-able fragments (e.g. PstI – MspI). Predicted locus yields were within 1 standard deviation of the mean observed recovery in all cases (see Table 1).

Conclusion

In silico digestion is useful for ddRAD projects in predicting the expected number of loci post-sequencing. The *a priori* estimates from FRAGMATIC may reduce wasted sequencing effort by optimizing size selection and enzyme choice. This is because coverage (i.e., sequencing depth) is substantially impacted by the number of simultaneously sequenced ddRAD loci. FRAGMATIC produces accurate predictions based on reference genomes that are relatively complete and closely related to the study species, however genetic distance (as estimated here) correlates only weakly with error in *in silico* estimates. A potential solution could involve a statistical correction based on genetic distances and genome sizes, however we suspect that additional aspects of genome architecture and quality of the reference assembly used are likely important in predicting the distribution of RAD sites. FRAGMATIC may be useful in exploring these effects by considering the placement of RAD sites relative to other genomic features such as repetitive elements.

FRAGMATIC can be easily applied to non-model organisms using an appropriate genome (e.g. most "phylogenetically-near" reference available), although a highly fragmentary input would potentially skew predictions. Based on the result of the *Danio rerio* sequencing, we suspect that spurious fragments represent a significant portion of wasted sequencing effort, and as such the bioinformatic treatment of the data post-sequencing should include stringent filtering loci by recovery depth to minimize the impact of fragments sequenced in error.

To minimize issues with coverage and repetitive content disproportionally affecting sequencing effort, the recommended usage of FRAGMATIC with non-model organisms and an appropriate reference genome (or ideally, multiple genomes) would involve: 1) Optimization of enzyme choice using *in silico* digests and inspection of FRAGMATIC output for locus yield and repetitive “peaks”; and 2) Comparison of candidate size selection ranges for the enzyme pair of choice to target a specific number of loci, scaling for differences in genome size and accounting for quality of genome assembly used. As aforementioned, a statistical correction involving genetic distance or some aspect of genome architecture may emerge in future study which could improve *in silico* estimates.

References

Andrews K.R., Good J.M., Miller M.R., Luikart G., Hohenlohe P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. advance on:81–92.

Baird N.A., Etter P.D., Atwood T.S., Currey M.C., Shiver A.L., Lewis Z.A., Selker E.U., Cresko W.A., Johnson E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 3:e3376.

DaCosta J.M., Sorenson M.D. 2014. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. PLoS One. 9:e106713.

Davey J.W., Hohenlohe P.A., Etter P.D., Boone J.Q., Catchen J.M., Blaxter M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12:499–510.

Eaton D.A.R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 30:1844–1849.

Edwards S. V., Shultz A.J., Campbell-Staton S.C. 2015. Next-generation sequencing and the expanding domain of phylogeography. Folia Zool. 64:187–206.

Gautier M., Gharbi K., Cezard T., Foucaud J., Kerdelhué C., Pudlo P., Cornuet J.-M., Estoup A. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol. Ecol. 22:3165–3178.

Heffelfinger C., Fragoso C.A., Moreno M.A., Overton J.D., Mottinger J.P., Zhao H., Tohme J., Dellaporta S.L. 2014. Flexible and scalable genotyping-by-sequencing strategies for population studies. BMC Genomics. 15:1–23.

Herrera S., Reyes-Herrera P.H., Shank T.M. 2015. Predicting RAD-seq marker numbers across the eukaryotic Tree of Life. Genome Biol. Evol. 7:3207–3225.

Howe K., Clark M., Torroja C., et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 496:498–503.

Kamps-Hughes N., Quimby A., Zhu Z., Johnson E. a. 2013. Massively parallel characterization of restriction endonucleases. Nucleic Acids Res. 41:1–8.

de Koning A.P.J., Gu W., Castoe T.A., Batzer M.A., Pollock D.D. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. PLoS Genet. 7:e1002384.

Lepais O., Weir J.T. 2014. SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. Mol. Ecol. Resour. 14:1314–1321.

Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. 2012. Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS One. 7:e37135.

Puritz J.B., Matz M. V., Toonen R.J., Weber J.N., Bolnick D.I., Bird C.E. 2014. Demystifying the RAD fad. Mol. Ecol. 23:5937–5942.

Rognes T., Flouri T., Nichols B., Quince C., Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 4:e2409v1.

Appendix

Table 1:Comparison of locus yield for 4 ongoing studies involving non-model organisms with the *in silico* estimates calculated as *s* + (*s/t)*m*, where *t* is the total number of observed sequence-able fragments (flanked by both restriction sites), *t* is the total number of observed fragments in the target size range, and *m* is the number of fragments which are artificially terminated by contig boundaries resulting from incomplete genome assembly.

Study taxon	**Reference taxon**	***N* samples**	**Size selection width**	**Predicted loci**	**Mean observed loci**	**σ^2 observed loci**
Crotalus viridis	*Crotalus mitchelli*	289	100bp	34527	34527	10625
Gila spp.	*Pimephales promelas*	236	100bp	28226	28226	8826
Rhinichthys spp.	*Pimephales promelas*	319	50bp	14475	14824	4330
Terrapene spp.	*Chrysemys picta*	116	55bp	38969	25555	15840

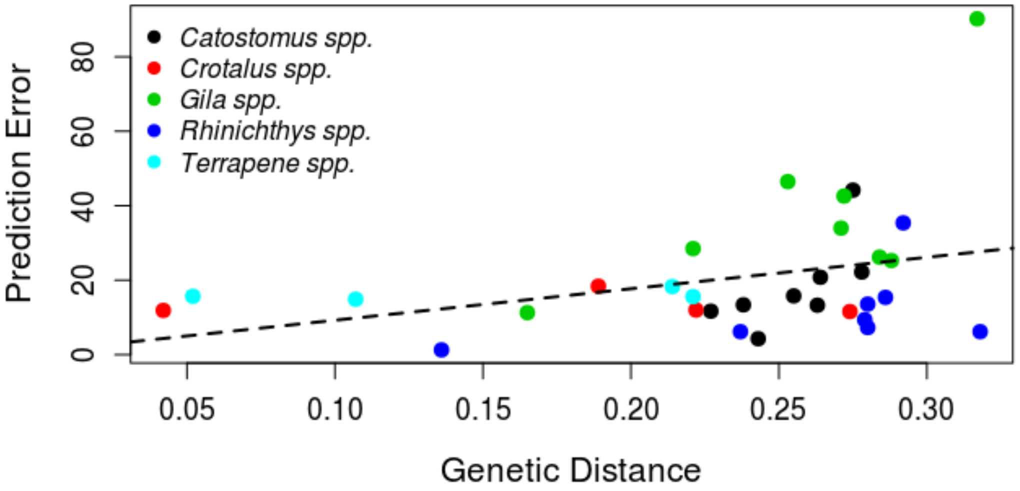

Figure 1: Relation between genetic distance and prediction error (the difference in number of average loci predicted vs observed, per megabase, based on a single sample of each sequenced *in vitro*). Linear model in R fitted as: *E* = 0.8159 + 84.501(*GD*), where *E* = predictive error and *GD*= mtDNA genetic distance

CHAPTER V

MRBAIT: Universal identification and design of targeted-enrichment capture probes

Chafin TK, Douglas MR, Douglas ME

Abstract

It is a non-trivial task to identify and design capture probes ("baits") for the diverse array of targeted-enrichment methods now available (e.g. ultra-conserved elements, anchored hybrid enrichment, RAD-capture). This often involves parsing large genomic alignments, followed by multiple steps of curating candidate genomic regions to optimize targeted information content (e.g. genetic variation), and to minimize potential probe dimerization and non-target enrichment. In this context, we developed MRBAIT, a user-friendly, generalized software pipeline for identification, design, and optimization of targeted-enrichment probes across a range of target-capture paradigms. MRBAIT is an open-source codebase that leverages native parallelization capabilities in Python and mitigates memory usage via a relational-database back-end. Numerous filtering methods allow comprehensive optimization of designed probes, including built-in functionality that employs BLAST, similarity-based clustering, and a graph-based algorithm that 'rescues' failed probes. Complete code for MRBAIT is available on GitHub (https://github.com/tkchafin/mrbait), and is also available with all dependencies via one-line installation using the conda package manager. Online documentation describing installation and runtime instructions can be found at: https://mrbait.readthedocs.io

Introduction

The application of next-generation sequencing methods to non-model organisms has been facilitated by a diverse array of novel 'reduced-representation' methods, whereby a consistent subset of the genome is targeted for sequencing across hundreds or thousands of individuals (Davey et al. 2011). One major trajectory for these methods is to target specific regions for sequencing, by utilizing the hybridization of oligonucleotide probes (or 'baits') to DNA fragments containing complementary sequences, followed by the subsequent separation of these target molecules (Mamanova et al. 2010). Although target-enrichment methods share this general design, numerous derivative methods have been developed and optimized for specific applications. For example, one commonly-applied paradigm is the enrichment of ultra-conserved genomic elements (UCEs), by identifying regions in divergent lineages with extremely low mutation accumulation, with the assay of genetic variation flanking these UCEs as the ultimate goal (e.g. Gnirke et al. 2009). Another popular approach is to specifically anchor probes to coding sequences (e.g. exon capture; Bi et al. 2012). Similarly, targeted fragmentation using restriction enzymes (per RADcap, Rapture) is also utilized, followed by a more specific reduction using capture probes (Ali et al. 2016; Hoffberg et al. 2016).

A universal requirement for these methods is that genomic resources be available *a priori*, or at least developed as a prerequisite to application, and from which probe sequences can then be designed. Transparent workflows are not always available (but see Faircloth, 2017 for such a treatment for UCEs), and are thus counter-productive to this endeavor. Some software does exist, but is often designed for a specific targeted-enrichment approach (Johnson et al. 2016; Faircloth 2017; Anil et al. 2018). One recently published option (BAITSTOOLS; Campana, 2017) is flexible enough to allow multiple inputs and enrichment schemes, yet does not natively incorporate post-processing steps to optimize bait-specificity. Here, we provide a flexible, user-friendly software, MRBAIT, that can be generalized to any targeted-enrichment paradigm. MRBAIT is not only open-source but also employs native Python parallelization. In addition, its memory usage, data management, portability, and iterative probe design are efficiently promoted through a relational database back-end using SQLITE.

Methods

Features and user interface

MRBAIT stores genomic regions or alignments, candidate target regions, and candidate probe sequences as an SQLITE relational-database with a Python wrapper and command-line interface (CLI). The database can be efficiently parsed then successively re-parsed, so as to allow fast exploration of numerous bait-design and filtering schemes. The general process is as follows:

(1) Build a consensus catalog of genomic regions by parsing alignments (as .xmfa, .maf, or .loci output of PYRAD) or genomes (as .fasta, annotated optionally with .vcf or .gff).

(2) Apply a sliding window along each consensus locus to find candidate target regions (depending on user specifications, e.g. indels allowed, frequency of flanking SNPs, etc.).

(3) Target filtering of regions (e.g. by GC content, maximum allowable pairwise identities, BLAST identity to potential contaminant genome), and resolve conflicts (if targets are within specified proximity along a scaffold or chromosome)

(4) Design a prospective bait set from passing target regions based on user-specified schema: tiling, or positional anchoring (e.g. centered or terminal within target region). If baits will be used for more distantly related taxa, polymorphism can be included to mitigate systematic bias in downstream molecular application

(5) Filtering and selection criteria (as in 3) are then applied to baits

(6) The pipeline can be resumed and any steps iteratively re-visited by providing the SQLITE database file (resulting in a significant reduction in runtime for successive runs)

Data are input in a variety of configurations: 1) Whole genomes (.fasta), with optional accompanying structural elements (as .gff) or variant information (.vcf); 2) multiple-genome alignment using the .maf output of MAFFT (Katoh and Standley 2013) or the .xmfa format of PROGRESSIVEMAUVE (Darling et al. 2010); or 3) reduced-representation alignments using the .loci format of PYRAD (Eaton 2014). Numerous filtering criteria are employed natively within MRBAIT and specified using the CLI, which allows target regions or designed probe sequences to be constrained in a variety of ways: with masking information from programs such as REPEATMASKER (Smit et al. 2013), via coordinates within a full genome to approximate all or a subset of specific genomic elements, by number of variant sites assayed (e.g. only retaining baits flanking known SNPs), or through other criteria (e.g. GC content, ambiguity or gap content). Targets or probes can be also filtered inclusively by optimizing specificity to a target genome, or exclusively by minimizing hits to a non-target (e.g. contaminant) genome using an internal call to NCBI-BLAST+ with a user-provided genome or database (Altschul et al. 1990). Probe-probe hybridization in downstream molecular application can also be circumvented using built-in clustering in MRBAIT via the VSEARCH algorithm (Rognes et al. 2016). Clustering results are used to build an undirected graph, with nodes as target regions (or baits), and edges representing pairwise alignments greater than some threshold identity and alignment length (user-provided). MRBAIT employs a naïve approach to identify the maximal independent set within this graph, optionally weighting nodes according to several user options so as to 'rescue' optimal targets without retaining edges. The motivation behind this approach is to retain a maximal number of baits without duplication. If undesired, this behavior can be easily disabled (or modified) using the CLI.

Benchmarking

Runtime and memory-usage were gauged using a ddRAD dataset generated for Whitetail deer (*Odocoileus virginianus*) from Arkansas. Samples (N=48) were digested with PstI and MspI restriction enzymes, size selected between ~375-525bp, and sequenced on an Illumina HiSeq 2500 with paired-end 150bp reads. Resulting data were assembled in pyRAD with 51,931 loci post-filtering. MRBAIT then processed these data. Requirements were as follows: A minimum per locus coverage of 25% for individuals; target regions with 1-10 flanking SNPs; and baits 60bp in length tiled across target regions at 1.5X coverage. These yielded 44,808 loci with a conserved region sufficient for bait design, with 27,102 candidate target regions flanking a sufficient number of SNPs. From these, a total of 43,342 baits were output in 392s across 4 threads on a 2014 iMac desktop. Identical runs with 1, 2 and 3 threads took 1182, 591, and 399 seconds, respectively, with a greater-than-linear speedup as core number increased. Peak memory usage increased sub-linearly with core count, at 120Mb for 1 thread and 300Mb for 4 threads on this dataset. For comparison, BAITSTOOLS (Campana 2017), with approximately comparable parameter settings, ran in 750 seconds using the 'SNP-targeting' strategy for PYRAD2BAITS (single-threaded) with no post-processing. The time discrepancy results from the initial setup of the relational database back-end (Step 1), which is the largest overhead for MRBAIT. Subsequent runs with re-parameterization for target selection, filtering, and bait design ran comparatively quickly. For example, the existing SQLITE database was passed to MRBAIT, with additional filtering on GC content for targets (between 0.3 and 0.7) and a new bait length of 80, in just 12 seconds (and resulted in 20,023 passed baits). This demonstrates the utility of the database approach in facilitating iterative probe design and exploring parameter values.

