THE ORIGINS AND MAINTENANCE OF GENOMIC VARIATION IN THE THREESPINE STICKLEBACK (*Gasterosteus aculeatus*)

by

THOMAS C. NELSON

A DISSERTATION

Presented to the Department of Biology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2017

DISSERTATION APPROVAL PAGE

Student: Thomas C. Nelson

Title: The Origins and Maintenance of Genomic Variation in the Threespine Stickleback (*Gasterosteus aculeatus*)

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Biology by:

Matthew Streisfeld	Chair
William Cresko	Advisor
Patrick Phillips	Core Member
John Postlethwait	Core Member
Kirstin Sterner	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School

Degree awarded June 2017

DISSERTATION ABSTRACT

Thomas C. Nelson

Doctor of Philosophy

Department of Biology

June 2017

Title: The Origins and Maintenance of Genomic Variation in the Threespine Stickleback (*Gasterosteus aculeatus*)

Genetic variation is the raw material of evolution. The sources of this variation within a population, and its maintenance within a species, have been mysterious since the birth of the field of evolutionary genetics. In this work, I study divergently adapted freshwater and marine populations of the threespine stickleback (*Gasterosteus aculeatus*) as an evolutionary model to track the origin of adaptive genetic variation and to describe the evolutionary processes maintaining variation across the genome. The stickleback is a small fish with a large geographic range encompassing the northern half of the Northern Hemisphere and composed of coastal marine habitats, freshwater lakes, and river systems. Populations of stickleback adapt rapidly to changes in habitat, and fossil evidence suggests that similar adaptive transitions have been ongoing in this lineage for at least ten million years. In this work, I develop a significant extension of restriction site-associated DNA sequencing (RAD-seq) to generate phased haplotype information to estimate gene tree topologies and divergence times at thousands of loci simultaneously. I find anciently derived clades of variation associated with marine and freshwater habitats in genomic regions involved in recent adaptive divergence; some divergence times extend to over ten million years ago. This history of adaptive divergence has had profound

effects on genetic variation elsewhere in the genome: chromosomes harboring freshwater-adaptive variants retain anciently derived variation in linked genomic regions, while marine chromosomes have much more recent ancestry. I present a conceptual model of asymmetric selective and demographic processes to explain this result, which will form a nucleus for future research in this species. Lastly, by incorporating genome-wide recombination rates estimated from multiple genetic maps, I describe a recombination landscape that is favorable to the maintenance of marine-freshwater genomic divergence. Low recombination rates in key chromosomal regions condense widespread divergence of the physical genome, encompassing many megabases, into a small number of Mendelian loci. Combined, my results demonstrate the interconnectedness of evolutionary processes taking place on ecological and geological timescales. The genetic variation available for adaptive evolution today is a product of the long-term evolutionary history of a species.

CURRICULUM VITAE

NAME OF AUTHOR: Thomas C. Nelson

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, Oregon
Kalamazoo College, Kalamazoo, Michigan

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2017, University of Oregon
Bachelor of Arts, Liberal Arts, 2009, Kalamazoo College

AREAS OF SPECIAL INTEREST:

Evolutionary Genetics
Population Genomics
Conservation Genetics

PROFESSIONAL EXPERIENCE:

Graduate Employee, Department of Biology, University of Oregon, Eugene, 2015-present

Graduate Teaching Fellow, Department of Biology, University of Oregon, Eugene, 2011-2012

Assistant Scientist, Essen BioScience, Inc., Ann Arbor, Michigan, 2009-2011

GRANTS, AWARDS, AND HONORS:

Doctoral Dissertation Improvement Grant, National Science Foundation, 2015-2017

Cancer Federation Scholarship Award, Cancer Federation, Inc., 2012

Graduate trainee in genetics, National Institutes of Health, 2012-2015

Magna cum Laude, Kalamazoo College, 2009

Phi Beta Kappa, Delta of Michigan, 2009

Summer Research Grants, Howard Hughes Medical Institute, 2006, 2007

PUBLICATIONS:

Kasimatis, K., T. C. Nelson, and P. C. Phillips. “The genomic signatures of sexual conflict”, *in review at Journal of Heredity*

Castilla, L., S. Warburton, W. Burggren, D. Muirhead, T. C. Nelson, and E. M. Dzialowski. “Circulatory changes associated with the closure of the ductus arteriosus in hatchling emu (*Dromaius novaehollandiae*)”. *Comp. Biochem. Physiol. A.* 191 (2016): 202-208

Nelson, T. C., K. Groth, P. R. Sotherland. “Maternal investment and nutrient use affect phenotype of American alligator and domestic chicken hatchlings.” *Comp. Biochem. Physio. A* 157 (2010): 19-27.

ACKNOWLEDGEMENTS

To begin, scientists and colleagues. First, I thank Bill not only for being my advisor and sounding board for the past five years, but for my first real introduction to evolution as an active process. I started many dark, rainy mornings during my first Oregon winter relishing strong coffee and sampling distributions, and I'll never be the same. Many thanks go to my committee, whose influence will not end here: Patrick, for the day trips into LandeLand and for sage advice that is always helpful and almost never solicited; John, for asking the questions I never would have thought to prepare an answer for; Kirstin, for keeping my eyes on the endgame; and Matt, if I can convince you of a result, I know it's good. Thanks to members of the Cresko Lab, current and former: Mark, Susie, Clay, Julian, Emily, Emily, Kristin, Kristin, Kate, Kat, Ann, Allison, and John. This work also benefitted greatly from conversations and input from John Willis, Lila Fishman, Matt Hahn, Martin Kreitman, John Reinitz, Katja Kasimatis, Madeline Chase, Sean Stankowski, and Kristin Potter and Sam Gerber from UO CAS-IT. A hearty thanks goes to my undergraduate advisor, Paul Sotherland, for all the adventures with archosaurs. And this work would never have been possible without those who work so hard just so the rest of us can get anything done: Mark Currey, Sara Nash, and Arlene Deyo.

This adventure would have been even more tortuous without a great deal of inspiration. My thanks to the artists and writers who also look for truth and understanding: to Douglas Adams, for your wit, curiosity, and wisdom; to David James Duncan, who would like that I am writing this in a cabin by a mountain stream; to Barbara Kingsolver, David Attenborough, Neil Tyson, Steve Jobs, Emma Watson, Aldo Leopold, Carolyn Porco, Richard Lewontin, and whoever wrote Zoo Books; to Carl Zimmer, for retweeting me that one time; to Lyle Lovett, Ryan Adams, and John Williams; to The Boss, for reasons that should be obvious; to Lin-Manuel Miranda, Aaron Sorkin, David Remnick, Joe Biden, John Favreau, and John Lovett; to Michelle and Barack Obama.

To my friends and family, who keep me grounded and remind me why the hard work is worth it. To Zachary Levi, the unlikeliest of wisemen. To Art Night, for your begrudging acceptance that making figures counts as making art. To those who have fomented my curiosity: to Tim, for shipping a scorpion to me when I was a kid (it was dead); to Susan, who may be the sole reason I've ever read anything. To those who have challenged me: to Bruce, who shows me other ways of knowing; to Zaide, who gave me my first D on a paper (and for good reason); to Jim, who knows the importance of detail; to Robert, for steeling my belief in the value of diversity; and to Hugo, for his humility and capability. To my parents, Vicki and Charles, for their endless support and confidence. If I can be a fraction as clever and hardworking as they think I am, I will make something of myself eventually. And to my sister, Liz, who is capable and intelligent in ways I will never understand and to whom I will always struggle to catch up.

Finally, I owe an extra debt of gratitude to Kristin Alligood, who for the past six years has been my labmate, my roommate, my cheerleader, my life coach, my wingman, my role model, and my friend. Thanks, KA!

OK, time's up. They're playing me off.