Comparison with existing methods

We parsed the existing Whitetail Deer ddRAD dataset so as to compare performance of bait design by MrBait versus BaitsTools and did so by maintaining maximum consistency in parameter settings between the two programs. We ran the 'pyrad2baits' program in BaitsTools with bait length of 80, a minimum of 20 individuals per locus, 50% overlap between tiled baits, and with baits containing gaps or ambiguous (N) characters excluded. These settings were replicated in MrBait, with no additional filtering to make comparison more appropriate. We also filtered the resulting bait sets by eliminating baits containing SNPs. This was accomplished natively within MrBait, and by using custom post-processing scripts for the BaitsTools output. The capacity of MrBait to filter targeted regions by 'informativeness' was not implemented, nor was BLAST-filtering for specificity. BaitsTools identified 14,276 non-variable bait sequences after manual post-processing in Python (successfully targeting 41.5% of the 20,912 loci with sufficient coverage), whereas MrBait found 12,084 baits, targeting 44% of loci. This demonstrates that both softwares can discover roughly equivalent sets of bait sequences, although in this case BaitsTools output required additional manual filtering while these steps were integrated in MrBait.

To compare accuracy of our bait design, we examined the data for 964 RAD loci from *Wisteria*, curated and assembled from paired-end sequencing data by Hoffberg et al. (2016). In parsing these loci, we excluded most of the native filtering methods in MrBait to keep results comparable. MrBait identified 1924 conservative 90-mer baits targeting all 964 loci, compared to the 1928 identified by Hoffberg et al. again indicating that MrBait will produce bait sets comparable to those from other existing methods.

However, users may find the additional utilities included natively in MrBait useful for reducing size of the total bait set, for example to improve specificity of the candidate baits (e.g. to reduce non-target enrichment), to reduce potential for ascertainment bias, or to reduce the overall number of sequences for synthesis (e.g. to meet budgetary requirements). For example, users may desire to remove baits which align to one another, as these can be non-specific to the intended locus (Faircloth 2017), or remove baits with extreme GC content which may show a phylogenetic bias when applied to broader taxa (Bossert et al. 2017). When applying a GC content filter (GC% >70 or <30), to the *Wisteria* dataset, 475 baits failed, while 25 failed when a conservative duplicate filter was applied (pairwise alignment of >80% identity over >80% of the bait length). Hoffberg et al. reported very high matrix occupancy with the designed bait set (99.8% of loci for 90% of samples, with a 4X coverage cutoff), however application of the uncurated bait set at a deeper phylogenetic scale could expose systematic bias associated with GC heterogeneity (e.g. Bossert et al. 2017), or with phylogenetic information content targeted by each bait, depending on the phylogenetic scale and intended method of downstream analysis (Meiklejohn et al. 2016). An additional major consideration is the potential for non-target capture from vastly different sources (e.g. bacterial contaminants), however extensive bioinformatic processing such as via native BLAST filtering in MrBait can significantly mitigate this (Bossert and Danforth 2018). Users are cautioned to consider any ascertainment biases which may be introduced, particularly when designing bait sets for a different phylogenetic scale than is available (e.g. as reference genomes) for bait design.

Conclusion

We provide a customizable and extensible open-source software (MRBAIT) that facilitates rapid and user-friendly bait development for an array of molecular applications (e.g. ultra-conserved elements, RAD-capture). It simultaneously identifies conservative 'target' regions in user-provided sequence data, designs probes to enrich them, and curates the resulting bait set. It also incorporates an array of native filtering strategies to help minimize downstream synthesis of problematic baits (e.g. duplicates), and to maximize specificity of baits to a target genome or desirable elements within them (e.g. known SNPs, or genomic features such as exons). MRBAIT adopts an SQL relational database back-end to minimize the problem of data files that necessitate high memory loads as well as significant I/O computational time. This allows users to rapidly re-parse the database with multiple different filtering criteria and promotes efficient exploration of parameter space and optimal bait sets for bait specificity and number (which affects synthesis cost). Comparisons with existing methods indicate that MRBAIT is similar in terms of quantity of targets discovered and runtime efficiency. Documentation and a full description of runtime options can be found at: https://mrbait.readthedocs.io

References

Ali O.A., O'Rourke S.M., Amish S.J., Meek M.H., Luikart G., Jeffres C., Miller M.R., Jeffres5 C., Miller M.R. 2016. RAD capture (Rapture): Flexible and efficient sequence-based genotyping. Genetics. 202:389–400.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–10.

Anil A., Spalinskas R., Åkerborg Ö., Sahlén P. 2018. HiCapTools: a software suite for probe design and proximity detection for targeted chromosome conformation capture applications. Bioinformatics. 34:675–677.

Bi K., Vanderpool D., Singhal S., Linderoth T., Moritz C., Good J.M. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics. 13.

Bossert S., Danforth B.N. 2018. On the universality of target-enrichment baits for phylogenomic research. Methods Ecol. Evol. 9:1453–1460.

Bossert S., Murray E.A., Blaimer B.B., Danforth B.N. 2017. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. Mol. Phylogenet. Evol. 111:149–157.

Campana M.G. 2017. BaitsTools: Software for hybridization capture bait design. Mol. Ecol. Resour.:1–6.

Darling A.E., Mau B., Perna N.T. 2010. Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 5.

Davey J.W., Hohenlohe P.A., Etter P.D., Boone J.Q., Catchen J.M., Blaxter M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12:499–510.

Eaton D.A.R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 30:1844–1849.

Faircloth B.C. 2017. Identifying conserved genomic elements and designing universal bait sets to enrich them. Methods Ecol. Evol.

Gnirke A., Melnikov A., Maguire J., et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27:182–189.

Hoffberg S.L., Kieran T.J., Catchen J.M., Devault A., Faircloth B.C., Mauricio R., Glenn T.C. 2016. RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. Mol. Ecol. Resour. 16:1264–1278.

Johnson M.G., Gardner E.M., Liu Y., Medina R., Goffinet B., Shaw A.J., Zerega N.J.C., Wickett N.J. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci. 4:1600016.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:727–744.

Mamanova L., Coffey A.J., Scott C.E., Kozarewa I., Turner E.H., Kumar A., Howard E., Shendure J., Turner D.J. 2010. Target-enrichment strategies for next-generation sequencing. Nat. Methods. 7:111–118.

McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. Genome Res. 22:746–754.

Meiklejohn K.A., Faircloth B.C., Glenn T.C., Kimball R.T., Braun E.L. 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements: Evidence for a bias in some multispecies coalescent methods. Syst. Biol. 65:612–627.

Rognes T., Flouri T., Nichols B., Quince C., Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 4:e2409v1.

Smit A., Hubley R., Green P. 2013. RepeatMasker 4.0. Seattle, WA Inst. Syst. Biol.

CHAPTER VI

FGTPARTITIONER: A rapid method for parsimonious delimitation of ancestry breakpoints in large genome-wide SNP datasets

Chafin TK

Software Description

Partitioning large (e.g. chromosomal) alignments into ancestry blocks is a common step in phylogenomic analyses (Springer and Gatesy 2018). However, current solutions require complicated analytical assumptions, or are difficult to implement due to excessive runtimes. Multiple approaches have been proposed for delimiting ancestry blocks in genomes (i.e. establishing recombination breakpoints), which generally fall into one of two categories: those which require dense or phased genotypic data (Liu et al. 2013); and those with complex analytical assumptions which require the definition of informative prior probability distributions and are computationally intensive (Dutheil et al. 2009). Both conditions are problematic for genome-scale studies of non-model species, where large-scale resequencing and phased reference data are unavailable, and genomes are often sequenced at low coverage.

I here describe a solution, FGTPARTITIONER, which is specifically designed for use with non-model genomic data without the need for high-quality phased reference data or dense population-scale sampling. FGTPARTITIONER delimits chromosome scale alignments using a fast interval-tree approach which detects pairwise variants which violate the four-gametes assumption (Hudson and Kaplan 1985), and rapidly resolves a most parsimonious set of recombination events to yield non-overlapping intervals which are both unambiguously defined and consistent regardless of processing order. These sub-alignments are then suitable for separate phylogenetic analysis, or as a 'first pass' which may facilitate parallel application of finer-resolution (yet more computationally intensive) methods.

After parsing user-inputs, the workflow of FGTPARTITIONER is as follows:

(1) For each SNP, perform four-gamete tests sequentially for rightward neighboring records, up to a maximal physical distance (if defined) and stopping when a conflict (='interval') is found. Intervals are stored in a self-balancing tree. When using multiprocessing, daughter processes are each provided an offset which guarantees a unique pairwise SNP comparison for each iteration

(2) Merge interval trees of daughter processes (if using optional parallel computation)

(3) Assign rank *k* per-interval, defined as the number of SNP records (indexed by position) spanned by each interval

(4) Order intervals by *k*; starting at min(*k*), resolve conflicts as follows: For each candidate recombination site (defined as the mid-point between SNPs), compute the depth *d* of spanning intervals. The most parsimonious breakpoint is that which maximizes *d*

These algorithm choices have several implications: indexing SNPs by physical position guarantees that the same recombination sites will be chosen given any arbitrary ordering of SNPs; and defining breakpoints as physical centerpoints between nodes means that monomorphic sites will be evenly divided on either side of a recombination event. Because monomorphic sites by definition lack phylogenetic information, they cannot be unambiguously assigned to any particular ancestry block, thus my solution is to evenly divide them. Heterozygous sites in diploid genomes are dealt with in multiple ways. By default, FGTPARTITIONER will randomly resolve haplotypes. The user can select an alternate resolution strategy which will either treat a SNP pair as failing if any resolution meets the four-gamete condition, or as passing if any possible resolution passes [i.e. the 'pessimistic' and 'optimistic' strategies (Wang et al. 2010)].

In conclusion, FGTPARTITIONER has several advantages over similar methods: 1) algorithmic and performance enhancements allow it to perform orders of magnitude faster, thus extending application to larger genomes; and 2) the flexibility of diploid resolution strategies precludes the need for haplotype phasing a priori. Validation using empirical data indicated the suitability of FGTPARTITIONER for highly distributed work on high-performance computing clusters, with parallelization easily facilitated by built-in options in the command-line interface. Additionally, runtime and memory profiling indicate its applicability on modern desktop workstations as well, when applied to moderately sized datasets. Thus, it provides an efficient and under-friendly solution to alignment pre-processing for phylogenomic studies, or as a method of breaking up large alignments in order to efficiently distribute computation for more rigorous recombination tests.

References

Dutheil J.Y., Ganapathy G., Hobolth A., Mailund T., Uyenoyama M.K., Schierup M.H. 2009. Ancestral population genomics: The coalescent hidden Markov model approach. Genetics. 183:259–274.

Hudson R.R., Kaplan N.L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 111:147–164.

Liu Y., Nyunoya T., Leng S., Belinsky S.A., Tesfaigzi Y., Bruse S. 2013. Softwares and methods for estimating genetic ancestry in human populations. Hum. Genomics. 7:1–7.

Springer M.S., Gatesy J. 2018. Delimiting coalescence genes (C-genes) in phylogenomic data sets. Genes. 9:1–19.

Wang J., Moore K.J., Zhang Q., de Villena F.P.-M., Wang W., McMillan L. 2010. Genome-wide compatible SNP intervals and their properties. Proc. First ACM Int. Conf. Bioinforma. Comput. Biol.:43–52.

CONCLUSION

A pervasive notion in the history of speciation research is the assumption that the solidification of reproductive boundaries between species is the focal benchmark in the evolutionary process. This is apparent in the pursuit of taxonomists to delimit discrete species, of phylogenetics to build bifurcating trees, and of evolutionary theory to define units of biodiversity. A shared aspect of these lofty ideals is the reverence with which reproductive isolation is held as a fundamental axis defining the 'speciation continuum'. This focus on reproductive isolation as the sole rate-limiting control on diversification is myopic, in that it relegates the role of gene flow to one that is primarily antagonistic to the proliferation of species [e.g. as a constraint on adaptive divergence (Futuyma 2010)].

Nature is rather more complex, with numerous controls on the accumulation of biodiversity over time (Dynesius and Jansson 2014; Rabosky 2016; Singhal et al. 2018; Harvey et al. 2019), to include pre- and post-speciation gene flow. In this dissertation I have demonstrated a variety of outcomes for post-speciation gene flow in species for which persistence in the Anthropocene is tenuous. I here discuss the implications of these outcomes with regards to hypothesized 'controls' of macroevolutionary patterns and how these contribute to iterative progress towards three primary trajectories in hybridization research (see Introduction).

Chapter I used a molecular assay—facilitated by the work presented in Chapter IV—to examine an empirical case wherein anthropogenic habitat alteration has been the driver for a local breakdown of reproductive isolation, resulting in either asymmetric introgression or hybrid swarm formation, depending on local context. In one case, a population facing intense genetic degradation was completely absent in more recent genetic surveys (Chafin et al. 2019), presumably as a result of continued admixture. The suggested mechanism involved a coupling between extrinsically-driven population decline and deleterious input of alleles, creating a negative fitness feedback, or ‘extinction vortex’ (Gilpin and Soulé 1986).

Yet, hybridization is also often implicated as a mechanism promoting evolutionary rescue (Stelkens et al. 2014; Fitzpatrick et al. 2016). I suggest the discrepancy lies in context-dependent details, and the interplay between rates of environmental change, gene flow, and reproductive isolation. Firstly, the probability of rescue is contingent on a sufficiently low rate of environmental change to allow adaptation (Lindsey et al. 2013), with gene flow having either a dampening or amplifying effect depending on the degree of extrinsic reproductive isolation. Secondly, although hybridization might supply beneficial adaptive variants under certain circumstances, it may also lead to a weakening of reproductive isolation and eventual assimilation of one species into another (Owens and Samuk 2020). I suggest that the case seen in *Gila robusta* and *G. cypha* (Chapter I) represents the early stages of this outcome. Studying cases such as *Gila*, where it is possible to directly sample a temporal transect through the *active* modulation of population persistence by hybridization, is a necessary step in understanding how outcomes play at the macroevolutionary scale.

In Chapter II I examined the outcomes of hybridization in *Gila* over such timescales and used a combination of phylogenetic approaches to quantify the degree to which hybridization (as opposed to alternate sources) drove patterns of discordance therein. I also substantiated methods by which empirical divergence in the ‘anomaly zone’ (see Degnan and Rosenberg 2009) can be detected. The implications for the study of hybridization at the macroevolutionary scale are two-fold. First, I used a scalable method for detecting phylogenetic reticulation in large-scale datasets by using low-resolution methods (e.g. D-statistic) as a ‘first pass’ to generate hypotheses, followed by computationally expensive network inference (see Solís-Lemus and Ané 2016) with massively parallel gene tree inference on HPC (high performance computing) systems. Secondly, I provided user-friendly means to explicitly assign the cause of 'star-like' phylogenies to divergence in the anomaly zone, where rapid speciation generates high discordance—as opposed to weak resolution caused by low differentiation. The latter is important because it aids in the unbiased discrimination of intraspecific lineages from poorly resolved interspecific lineages, while the former is a critical framework for inferring large-scale phylogenetic networks.

Finally, I used a novel method described in Chapter VI to delimit ancestry blocks in the genome of the endangered red wolf in order to understand how local genome structure biases retention of introgressed alleles. I showed how recombination interacts with selection to create 'refugia' on the X-chromosome which retained a species tree pattern reflecting more ancient divergence despite a genomic mosaic reflecting hybridity. This confirmed expectations of the so-called 'large-X' effect (Coyne and Orr 1989) and also served to reject an hypothesized hybrid origin of the red wolf in favor of one in which secondary introgression masked signatures of prior isolation. I then showed how this process of 'ancestry swamping' in the autosomes misled prior analyses (aided in part by a method described in Chapter V).