To Ellie.
My understanding of Nature depends on imperfection.
You will be a difficult one to explain.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Evolution by natural selection generates and shapes biodiversity (Darwin 1859, Nosil 2012). The raw material of evolution is genetic variation, which has its ultimate source in random mutation. Once thought to be rare in natural populations (Fisher 1930), we now know that genetic variation is ubiquitous and highly structured throughout the genome and across geography (Hubby and Lewontin 1966, Lewontin and Hubby 1966, Lewontin 1974, Endler 1977, Langley, Stevens et al. 2012). This abundance of raw material supports another surprising observation: natural populations adapt quickly, responding to strong selection pressures on short, ecologically-relevant timescales (Grant and Grant 2002, Barrett, Rogers et al. 2008, Lescak, Bassham et al. 2015). In these bouts of evolution, natural selection quickly and effectively structures and filters variation, much of it likely pre-existing. But this leaves the field of evolutionary biology with a more general lack of knowledge because, except in rare cases, we do not know the origins of this standing adaptive genetic variation. Nor do we have a complete understanding of which evolutionary forces and processes maintain variation within a species or influence its genomic and geographic distribution. To address questions of the origins and the maintenance of genetic variation, which are explicitly historical questions, it is not sufficient to identify those variants or to treat them in isolation. Addressing these questions requires us to reconstruct the evolutionary history of genetic variants and to do so in the context of the genomes in which they reside.

The sources of genetic variation available during adaptive evolution in large part determine the pace and paths available to adaptation. Much of classical evolutionary theory works under the assumption that adaptation proceeds via the fixation of new mutations that arise during the process of adaptation (Fisher 1930, Orr 2005). This model can certainly hold true in lineages with large population sizes, including prokaryotes (Tenaillon, Barrick et al. 2016). But for many plants and animals, a sole reliance on new mutation would lead to adaptation being mutation-limited, particularly when

environments change rapidly (Charlesworth 2009, Karasov, Messer et al. 2010). Indeed, more and more studies of adaptation in the wild are finding that adaptation draws on *standing genetic variation* — the pool of segregating variation present in a population before a selection pressures arises (Barrett and Schluter 2008). If the goal is to understand the sources of adaptive variation, however, saying that a variant was standing is only a partial answer. When did that variant originally arise as a new mutation (Figure 1.1)? Was it evolving neutrally or had it already been filtered by selection? By studying variation in haplotypes at genes known to be under selection, recent studies in plants (Stankowski and Streisfeld 2015); insects (Fontaine, Pease et al. 2015, Wallbank, Baxter et al. 2016); and mammals (Linnen, Kingsley et al. 2009), including humans (Huerta-Sánchez, Jin et al. 2014), have identified complex histories behind alleles involved in adaptation — histories that span millions of years and involve genetic exchange among multiple populations and even separate species. But whether these case studies represent the larger pool of standing genetic variation or are interesting outliers remains unknown. To address those questions