Together, these chapters create a framework to discriminate phylogenetic patterns of hybridization in a scalable manner (Chapters I, II, VI, V), and to further categorize different outcomes of hybridization (Chapters III and VI). Moving forward, my research offers a blueprint to synthesize modern trajectories in hybridization research into a broader, comparative fabric. First, in order to assess rates of hybridization, and the respective fates of hybrid lineages, a targeted molecular approach is required (e.g. as facilitated by Chapters IV and V) that circumvents cost limitations of sequencing full genomes of non-model organisms. Second, hybridization must be categorized into components that distinguish ancient *versus* contemporary (Chapter I and III), and generative *versus* introgressive (Chapter III). Finally, large-scale networks (e.g. hundreds or thousands of taxa) can be constructed using the approach of Chapter II, wherein hybridization was explicitly separated from alternative sources of phylogenetic variation in a massively parallel computational pipeline. Because this framework facilitates the broad-scale quantification of hybridization (e.g. across large complete clades), it directly contributes to future work aiming to test the role of hybridization in mediating macroevolutionary patterns of diversification and trait evolution (e.g. Maddison et al. 2007; Fitzjohn 2010; Rabosky 2014; Beaulieu and O'Meara 2016; Bastide et al. 2018; Harvey and Rabosky 2018). This approach can best be contrasted with phylogenetic comparative methods for identifying correlations between species traits and diversification; for example, if hybridization generally promotes adaptive variation (as in Meier et al. 2017), then lineages having an increased rate of hybridization should show either lower extinction rates or higher speciation rates. Herein lies the general framework from which the role of hybridization as a large-scale driver of biodiversity can be parsed (e.g. Folk et al. 2018).

References

Bastide P., Solís-Lemus C., Kriebel R., Sparks K.W., Ané C. 2018. Phylogenetic comparative methods on phylogenetic networks with reticulations. Syst. Biol. 67(5):800–820.

Beaulieu J.M., O'Meara B.C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst. Biol. 65:583–601.

Chafin T.K., Douglas M.R., Martin B.T., Douglas M.E. 2019. Hybridization drives genetic erosion in sympatric desert fishes of western North America. Heredity. 123:759–773.

Coyne J.A., Orr H.A. 1989. Patterns of speciation in Drosophila. Evolution. 43:362–381.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Dynesius M., Jansson R. 2014. Persistence of within-species lineages: A neglected control of speciation rates. Evolution. 68:923–934.

Fitzjohn R.G. 2010. Quantitative traits and diversification. Syst. Biol. 59:619–633.

Fitzpatrick S.W., Gerberich J.C., Angeloni L.M., et al. 2016. Gene flow from an adaptively divergent source causes rescue through genetic and demographic factors in two wild populations of Trinidadian guppies. Evol. Appl. 9:879–891.

Folk R.A., Soltis P.S., Soltis D.E., Guralnick R. 2018. New prospects in the detection and comparative analysis of hybridization in the tree of life. Am. J. Bot. 105:364–375.

Futuyma D.J. 2010. Evolutionary constraint and ecological consequences. Evolution. 64:1865–1884.

Gilpin M.E., Soulé M.E. 1986. Minimum viable populations: Processes of species extinction. In: Soulé M.E., editor. Conservation Biology: The Science of Scarcity and Diversity. Sunderland, MA: Sinauer. p. 19–34.

Harvey E., Harvey M.G., Singhal S., Rabosky D.L. 2019. Beyond reproductive isolation : Demographic controls on the speciation process. Ann. Rev. Ecol. Evol. Sys. 50:75-95.

Harvey M.G., Rabosky D.L. 2018. Continuous traits and speciation rates: Alternatives to state-dependent diversification models. Methods Ecol. Evol. 9:984–993.

Lindsey H.A., Gallie J., Taylor S., Kerr B. 2013. Evolutionary rescue from extinction is contingent on a lower rate of environmental change. Nature. 494:463–467.

Maddison W.P., Midford P.E., Otto S.P. 2007. Estimating a binary character's effect on speciation and extinction. Syst. Biol. 56:701–710.

Meier J.I., Marques D.A., Mwaiko S., Wagner C.E., Excoffier L., Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat. Commun. 8:14363.

Owens G.L., Samuk K. 2020. Adaptive introgression during environmental change can weaken reproductive isolation. Nat. Clim. Chang. 10:58–62.

Rabosky D.L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PLoS One. 9.

Rabosky D.L. 2016. Reproductive isolation and the causes of speciation rate variation in nature. Biol. J. Linn. Soc. 118:13–25.

Singhal S., Huang H., Grundler M.R., Marchán-Rivadeneira M.R., Holmes I., Title P.O., Donnellan S.C., Rabosky D.L. 2018. Does population structure predict the rate of speciation? A comparative test across Australia's most diverse vertebrate radiation. Am. Nat. 192:432–447.

Solís-Lemus C., Ané C. 2016. Inferring phylogenetic networks with Maximum Pseudolikelihood under incomplete lineage sorting. PLoS Genet. 12:1–21.

Stelkens R.B., Brockhurst M.A., Hurst G.D.D., Greig D. 2014. Hybridization facilitates evolutionary rescue. Evol. Appl. 7:1209–1217.

Appendix – Supplementary Tables and Figures

Table S1: Sampling Locations and Drainages for *N*=386 *Gila* Individuals and Outgroups.

Site	Major Drainage	Location	*N*
Gila atraria	Bonneville, Snake	Multiple localities	14
Gila cypha			
C1	Yampa	Yampa Canyon, Moffat Co., CO	1
C2	Colorado	Black Rocks & Westwater Canyons, Colorado	2
C3	Colorado	Grand Canyon, Coconino Co., AZ	9
C4	Little Colorado	Little Colorado R. mainstem, Coconino Co., AZ	4
Gila ditaenia	de la Concepcion	Rio Magdalena, Sonora, MX	2
Gila elegans	Hatchery	SNARRC, Dexter, NM	11
Gila eremica	Rio Sonora	Rio Sonora, Sonora, MX	2
Gila intermedia			
I1	Verde	Spring Creek, Yavapai Co., AZ	10
I2	Agua Fria	Silver Creek, Yavapai Co., AZ	8
I3	Agua Fria	Indian Creek, Yavapai Co., AZ	9
I4	Agua Fria	Sycamore Creek, Yavapai Co., AZ	10
I5	Gila	Mineral Creek, Pinal Co., AZ	10
I6	Gila	Eagle Creek, Greenlee Co., AZ	10
I7	Gila	Hardin-Cienega Creek, Greenlee Co., AZ	10
I8	San Pedro	Turkey Creek, Santa Cruz Co., AZ	10
Gila jordani	Pahranagat/ White	Key-Pittman WMA refuge population (brood stock	21
Gila minacae	Rio Yaqui	Rio Bavispe, Chihuahua, MX	7
Gila nigra			
N1	Verde	Fossil Creek, Yavapai Co., AZ	11
N2	Verde	Weber Creek, Gila Co., AZ	8
N3	Verde	Verde River headwaters, Gila Co., AZ	5
N4	Salt	Gordon Creek, Gila Co., AZ	10
N5	Salt	Gun Creek, Gila Co., AZ	11
N6	Salt	Marsh Creek, Gila Co., AZ	10
Gila nigrescens	Multiple	Multiple localities	5
Gila pandora	Rio Grande	Palomas Creek, Sierra Co., NM	6
Gila pulchra	Rio Yaqui	Rio Tomochic, Chihuahua, MX	5
Gila purpurea	Rio Yaqui	San Bernadino NWR, Cochise Co., AZ	2
Gila robusta			
R1†	Bill Williams	Trout Creek, Mohave Co., AZ	9
R2†	Bill Williams	Francis Creek, Yavapai Co., AZ	8
R3†	Verde	Verde River mainstem, Yavapai Co., AZ	8
R4†	Verde	Verde River mainstem, Yavapai Co., AZ	4
R5†	Verde	Verde River mainstem, Yavapai Co., AZ	3
R6†	Verde	West Clear Creek, Yavapai Co., AZ	5
R7†	Salt	Salt River mainstem, Maricopa Co., AZ	5
R8†	Salt	Cherry Creek, Gila Co., AZ	8
R9†	San Pedro	Aravaipa Creek, Pinal Co., AZ	10
R10‡	Green	Upper Green River tributaries, Wyoming	21
R11‡	Yampa	Little Snake River, Wyoming	5
R12‡	Yampa	Upper Yampa River tributaries, Moffat Co., CO	5
R13‡	Green	San Rafael River, Utah	4
R14‡	Colorado	Upper Colorado River mainstem, Colorado	4
R15‡	San Juan	Navajo River, Colorado	2
R16‡	Little Colorado	East Clear Creek, Coconino Co., AZ	16
V1	Virgin	Muddy (Moapa) River, Clark Co., NV	19
V2	Virgin	Virgin River mainstem, Washington Co., UT	17
Ptychocheilus			
P. grandis	Eel River	South Fork Eel River, Humboldt Co., CA	2
P. lucius	Colorado	Yampa River, Moffat Co., Colorado	8

†Lower Colorado River basin (below Grand Canyon) ‡Upper Colorado River basin (above Grand Canyon)

Table S2: Ancestry proportions of the red wolf genome at varying filtering thresholds and metrics of inclusion. Values are reported as proportion of bases, with proportion of blocks in parentheses. Note that here, 'Unassigned' reflects regions in which neither ancestries could be assigned (e.g. red wolf sister to outgroup).

Filtering criterion	Coyote	Gray Wolf	Heterozygous	Unassigned
No filters	0.252 (0.204)	0.232 (0.365)	0.337 (0.365)	0.179 (0.214)
Exc. unassigned blocks	0.311 (0.262)	0.417 (0.470)	0.272 (0.268)	-
+ 0.9 > int.het < 0.1	0.426 (0.310)	0.376 (0.341)	0.198 (0.349)	-
+ bp.RELL < 0.10	0.511 (0.390)	0.280 (0.284)	0.208 (0.326)	-
Exc. heterozygous	0.437 (0.377)	0.386 (0.415)	-	0.177 (0.207)
+ bp.RELL < 0.10	0.431(0.264)	0.234 (0.185)	-	0.335 (0.550)
Exc. unassigned and heterozygous	0.531 (0.476)	0.469 (0.524)	-	-
+ bp.RELL < 0.10	0.648 (0.588)	0.352 (0.412)	-	-

Table S3: Mutation-scaled and absolute divergence time estimates from g-PhoCS. Parameters are as follows: population divergence time for red wolf and coyote (τ_{COY}); divergence time for red wolf and gray wolf (τ_{WOLF}); divergence time for all *Canis* species (τ_{CANIS}); and the divergence time for the root (τ_{ALL}). Values shown are the raw arithmetic mean estimates, with calibrated estimates in years in parentheses, assuming a generation time of three years and an average per-generation mutation rate of μ=4x10^{-9} / base pair.

Analysis	τ_{COY}	τ_{WOLF}	τ_{CANIS}	τ_{ALL}
Coyote subset	3.8x10^{-5} (28,500)	-	7.3x10^{-4} (547,500)	2.5x10^{-3} (1,875,000)
Gray wolf subset	-	5.9x10^{-6} (4,425)	7.6x10^{-4} (570,000)	2.9x10^{-3} (2,175,000)

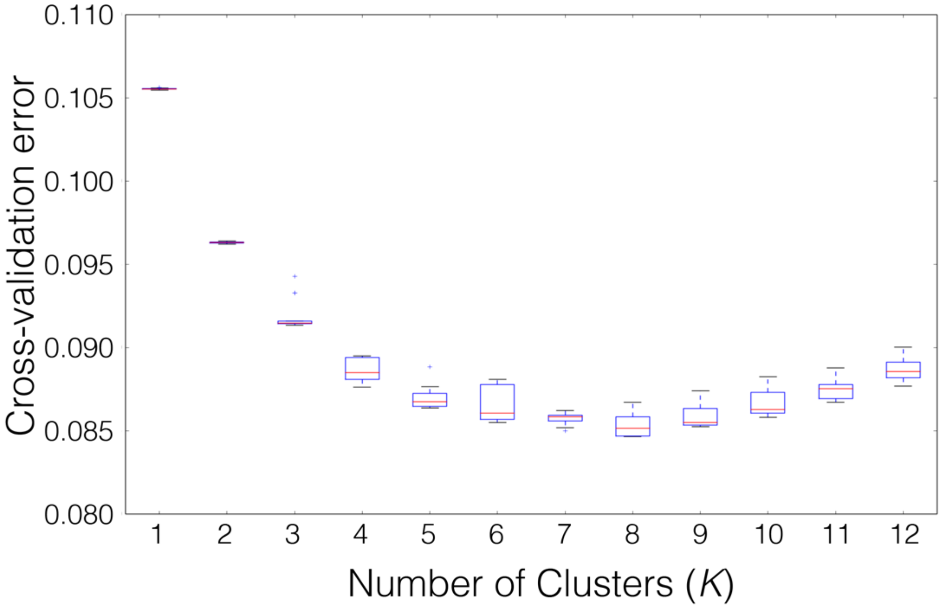

Figure S1: Cross-validation error analysis for ADMIXTURE, as represented by number of clusters (*K*)

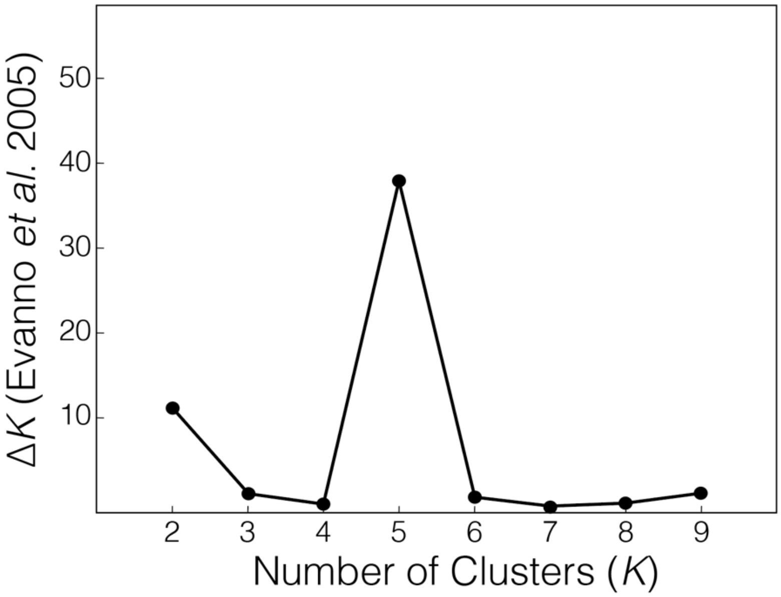

Figure S2: Change in model likelihood by *K* derived from a STRUCTURE analysis.