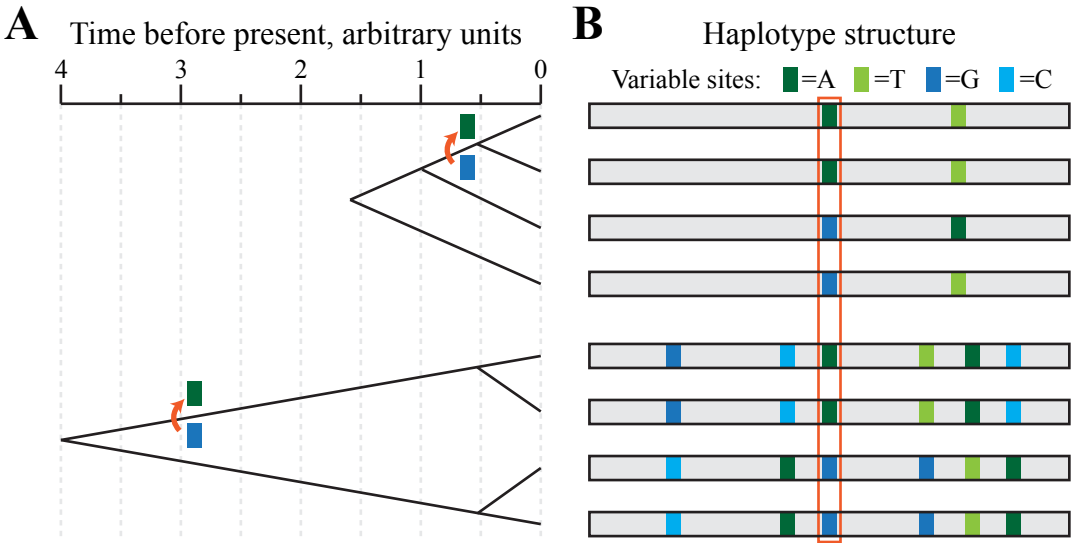

Figure 1.1. Haplotype structure illuminates evolutionary history. Panel A: Two gene trees representing the branching history of a non-recombining segment of DNA. An A-to-G transition occurs on both gene trees. Panel B: Example haplotype structure of DNA sequences given the gene trees in (A). The highlighted polymorphism is uninformative of the true history, which is revealed by examining linked variation.

we now require studies of the genetic history of adaptive evolution across entire genomes.

Even those few case studies speak to a second point: the alleles involved in adaptation can have complex histories, involving multiple populations and even multiple species. Hence, the structuring of populations across geography and through time means that adaptation often occurs in the context of a metapopulation (Pannell and Charlesworth 2000). Metapopulation dynamics have been part of evolutionary theory in some form since the beginning of the Modern Synthesis (Wright 1932). For much of the history of evolutionary genetics biologists simply lacked the empirical data supporting the relevance of these models to our understanding of adaptation in nature. Now, genomically enabled studies of natural populations are reintroducing the importance of population structure and demographic history to the trajectory of adaptive evolution, but with the new understanding that the effects of these processes can be heterogeneously realized across the genome (Hohenlohe, Bassham et al. 2010, Nosil and Feder 2012, Soria-Carrasco, Gompert et al. 2014). Genetic differentiation of populations across geography is rampant, occurring on local and global scales (Lekberg, Roskilly et al. 2012, Skoglund, Mallick et al. 2015) and driven by both neutral (Leslie, Winney et al. 2015) and selective processes (Feder, Egan et al. 2012, Lekberg, Roskilly et al. 2012). In an even more fundamental shift in our understanding, species boundaries can be alarmingly porous, leading to genetic exchange between taxa that were long thought to be isolated evolutionary lineages (Kane, King et al. 2009, Prüfer, Racimo et al. 2014, Beck, Thompson et al. 2015).

All genetic variation evolves in the context of a genome. The purely physical nature of the genetic material means that variants residing on the same molecule of DNA are co-inherited, except when physical linkage is broken by meiotic recombination. This simple fact has drastic consequences for evolution by natural selection. Because selection acts on phenotypes, the fate of any single variant is tied not only to its direct effect on phenotype but to the combined effects of, and interactions between, the variants throughout the genome in which it resides and with which it is inherited. Selection

therefore sorts variation in genomes, making the genome the unit of selection and the proper scale of inference (Lewontin 1974).

Studying the genetic basis and evolutionary history of adaptive evolution genome-wide was until recently an extremely arduous and expensive endeavor confined to only laboratory model organisms and humans (Jensen-Seaman, Furey et al. 2004, Altshuler, Brooks et al. 2005, Harr 2006). Therefore, many of these core questions regarding the history of adaptive evolution have been unanswered in natural populations. However, over the past decade, thanks to the DNA sequencing revolution, our ability to detect genetic variation across the genome has become nearly unlimited, leading to efforts to detect the genetic targets of selection across the tree of life (Yi, Liang et al. 2010, Blount, Barrick et al. 2012, Prasad, Song et al. 2012, Vitti, Grossman et al. 2013). Unfortunately, genome-scale descriptions of genetic variation have far outpaced descriptions of *genomic* variation. Until we are able to describe the available variation in genomes that are seen by selection, our understanding of adaptation will remain incomplete. My dissertation helps to fill this gap in knowledge.

In the chapters that follow, I study the threespine stickleback fish (*Gasterosteus aculeatus*) as a model system to explore evolutionary histories and patterns of inheritance of genomic variation involved in and influenced by adaptive evolution. The stickleback is a small fish distributed throughout the Northern Hemisphere in coastal marine habitats and freshwater lakes, streams, and rivers (Bell and Foster 1994, McKinnon and Rundle 2002). Marine and freshwater populations are locally adapted and partially reproductively isolated from one another (Boughman 2001, Boughman 2009, Lackey and Boughman 2017), although gene flow does occur between neighboring populations in the absence of geographic barriers (Hendry and Taylor 2004, Roesti, Gavrilets et al. 2014). Most, if not all, freshwater populations that have been studied to date were founded recently in the history of the species (~15,000 years ago or less) by phenotypically and genetically marine fish (Bell and Foster 1994, Cresko, Amores et al. 2004), but phenotypic transitions typical of recently derived populations have been observed in the fossil record extending back millions of years (Bell 1994, Bell, Stewart et al. 2009).

The threespine stickleback is an ideal system in which to address the history of genomic variation for several reasons. First, adaptation in the stickleback relies heavily on standing genetic variation (Schluter and Conte 2009, Roesti, Gavrilets et al. 2014, Lescak, Bassham et al. 2015). But only in a single instance, the major-effect locus controlling a presence-absence polymorphism in bony plate armor (*Ectodysplasin A* [*Eda*]), have authors described the evolutionary history of the genetic variation at the causative locus (Colosimo, Hosemann et al. 2005). Second, the species exists as a metapopulation, with both selective and neutral processes structuring variation across geography (Catchen, Bassham et al. 2013, Roesti, Kueng et al. 2015). Third, genomic divergence between marine and freshwater populations is highly parallel across populations, heterogenous across the genome, and involves large chromosomal regions. This conserved genomic architecture hints at the selective maintenance of broader patterns of genomic variation.

In this work, I study adaptive divergence in two separate freshwater lake populations founded from an ancestral marine population in Cook Inlet, Alaska, at the end of the Pleistocene glaciation, approximately 12,000 years ago (Francis, Baumgartner et al. 1986, Reger and Pinney 1996, Cresko, Amores et al. 2004, Figure 1.2). I use a combination of reduced-representation and whole-genome sequencing of natural isolates, the development of new sequencing methodologies, and laboratory genetic mapping crosses to address the following questions:

1) What was the evolutionary history of the genetic variation involved in recent adaptive divergence?
2) How has selection structured and maintained genetic variation, both across geography and throughout the genome?
3) How does variation in the recombination rate influence coinheritance of genomic variation? Do these patterns differ in regions affected by divergent selection?