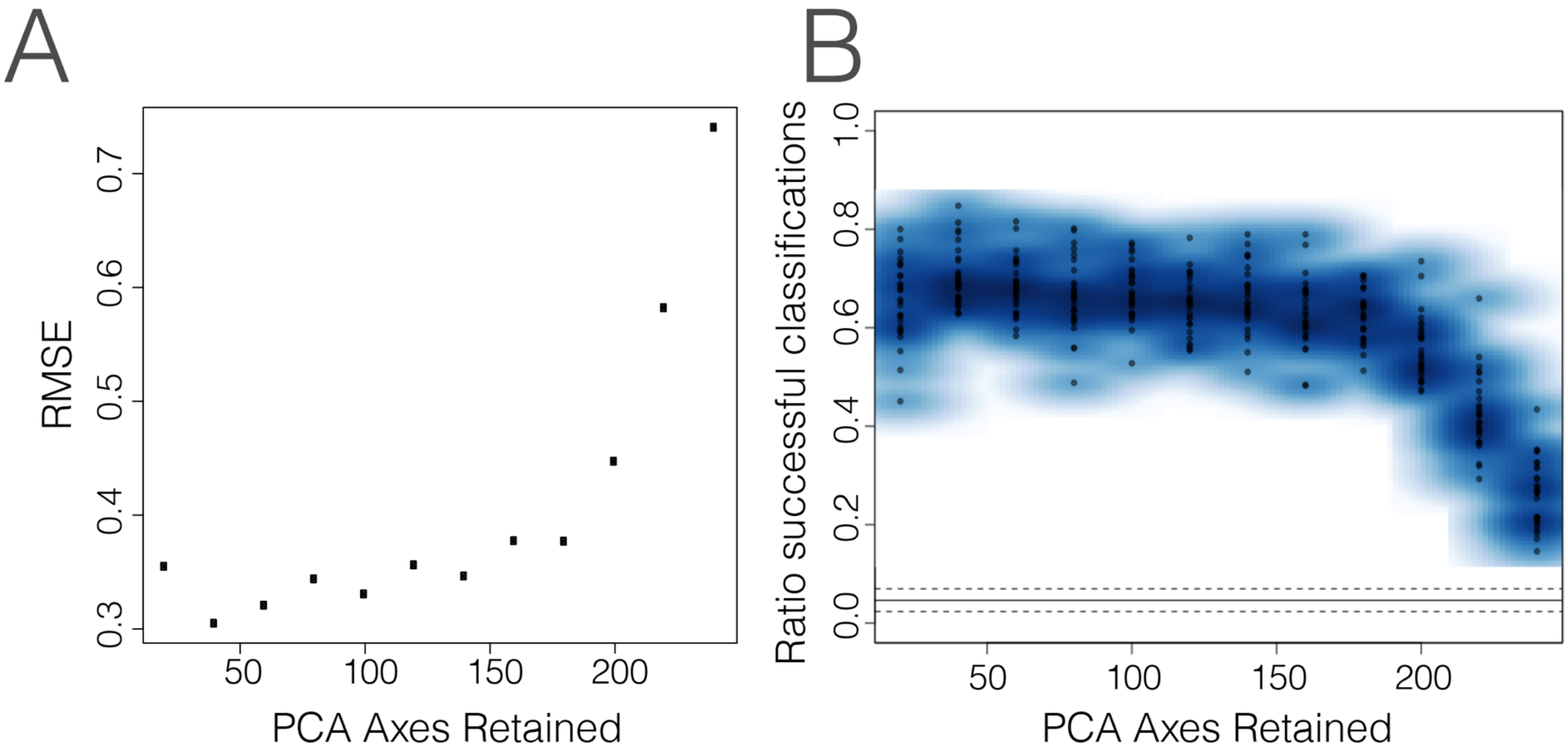

Figure S3: Results of cross-validation analysis in Discriminant Analysis of Principal Components (DAPC), depicting (A) root-mean-square error (RMSE) for classifications under varying number of Principal Component axes (PC's) retained, and (B) proportion of successful classifications for 20 replicates with varying number of PC's retained.

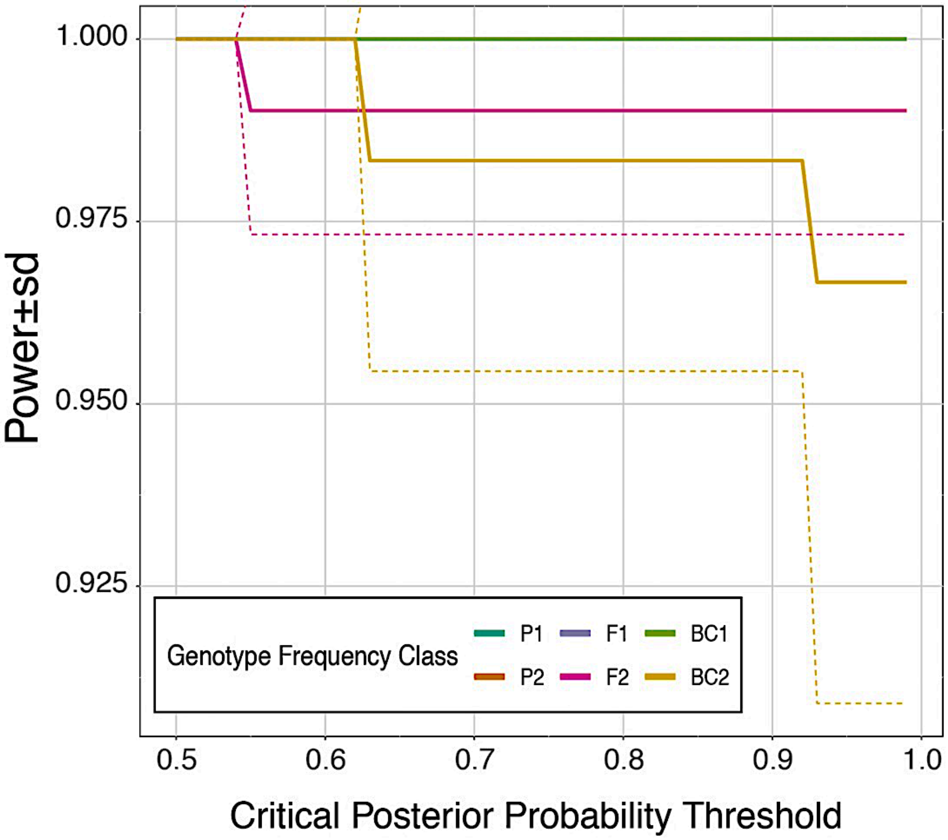

Figure S4: Plot of statistical power of assignment for each of six genotype frequency classes. NewHybrids of simulated hybrid genotypes in HybridDetective, using various critical thresholds (from 0.5 to 1.0) for posterior assignment probabilities. Solid lines indicate mean power, while dashed lines are standard deviations across replicates. Note that accuracy of assignment is 100% in all cases (not shown).

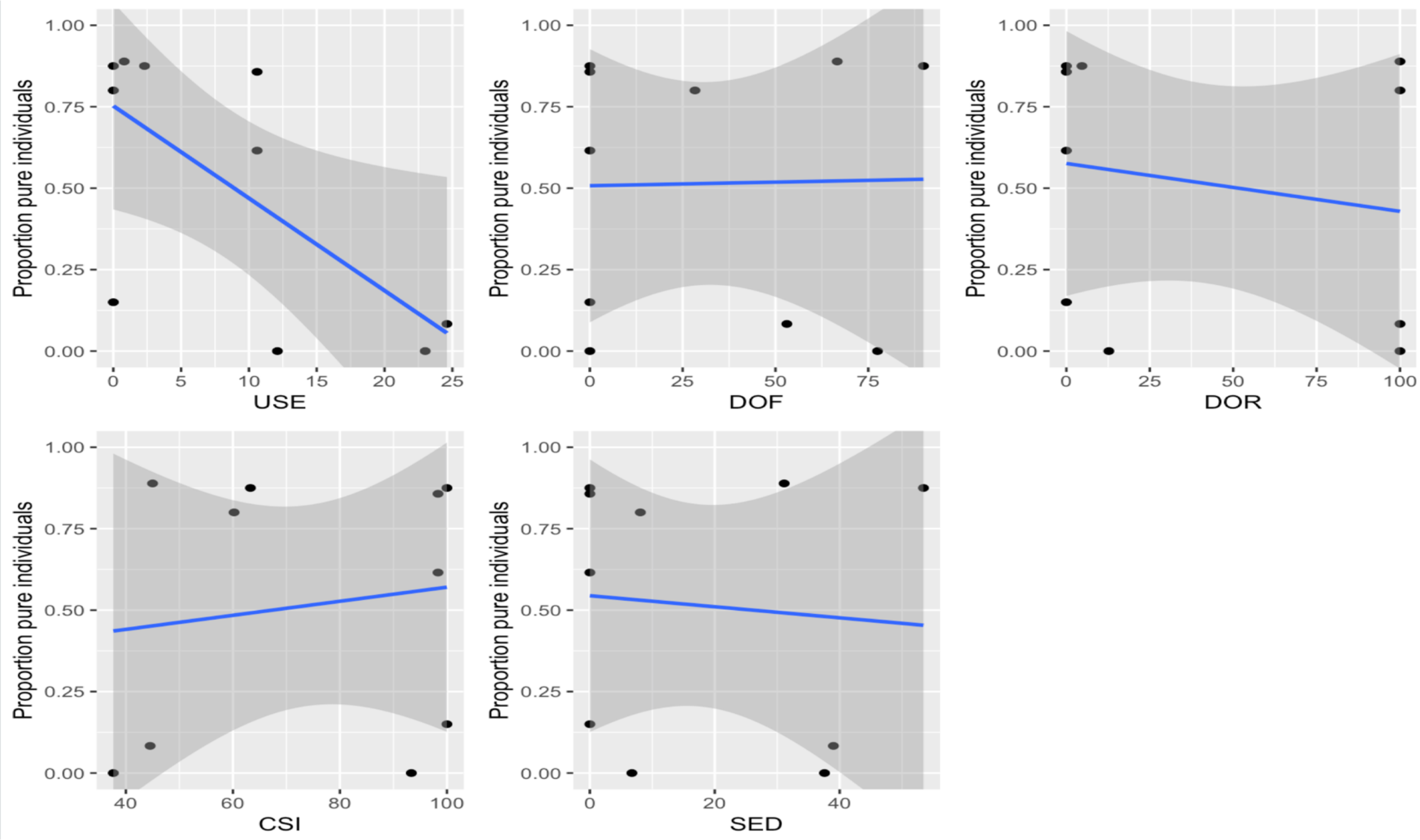

Figure S5: Relationships between five indices of anthropogenic pressure on stream reaches, and the proportion of genetically pure individuals sampled therein. DOF=Degree of fragmentation; DOR=degree of regulation; SED=degree of sediment trapping; USE=percent consumptive water use; CSI=connectivity status index. Scales for DOF, DOR, SED, and USE reflect a proportion of effect, with 100 = 100% impact. CSI scales from 0 to 100, with 100 being full connectivity (i.e. not detectable human impact).

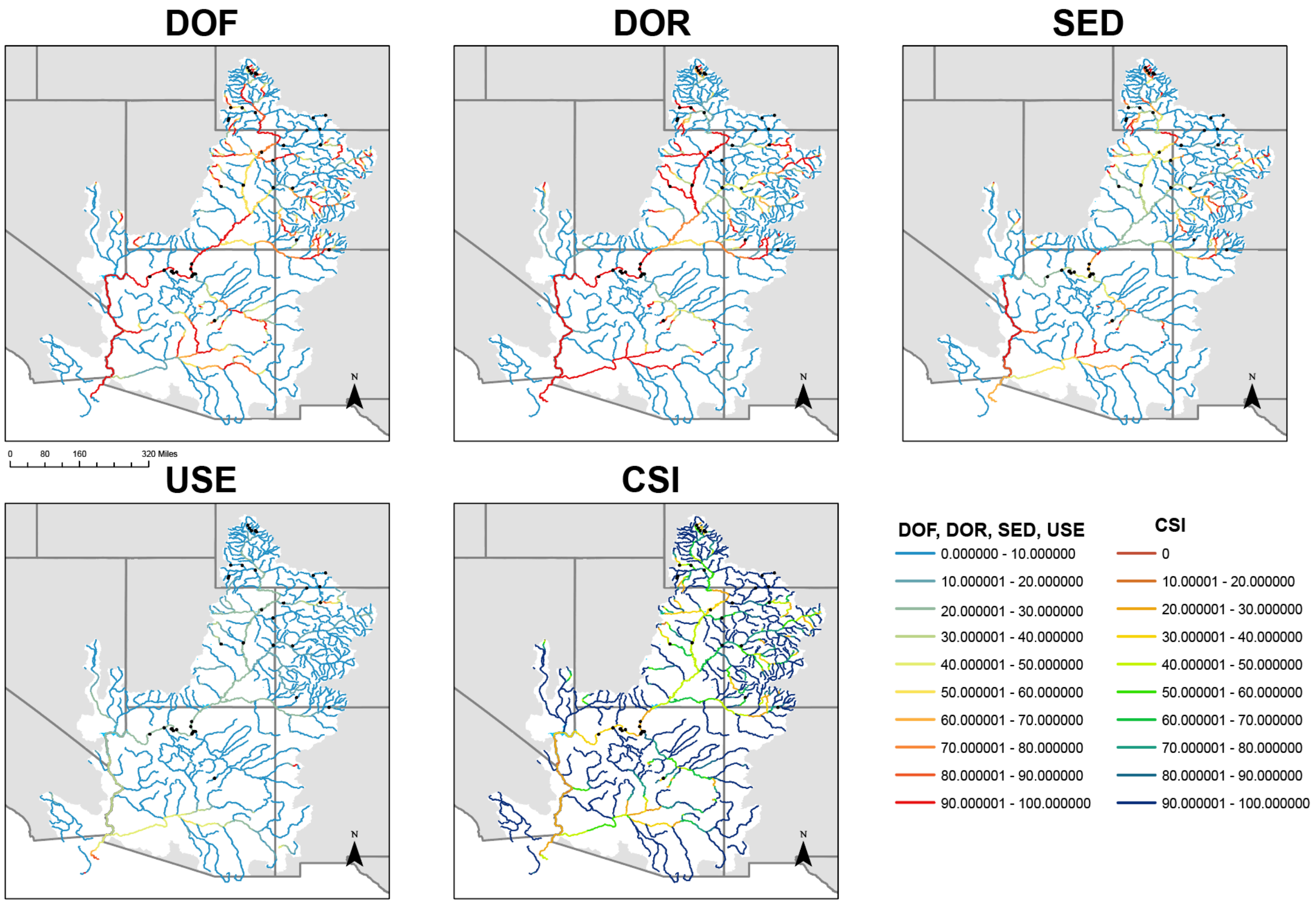

Figure S6: Human pressure indices plotted onto stream reaches in the Colorado River. DOF=Degree of fragmentation; DOR=degree of regulation; SED=degree of sediment trapping; USE=percent consumptive water use; CSI=connectivity status index. Scales for DOF, DOR, SED, and USE reflect a proportion of effect, with 100 = 100% impact. CSI scales from 0 to 100, with 100 being full connectivity (i.e. not detectable human impact).

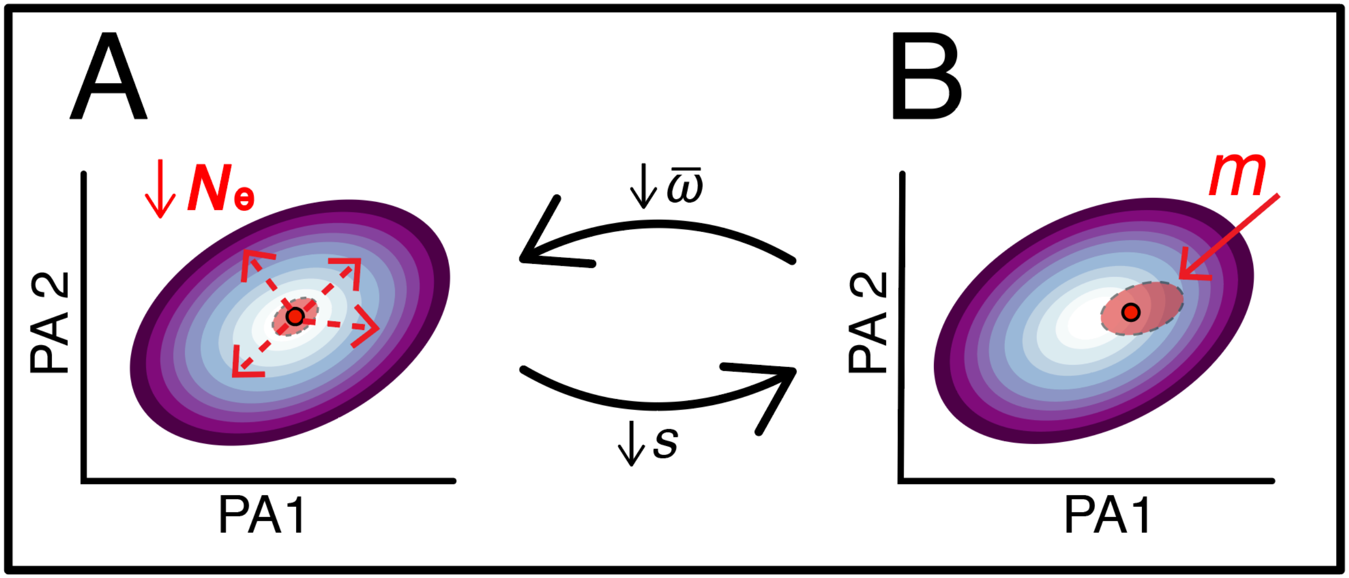

Figure S7: Extinction vortex via negative feedback of inbreeding and outbreeding depression. Shown is the population mean fitness ($\overline{\omega}$; red dot) and variance (red ellipse) within a fitness gradient from low (purple) to high (white), as a function of two arbitrary phenotypic axes (PA1, PA2). (A) Decreasing effective populations size (N_e) lowers the strength of selection (s) and reduces both $\overline{\omega}$ and genetic variance (=inbreeding depression). (B) When introgression is maladaptive, increased gene flow (m) bolsters maladaptive genetic variance while driving a further reduction in $\overline{\omega}$ (=outbreeding depression). This, in turn, prompts further depreciation in N_e, with weakened purifying selection and a relatively greater influence of maladaptive m as a consequence. Persistent coupling of these processes can then drive population extirpation, especially given extrinsic effects on $\overline{\omega}$, such as rapid environmental change.

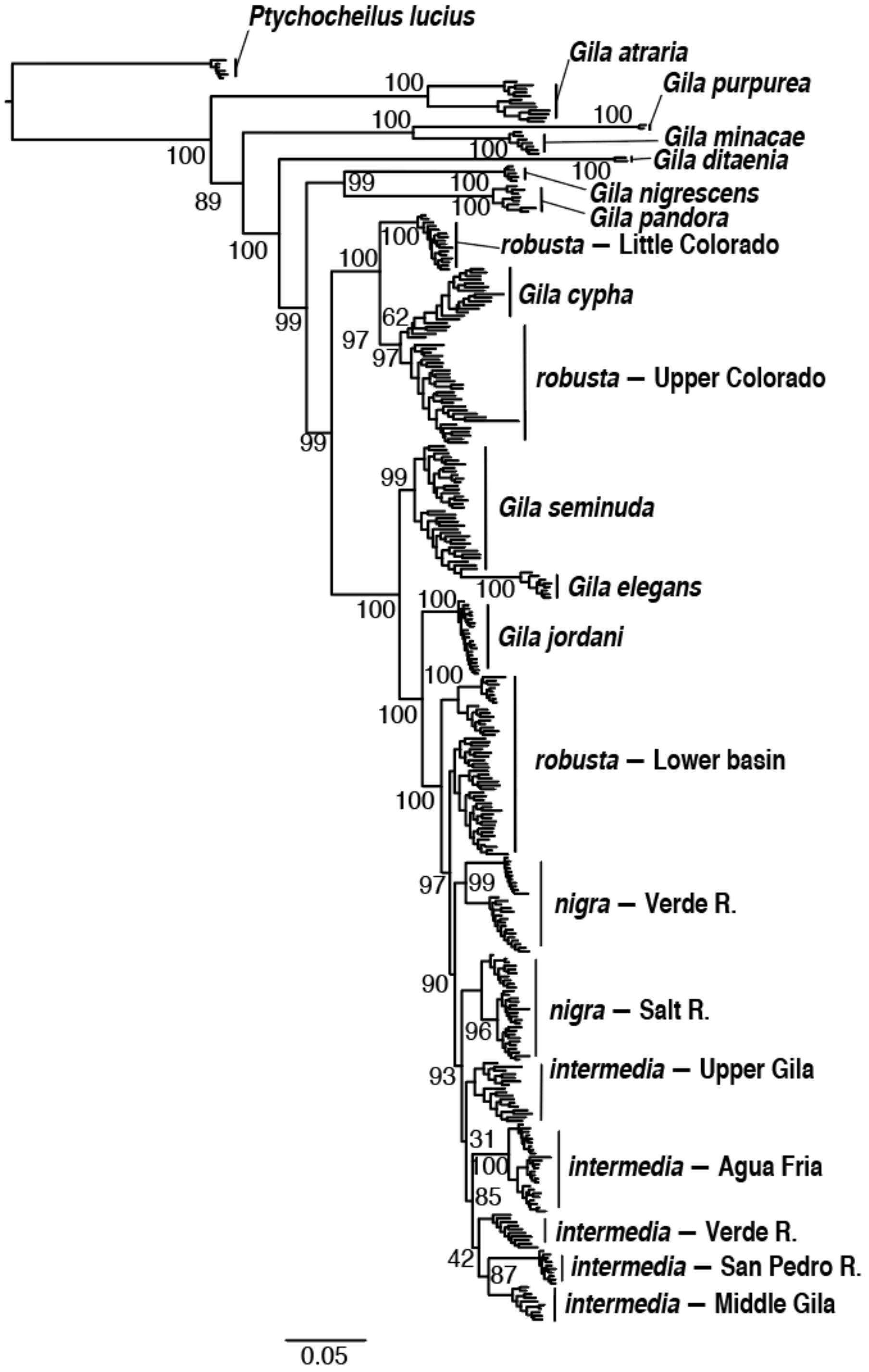

Figure S8: Phylogram showing results from an unconstrained search using 21,717 concatenated SNPs in IQ-TREE. Focal nodes are annotated with bootstrap support (values for shallow nodes omitted for clarity). For specific locality information, refer to Table S1.

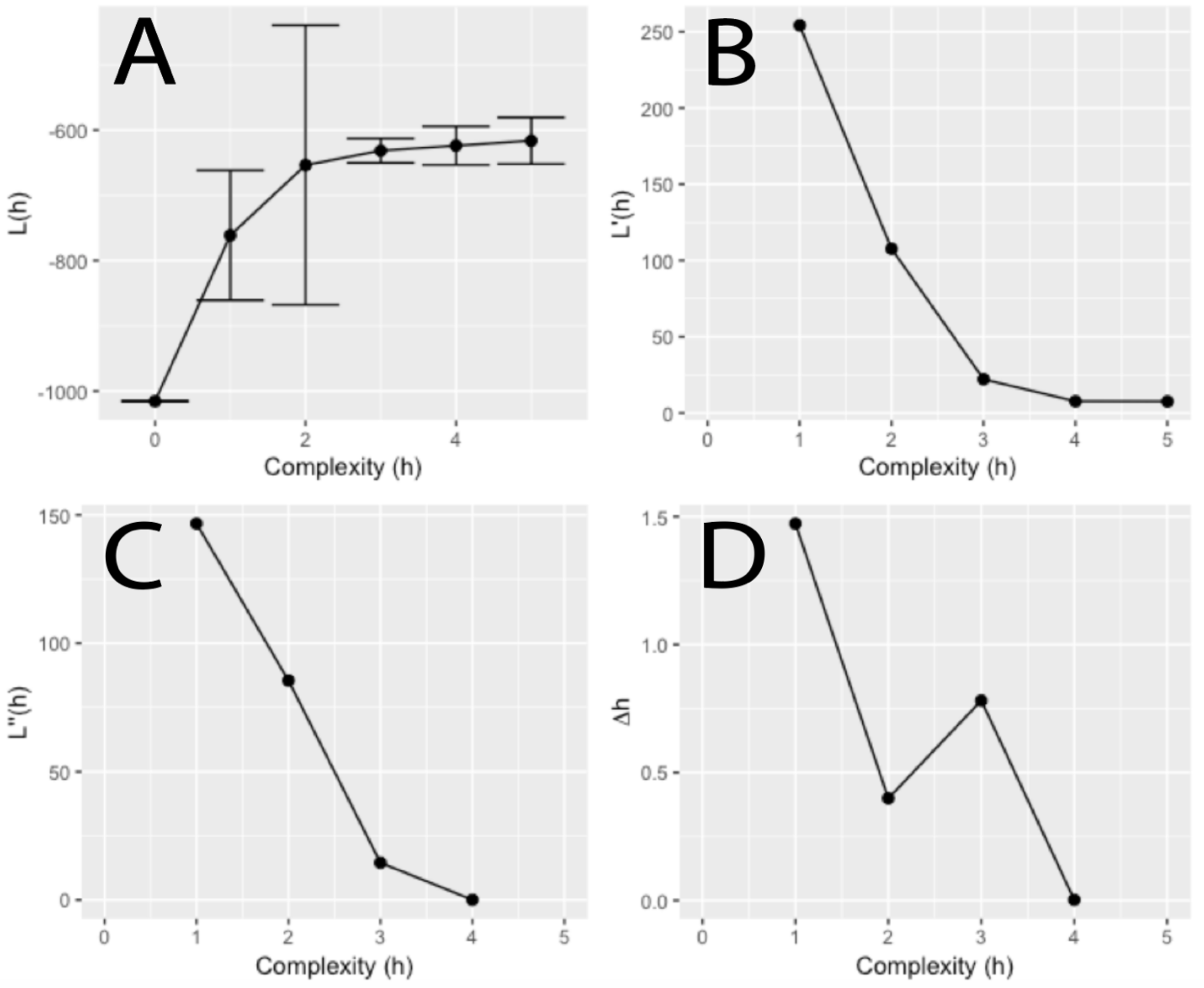

Figure S9: Model selection results for SNaQ/ PhyloNetworks; *h*=maximum number of hybrid edges allowed per model; (A) L(*h*) = -log likelihood for the best network of *N*=48 replicate runs per value of *h*; (B) L'(*h*) = 1st order change in L(*h*) = L(*h*) – L(*h*-1); (C) L''(*h*) = 2nd order change in L(*h*) = L'(*h*+1) – L'(*h*); and (D) Δ*h* = L''(*h*) / s(*h*) where s(*h*) is the standard deviation in L(*h*) among replicates.

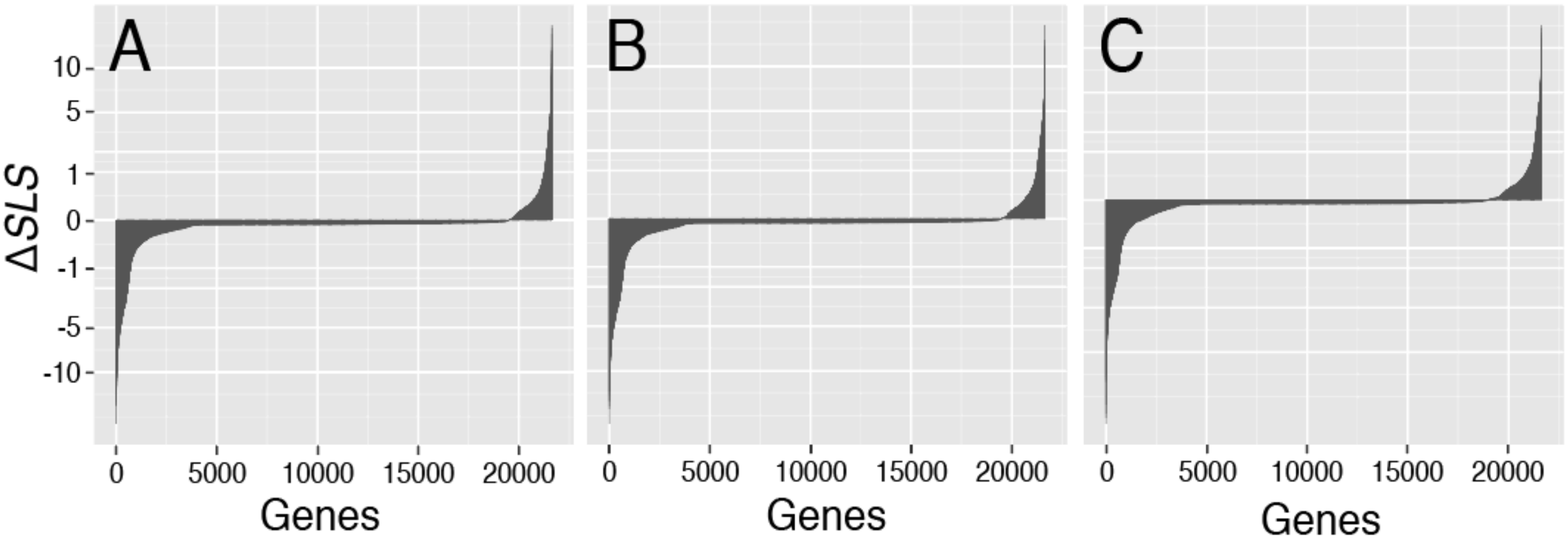

Figure S10: Site-wise log-likelihood differences (Δ*SLS*) for (A) SVDQuartets, (B) PoMo, and (C) TICR topologies as compared to an unconstrained concatenated tree. Δ*SLS* values are transformed as signed square-roots, with positive values indicating increased site-likelihood under the constrained model, and negative values having increased likelihood under the unconstrained concatenated model.