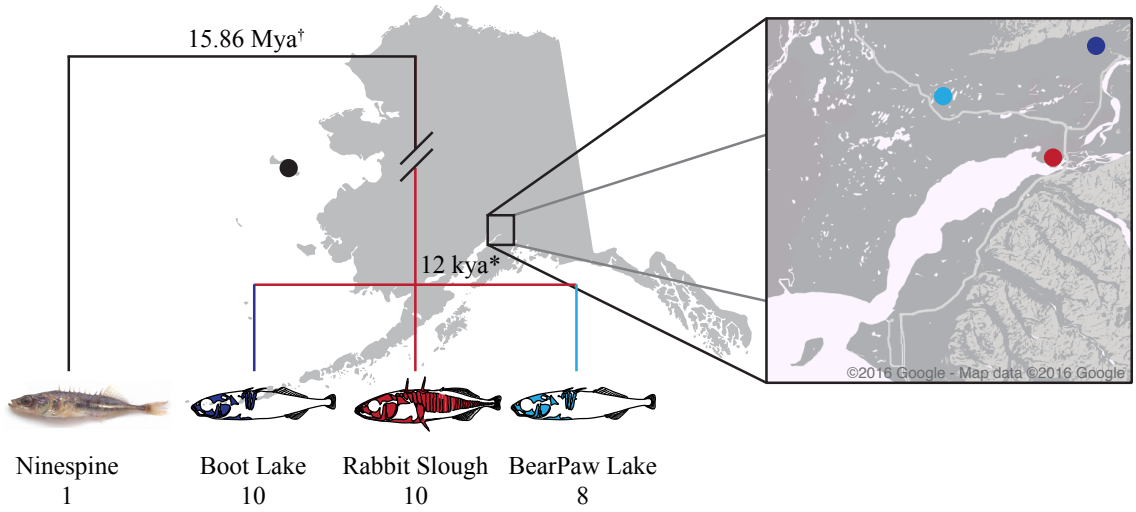

Figure 1.2. Stickleback populations used in this work. The phylogeny shows population-level relationships among stickleback populations, with approximate divergence times shown at nodes separating ninespine stickleback (*Pungitius pungitius*) from threespine stickleback (†Aldenhoven *et al*., 2010) and divergence times among Alaskan threespine stickleback populations (*Reger and Pinney, 1996). Branches within the threespine lineage are colored by the predominant phenotype in that population: Red, marine; blue, freshwater. Line drawings of threespine stickleback highlight stereotypical bony morphologies of marine and freshwater ecotypes. Actual sampling locations are shown as dots on the map of Alaska. Cook Inlet sampling locations are shown in the enlarged map. Line drawings courtesy of Kristin Alligood. Ninespine stickleback image from the Canadian Register of Marine Species (CaRMS photogallery 2010; www.marinespecies.org).

In chapter II, I describe patterns of DNA sequence variation and divergence associated with the recent colonization of new freshwater habitat. Toward this end, I develop a novel form of restriction site-associated DNA sequencing (RAD-seq) designed to construct local haplotypes (hRAD-seq) and analyze patterns of molecular evolution. This technique is not only critical for my work, but enables similar genome-scale genealogical analyses in many other organisms and will therefore have broad impact on the field of evolutionary genomics. With this technique I demonstrate that recent adaptation drew on a store of anciently derived alleles at loci that are under long-term selection toward alternative fitness optima in marine and freshwater environments. This work provides the first direct estimates of sequence divergence between freshwater and marine alleles at a genomic scale, indicating that the evolutionary history of the *Eda* locus is not an outlier,

but characteristic of a larger suite of anciently derived variation defining marine and freshwater stickleback forms. In chapter III, I extend these results into a coalescent framework and estimate the time to the most recent common ancestor (T_{MRCA}) of genomic variation affected by both divergent selection and parallel adaptation. While divergent selection maintains among-habitat variation over millions of years, parallel adaptation increases identity-by-descent among independently derived freshwater populations. Surprisingly, these conserved freshwater alleles reside on hyper-diverse genetic backgrounds which do not exist on marine chromosomes. I end chapter III with a conceptual model to explain this asymmetry. In chapter IV, I integrate estimates of genome-wide recombination rate variation from laboratory mapping crosses and find that low recombination rates reduce the dimensionality of marine-freshwater divergence, as large genomic regions collapse to a small number of Mendelian loci, inherited as discrete units, likely with large effects on fitness. These data provide a proximate mechanism for the maintenance of large, highly divergent genomic regions and suggest ongoing adaptive evolution of the recombination landscape itself. Finally, in chapter V I discuss the implications of these findings for our understanding of adaptation and point to critical areas for future work.

CHAPTER II

WISDOM OF THE ANCIENTS: LONG-STANDING GENETIC VARIATION DRIVES RAPID ADAPTATION IN THREESPINE STICKLEBACK

INTRODUCTION

Natural populations harbor abundant genetic variation, which is partitioned across geography that often includes multiple ecologically distinct habitats (Lewontin 1974, Avise 2000, Hedrick 2006, Nosil 2012). While neutral processes play an important role in structuring genetic variation, observed patterns of variation across geography and across the genome are now often thought to be either the direct result of, or indirectly impacted by, the action of natural selection (Begun, Holloway et al. 2007, Hahn 2008, Burri, Nater et al. 2015, Martin, Möst et al. 2016). The neutral and selective history of the standing variation within a species therefore impacts the tempo and mode of adaptation to new environments, for example by maintaining readily available adaptive alleles or by depleting the pool of standing genetic variation. Recently, the field of population genomics, spurred by advances in DNA sequencing technologies, has begun revealing diverse sources of adaptive variation in natural populations, including contributions from standing genetic variation and introgressed alleles (Domingues, Poh et al. 2012, Huerta-Sánchez, Jin et al. 2014, Malinsky, Challis et al. 2015). A main goal of evolutionary biology is now to understand this history of adaptive genetic variation (Fontaine, Pease et al. 2015, Stankowski and Streisfeld 2015, Wallbank, Baxter et al. 2016) both to define its origin and to better understand how this history might influence future change.