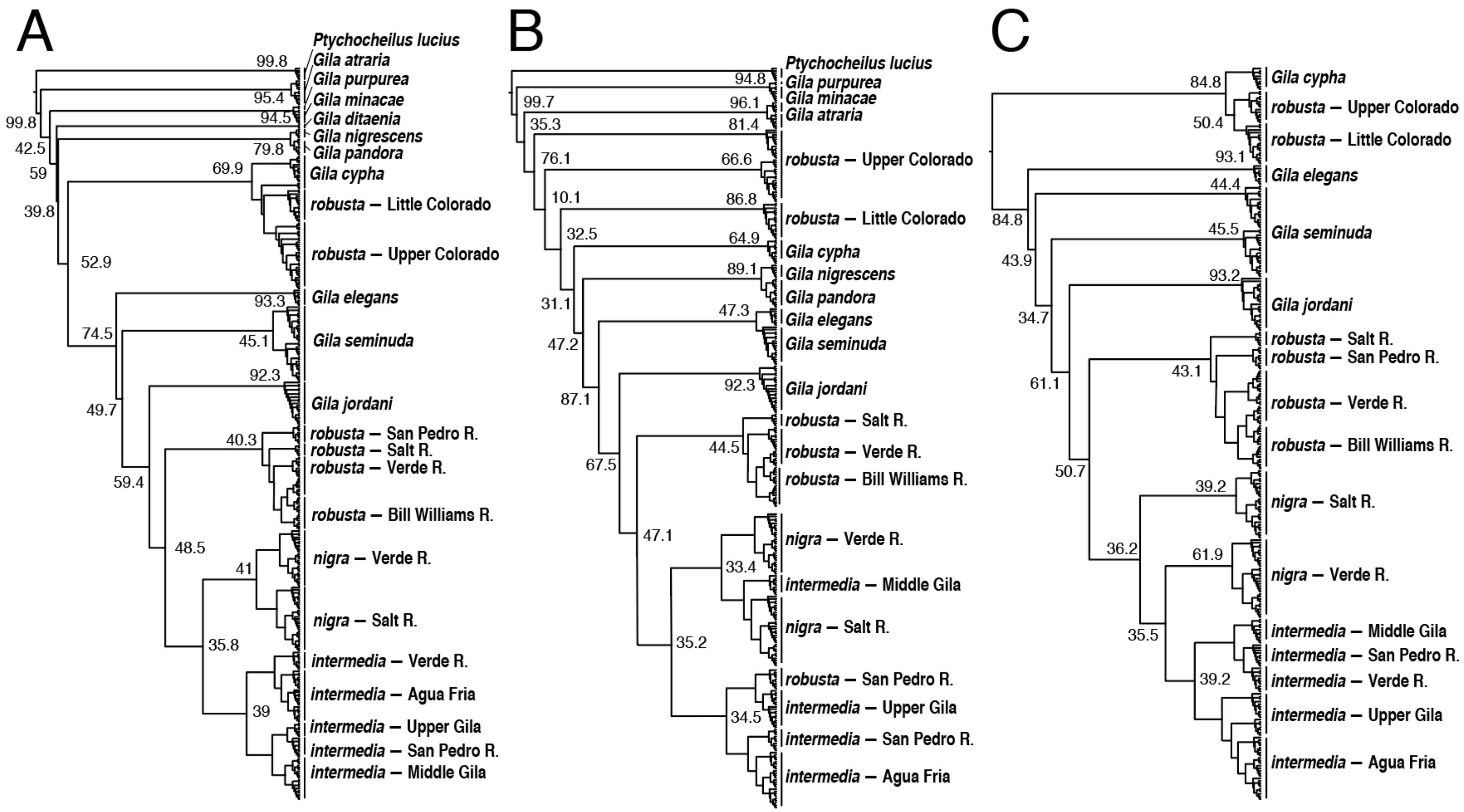

Figure S11: Site-wise concordance factors (s*CF*) for lineage trees produced in IQ-TREE under topological constraints for the (A) SVDQUARTETS, (B) POMO, and (C) TICR results. For details, see Methods.

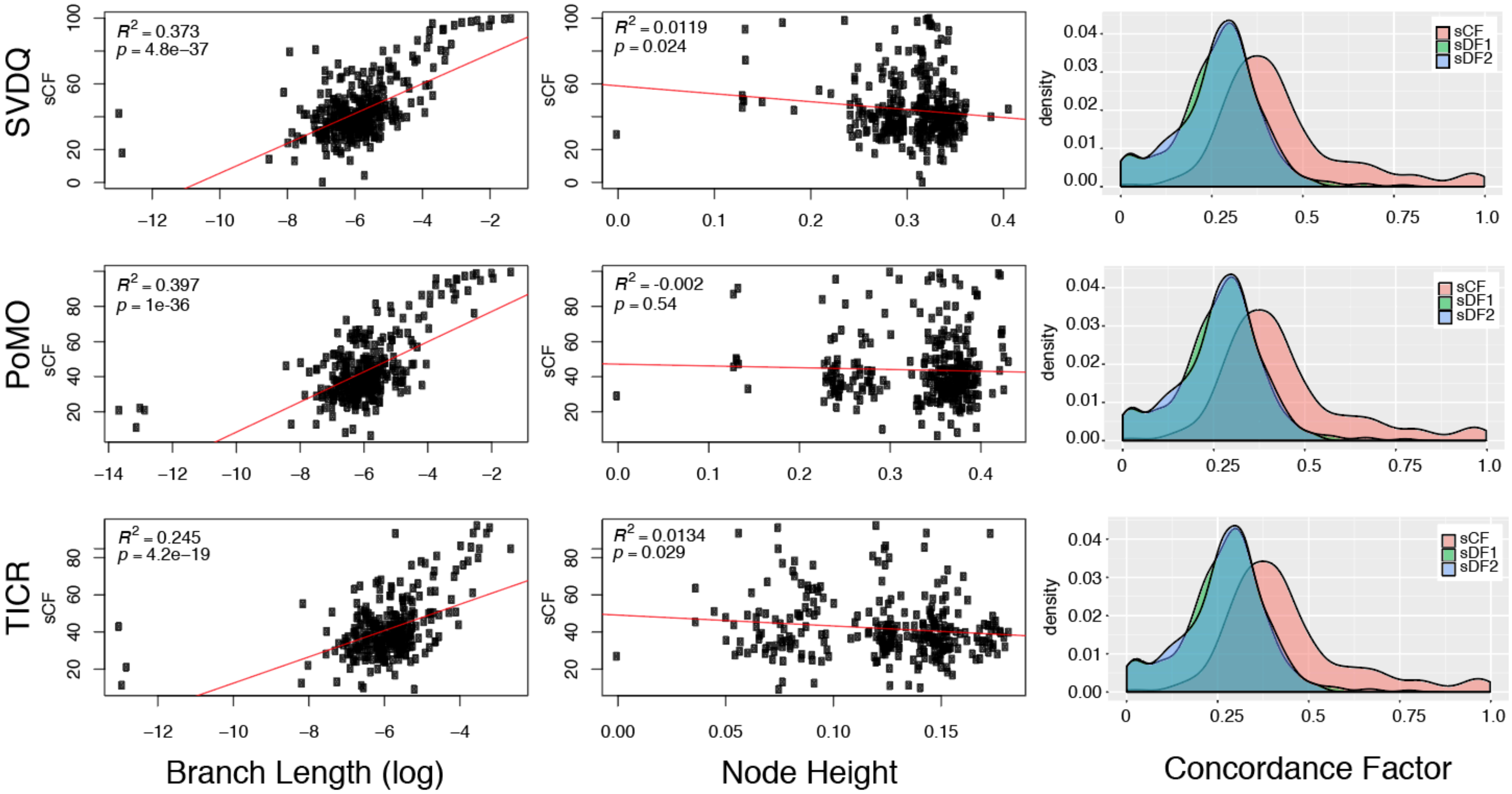

Figure S12: Characterization of site-wise concordance (*s*CF) factors for SVDQUARTETS, POMO, and TICR phylogenies. Panels show (left to right): Linear regression of subtending branch lengths (log-transformed) with *s*CF; node height (cumulative branch lengths from root to focal node); and densities of *s*CF across nodes as compared to the discordance factors for the two conflicting quartet resolutions (*s*DF1 and *s*DF2).

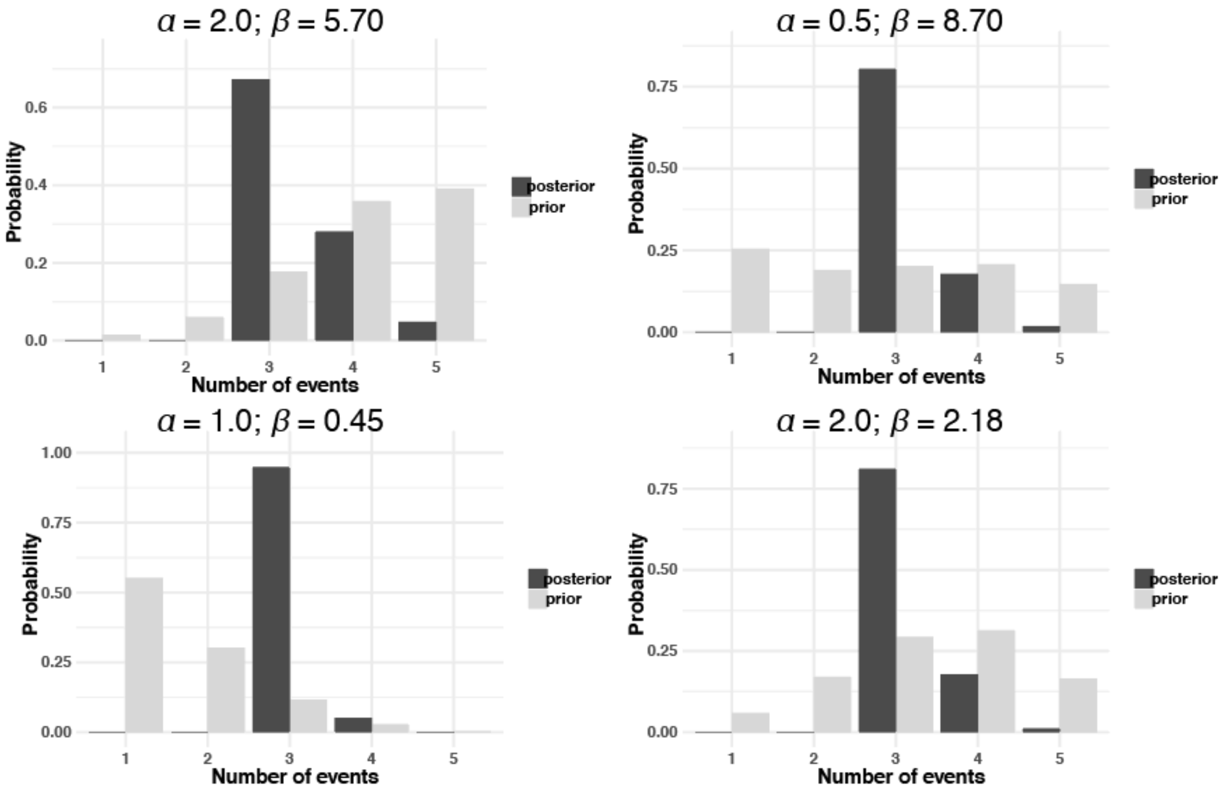

Figure S13: Prior and posterior probabilities for number of independent divergence events in ECOEVOLITY co-divergence models for *Gila*. Parameters across all runs were identical, except for the shape (α) and scale (β) of the gamma-distributed prior on the Dirichlet process concentration.

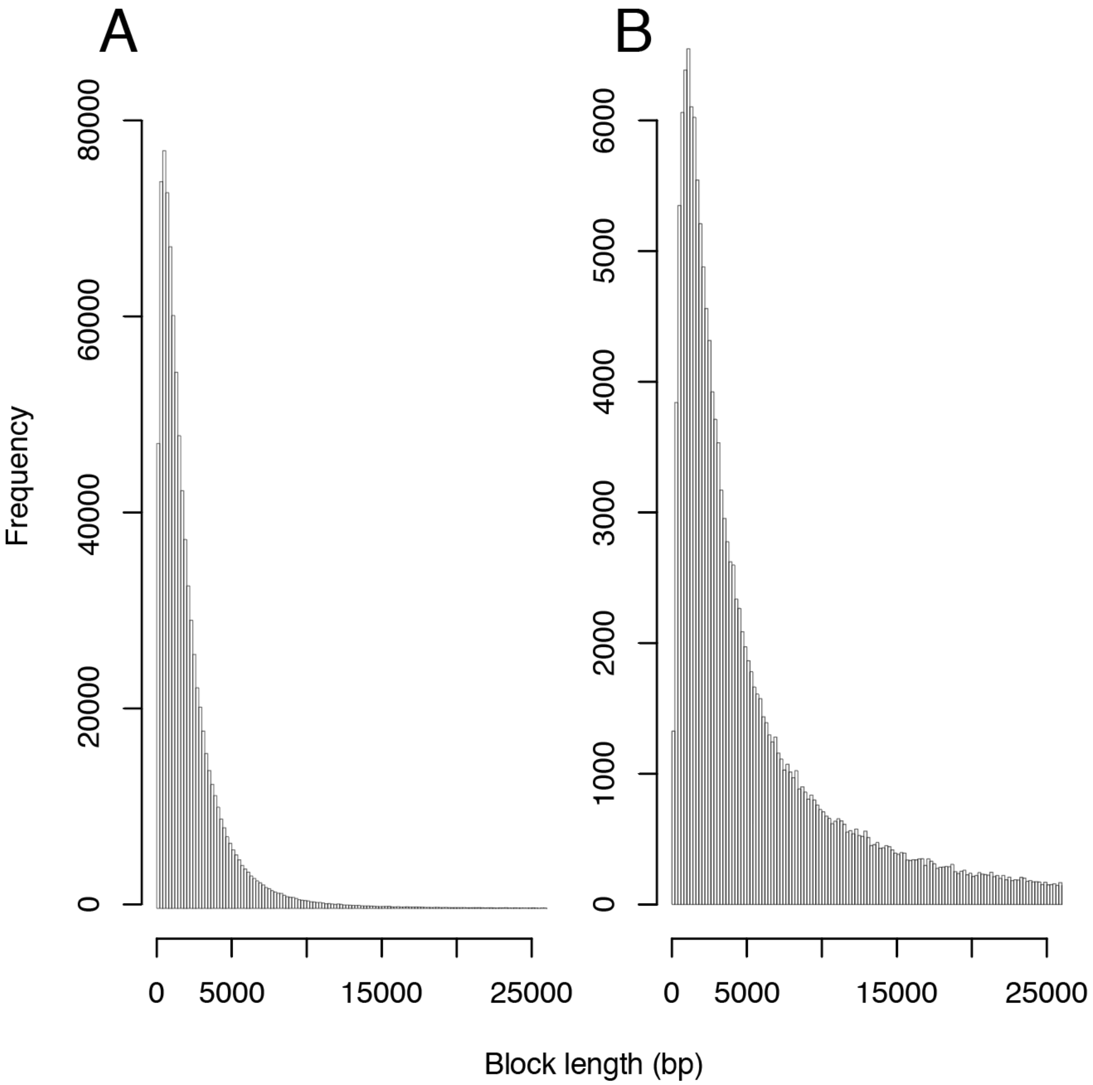

Figure S14: Ancestry block lengths in the red wolf genome before (A) and after (B) merging consecutive blocks of the same ancestry. Note truncation of the x-axis for interpretability.