Studies of adaptation in the wild often conflict with the standard genetic models of adaptation via new mutations (Barrett and Schluter 2008). The standard model envisions adaptation toward a single fitness optimum — that is, driven by the sequential fixation of new mutations (Fisher 1930, Orr 2005). In contrast, natural populations often respond remarkably quickly to selective pressures (Palumbi 2001, Grant and Grant 2002, Barrett, Rogers et al. 2008, Lescak, Bassham et al. 2015), showing rapid phenotypic

shifts and casting doubt on the ability of new mutations to contribute to such adaptive evolution on ecological timescales. Also contrary to these classical theories, adaptation in the wild often occurs across a landscape (Gillespie 1973, Endler 1977), with diverse ecologies leading to adaptation toward multiple fitness optima (Felsenstein 1976, Ewing 1979, Hedrick 2006). Migration among populations leads to the formation of metapopulation dynamics (Levin 1995). In comparison to a single panmictic population, genetic differentiation among populations can retain substantial amounts of standing genetic variation (Charlesworth, Charlesworth et al. 2003), which can traverse the metapopulation via gene flow.

A key finding from studies of the genetics and genomics of natural populations is that adaptation often relies extensively on standing genetic variation (reviewed in Barrett and Schluter 2008). Standing variation is now thought to be a common source of adaptive variation because alleles exist at higher frequencies than new mutations and are available immediately during a selective episode (Hermisson and Pennings 2005). Experimental and observational studies have demonstrated adaptation to novel environments occurring within one or a few generations (Grant and Grant 2002, Ingram, Svanbäck et al. 2012) and involving both Mendelian (Barrett, Rogers et al. 2008) and quantitative traits (Brodie 1992, Grant and Grant 2002). Standing genetic variation can also include larger structural genomic features such as chromosomal inversions, which can maintain linkage disequilibrium between adaptive alleles at many loci (Kirkpatrick and Barton 2006, Kirkpatrick 2010, Lowry and Willis 2010). The extent to which segregating adaptive variation influences longer-term patterns of evolution is now evident in phylogenomic studies of higher level taxonomic groups, where species separated by tens of millions of years can retain adaptive variation that was present in ancestral populations (Hahn and Nakhleh 2016, Pease, Haak et al. 2016).

Despite its importance to adaptation on ecological timescales, we know little about either the evolution of standing genetic variation or its distribution across geography and across the genome. Did selection act on standing variants that were evolving neutrally prior to the observed adaptive episode or has selection re-used genetic

variants that had been under selection previously? Are standing variants recently derived, existing as few copies on limited genetic backgrounds, or have they persisted for longer periods of time? Is standing genetic variation dispersed or clustered in genomes and, if so, why? Answers to these questions require not only the identification of the allele or genomic region involved in adaptation, but also patterns of haplotype variation that reveal the evolutionary history of that region. Where authors have dissected the histories of adaptive variants, they have found complex and varying answers (Colosimo, Hosemann et al. 2005, Stankowski and Streisfeld 2015, Wallbank, Baxter et al. 2016). But in part because extracting informative haplotype information is time- and labor-intensive, these studies typically analyze a single locus or genomic region and we know much less about whether these patterns reflect those that are common across the genome.

Here, we use the threespine stickleback fish (*Gasterosteus aculeatus*) as a model to investigate the structure and history of adaptive genomic variation. The stickleback is a natural model system for the genomics of adaptation in the wild (Hohenlohe, Bassham et al. 2010, Jones, Grabherr et al. 2012), with phenotypically and ecologically diverse populations adapted to coastal marine environments as well as freshwater ponds, lakes, and stream systems (Bell and Foster 1994, Cresko, Amores et al. 2004, Roesti, Hendry et al. 2012). In many regions throughout the Northern Hemisphere, stickleback from ancestrally oceanic populations have independently colonized freshwater habitats and adapted with remarkable speed (Cresko, Amores et al. 2004, Schluter, Marchinko et al. 2010, Lescak, Bassham et al. 2015). Independent freshwater colonization events often involve convergent phenotypic evolution that is mirrored by convergent genomic signatures of selection (Hohenlohe, Bassham et al. 2010, Jones, Grabherr et al. 2012, Roesti, Kueng et al. 2015). The speed and parallelism of freshwater divergence in stickleback suggest that alleles adaptive to freshwater likely exist as standing genetic variation in the ocean ancestors (Schluter and Conte 2009, Lescak, Bassham et al. 2015). In a now classic example, the loss of bony lateral plate armor in multiple freshwater populations was associated with fixation of common allelic variants of the *Ectodysplasin A* (*Eda*) locus that were identical-by-descent and anciently derived from the oceanic