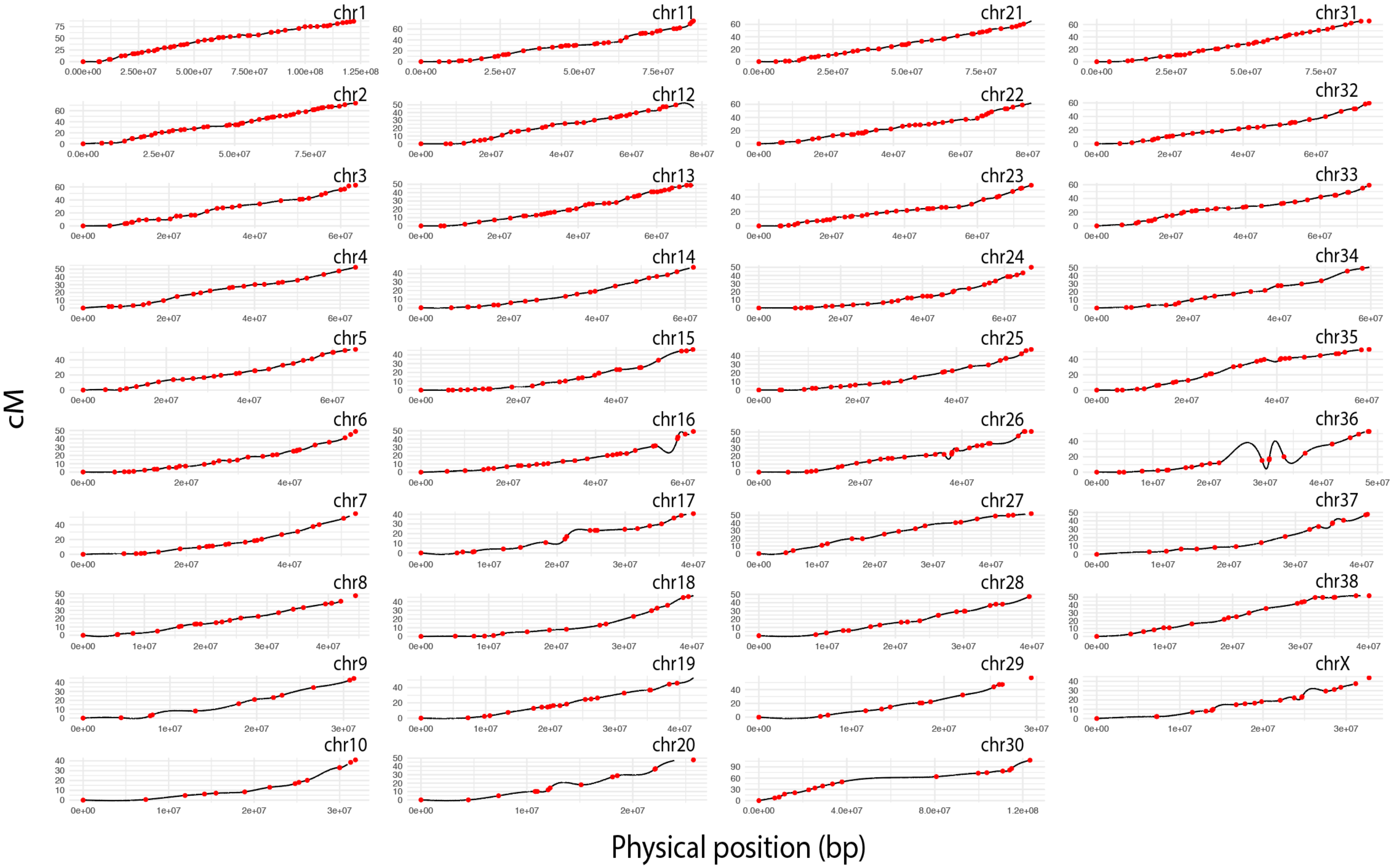

Figure S15: Cubic interpolation models for red wolf chromosomes (black) with points depicting data points from the genetic map (red)

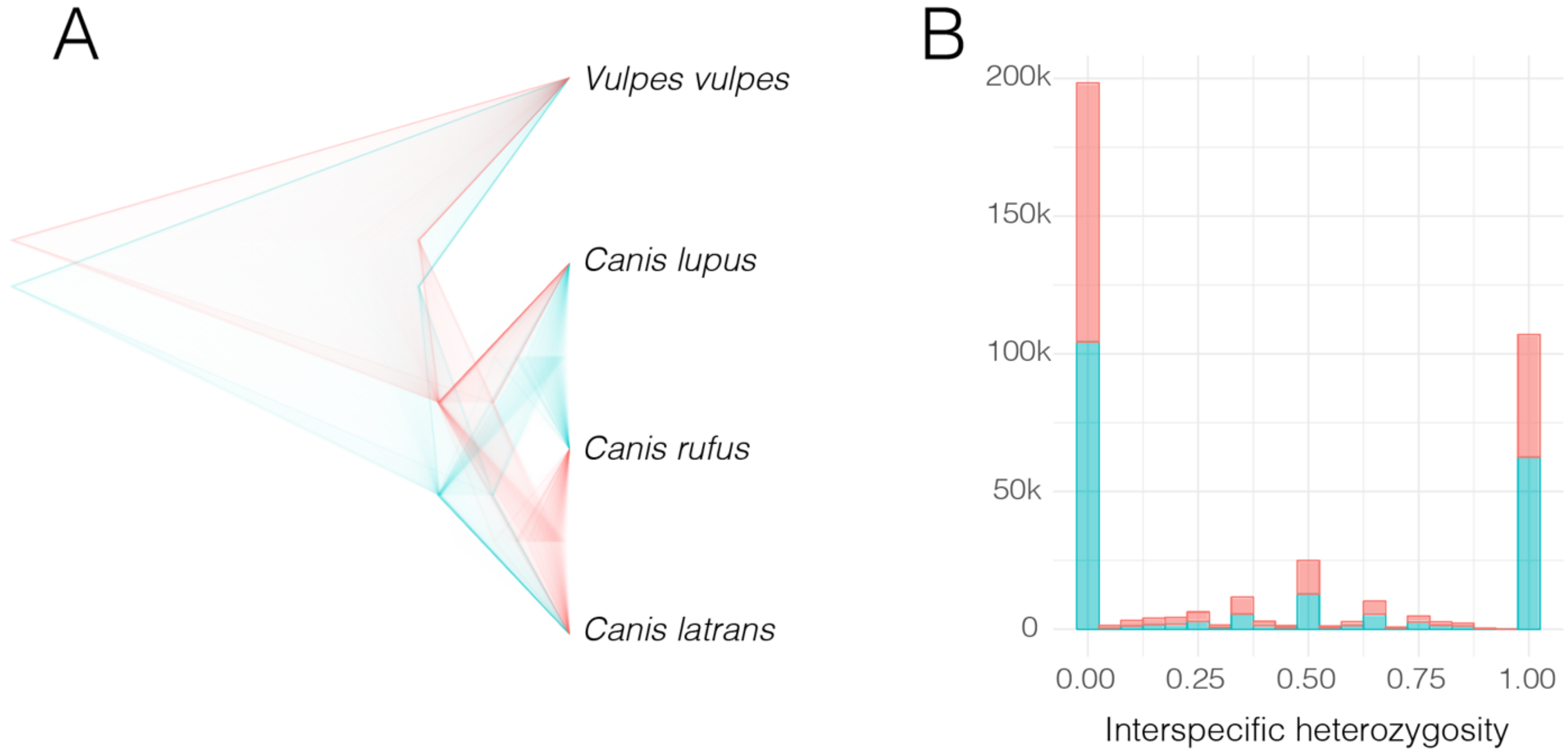

Figure S16: Densitree plot (A) of gene trees and distribution of interspecific heterozygosity (B) among sampled genomic ancestry blocks in the red wolf genome. Gene trees are restricted to those which were significantly supported by approximated bootstrap proportions (e.g. <10% of trees supporting an alternate topology), whereas interspecific heterozygosity is reported for all blocks.

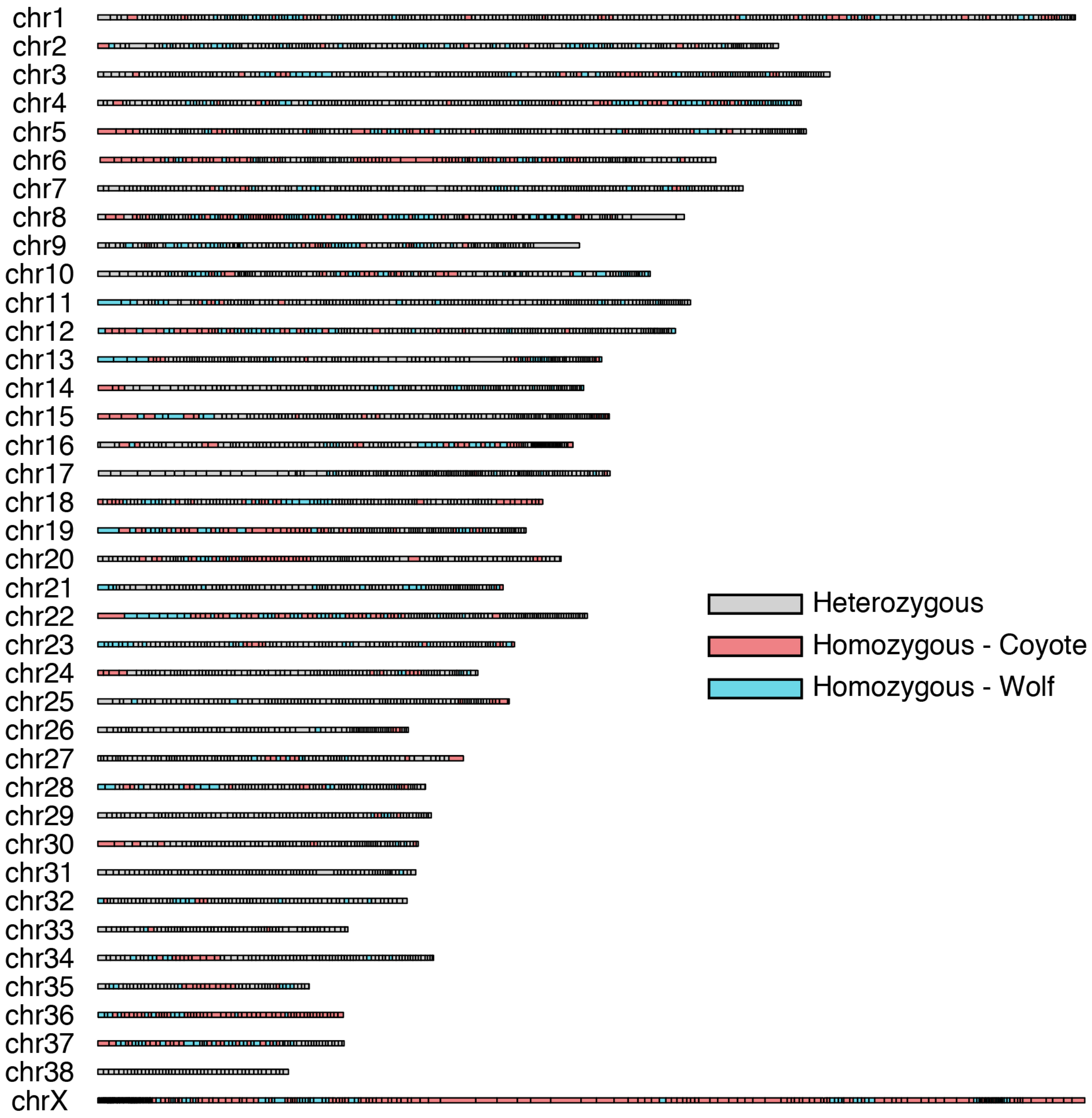

Figure S17: Dominant ancestries summarized along red wolf chromosomes, as determined via 'majority-rule' among delimited ancestry blocks merged into 500kb segments, excluding blocks for which ancestry could not be decisively determined.

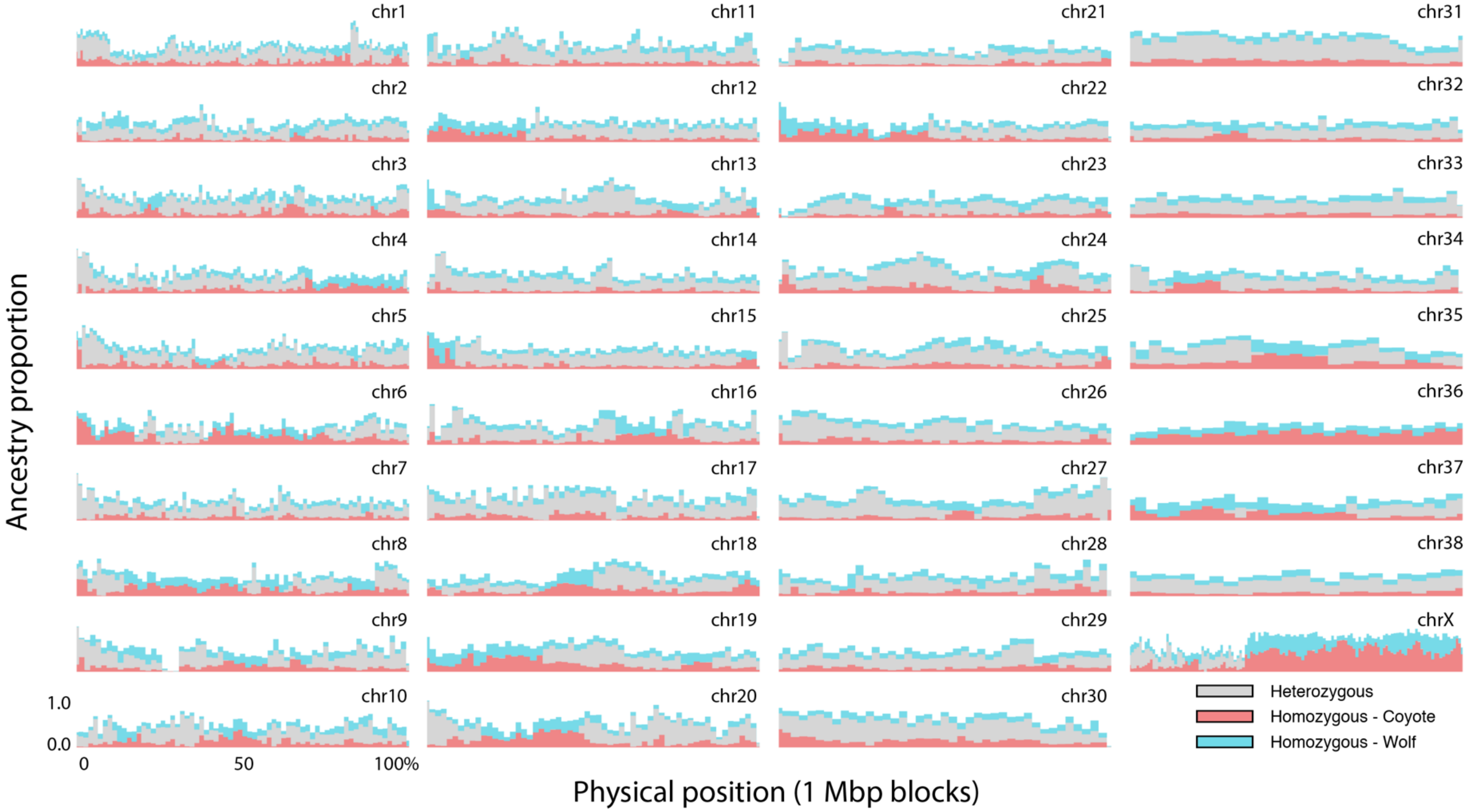

Figure S18: Proportion of assigned ancestries in the red wolf genome in 1 megabase blocks per chromosome, with blocks that could not be conclusively called as heterozygous or parental-homozygous excluded.

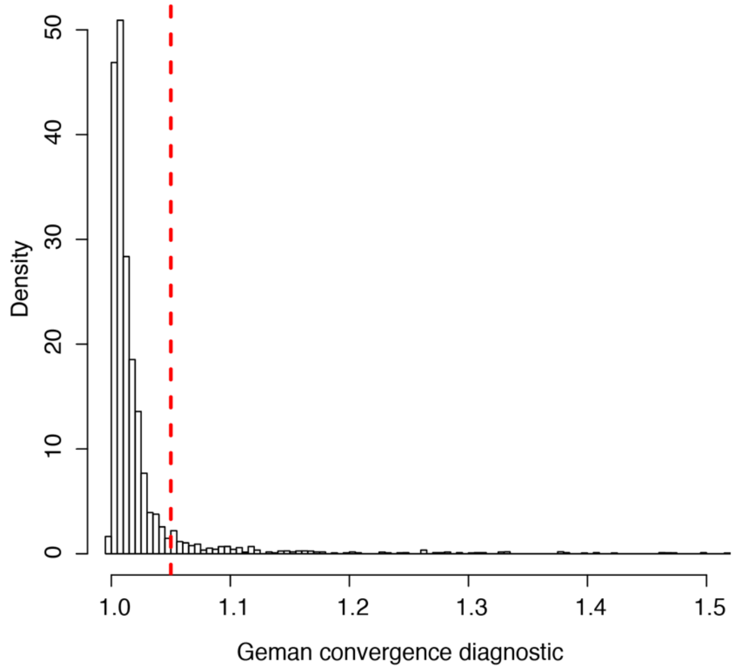

Figure S19: Histogram of Gelman convergence diagnostics across MCMC chains (each having two replicates), showing a cutoff threshold (red) of 1.05. Values shown are post burn-in, following an automated iterative procedure testing burn-in values according to the Geweke diagnostic.

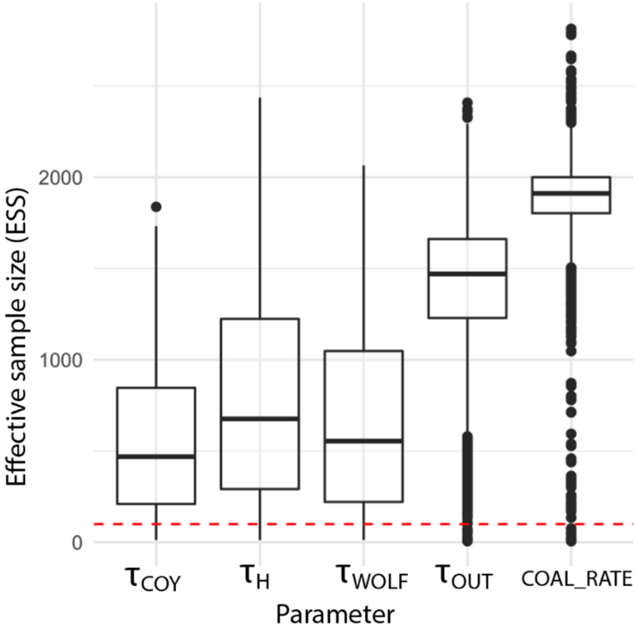

Figure S20: Effective sample sizes summarized across coalHMM MCMC chains passing Gelman-Rubin convergence threshold of 1.05. The minimum ESS threshold (red) of 100 was used to filter coalHMM results.

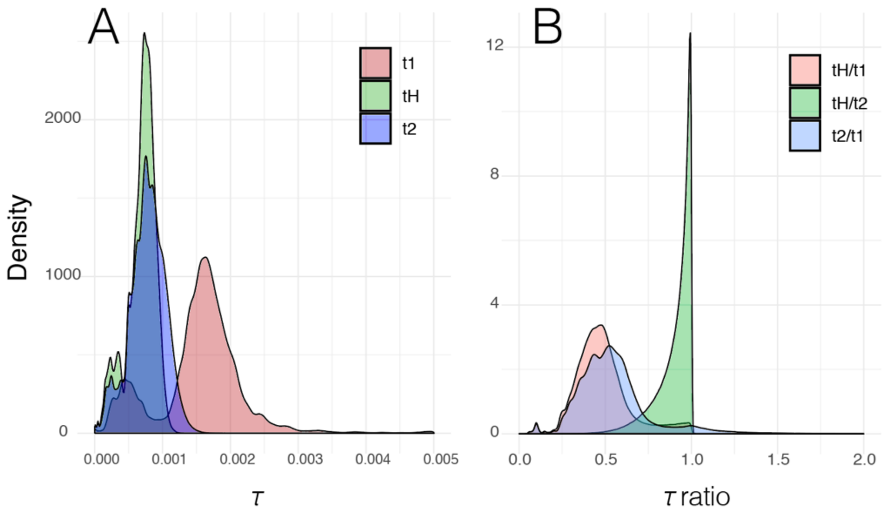

Figure S21: Posterior distributions of coalescent times inferred using coalHMM within 1Mb blocks of the red wolf genome (A) and the ratios among dates (B)

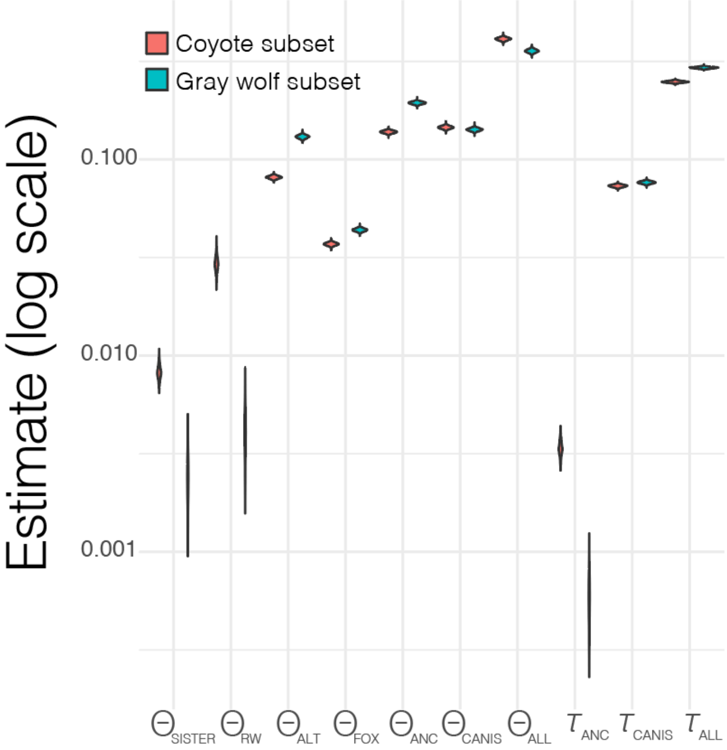

Figure S22: Full set of parameter estimates from g-PhoCS. Results are grouped by putative sub-genomes, showing mutation-scaled effective population size (Θ) and divergence times (τ). Parameters are displayed on a log scale, as follows: Effective size per subset (Θ_{SISTER}); red wolf (Θ_{RW}); non-source (Θ_{ALT}); red wolf and source (Θ_{ANC}); all *Canis* (Θ_{CANIS}); and root (Θ_{ALL}), as well as divergence time for red wolf and sister (τ_{ANC}); all *Canis* (τ_{CANIS}); and root (τ_{ALL}).

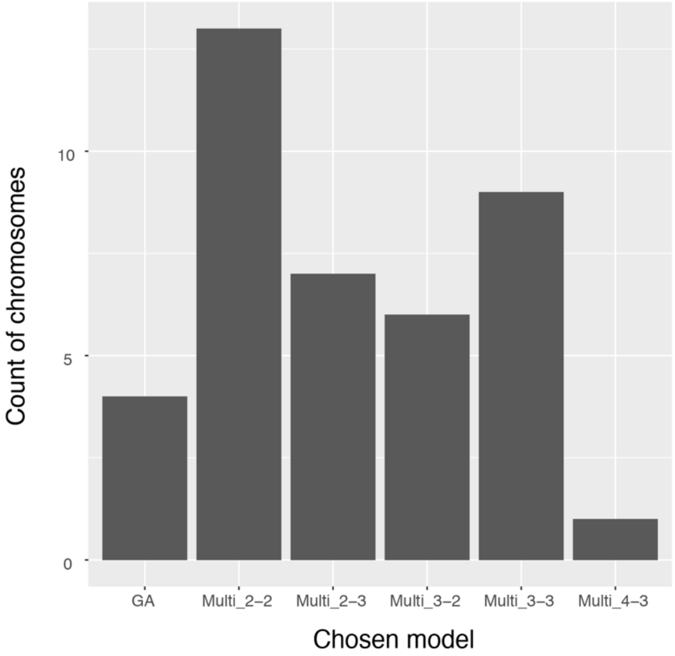

Figure S23: Total count of chromosomes choosing categories of admixture models, where GA=gradual admixture; and Multi_*x*-*y* represents multiple-pulse admixture models where *x*=number of inferred coyote admixture events and *y*=number of gray wolf events.

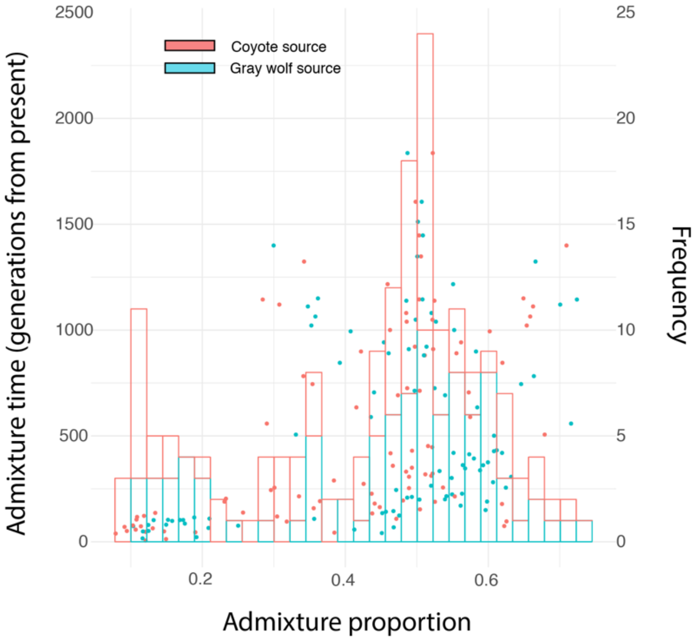

Figure S24: Admixture proportion for inferred admixture events inferred at different times (measured in generations before the present). The left axis (=points) show the measured admixture times as a function of admixture proportions, which the right axis (=histogram) shows frequency of admixture proportions across all events.

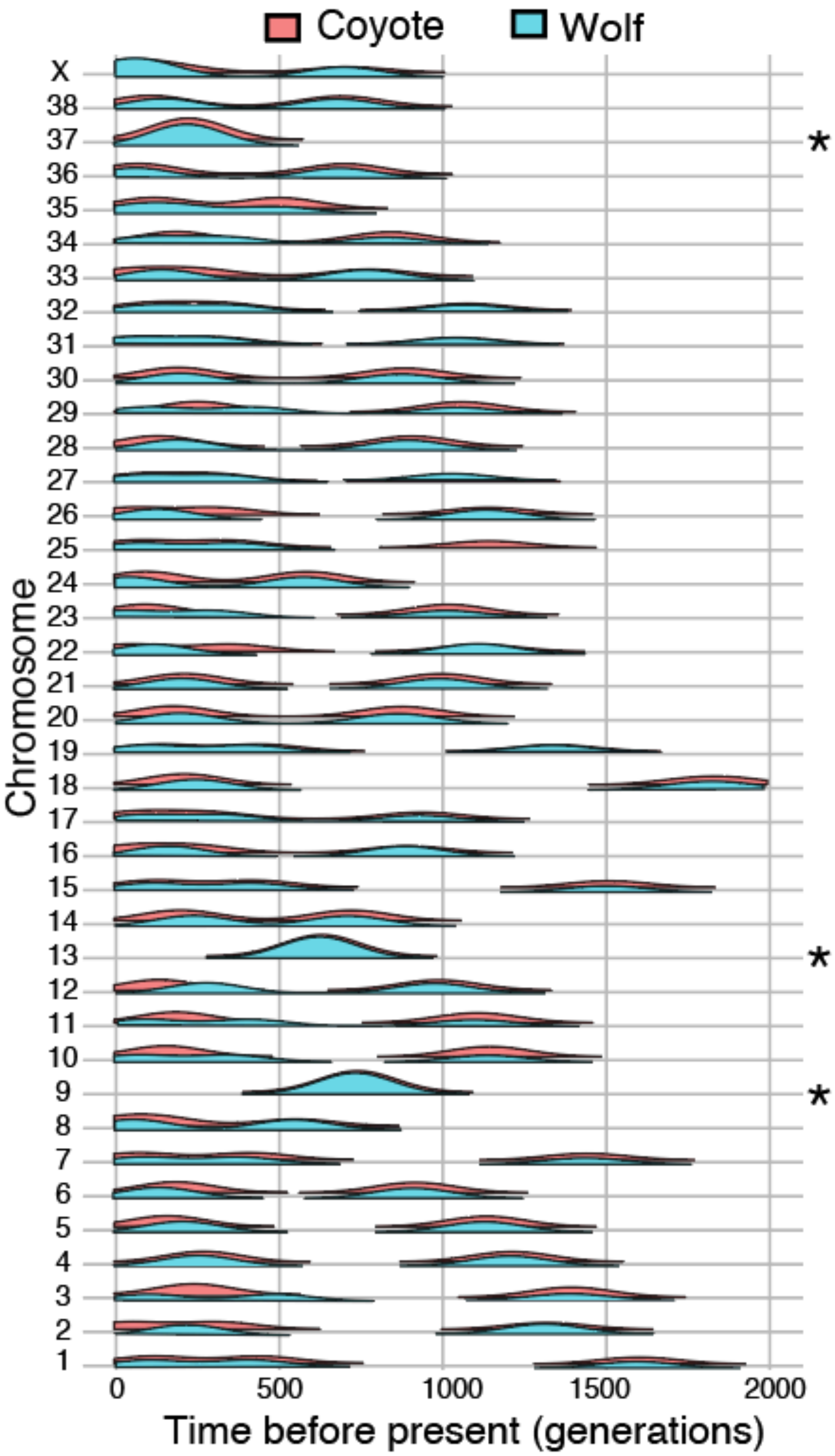

Figure S25: Admixture times inferred from ancestry block lengths in a multiple-pulse hybridization model. Time is shown in generations before the present, with the *y*-axis depicting densities based on 100 replicate datasets per chromosome, wherein heterozygous blocks were randomly assigned ancestry. Note that chromosomes best fitting a single-pulse model (*) depict the inferred start time for a gradual admixture (GA) scenario in which migration rates thereafter towards the present.