alleles (Colosimo, Hosemann et al. 2005). A strikingly similar pattern was recently observed associated with a chromosomal inversion on chromosome 1, indicating that deeply divergent haplotypes may be a common phenomenon in this system (Roesti, Gavrilets et al. 2014). However, convergent phenotypes with independent genetic origins have also been documented, as in the case of reduced pelvic structural features in freshwater populations (Chan, Marks et al. 2010). These data point to the diverse molecular evolutionary histories of adaptive variation in this system even as they account for a small fraction of the genomic regions involved in adaptation.

To identify patterns of molecular evolution at a genomic scale, we extend restriction site-associated DNA sequencing (RAD-seq) to generate high-confidence local haplotypes for sequence analysis, a variant we refer to as hRAD-seq. RAD-seq has proven to be a powerful technique to identify SNPs in natural populations (Andrews, Good et al. 2016), and hRAD-seq extends this utility to the measurement of DNA sequence variation in a way that was limiting in previous RAD-seq approaches. Specifically, we estimate sequence divergence within and among two divergently adapted stickleback populations in two locations in Cook Inlet, Alaska, USA: Boot Lake and Rabbit Slough. The Boot Lake population was founded approximately 12,000 years ago and displays stereotypical freshwater morphology, including reduced lateral plate armor and a reduced pelvic structure, and stereotypical patterns of genomic divergence (Cresko, Amores et al. 2004, Hohenlohe, Bassham et al. 2010). The Rabbit Slough population is an anadromous stickleback population that displays stereotypical marine stickleback morphology and is representative of the marine stickleback that likely colonized Boot Lake (Cresko, Amores et al. 2004). Below, we confirm that adaptive divergence in the Boot Lake population occurred primarily via standing genetic variation and demonstrate that allelic divergence in the genomic regions involved in the marine-to-freshwater transition can be ancient, pre-dating most of the genetic variation across the rest of the stickleback genome. These data suggest common mechanisms for the long-term maintenance of adaptive variation within a species.

METHODS

Sample Collection and Library Preparation

Wild threespine stickleback were sampled from two locations in Cook Inlet, Alaska, USA. We sampled five phenotypically freshwater fish from Boot Lake (N 61.7167, W 149.1167) and five oceanic fish from Rabbit Slough (N 61.5595, W 149.2583) (Cresko, Amores et al. 2004, Hohenlohe, Bassham et al. 2010). DNA was extracted from fin clips preserved in 95% ethanol using either Qiagen DNeasy spin column extraction kits or Ampure magnetic beads (Beckman Coulter, Inc) following manufacturer's instructions. Yields averaged 1-2 μg DNA per extraction (~30 mg tissue).

We designed our library preparation strategy to simplify downstream sequence processing and analysis by taking advantage of the phase information captured by paired-end sequencing. We generated RAD libraries from these samples using the single-digest sheared RAD protocol from Baird et al. (Baird, Etter et al. 2008) with the following specifications and adjustments: 1 μg of genomic DNA per fish was digested with the restriction enzyme PstI-HF (New England Biolabs), followed by ligation to P1 Illumina adaptors with 6 bp inline barcodes. Ligated samples were multiplexed and sheared by sonication in a Bioruptor (Diagenode). To ensure that most of our paired-end reads would overlap unambiguously and produce longer contiguous sequences, we selected a narrow fragment size range of 425-475 bp. The remainder of the protocol was per Baird et al (2008). All fish were sequenced on an Illumina HiSeq 2500 using paired-end 250 bp sequencing reads at the University of Oregon's Genomics and Cell Characterization Core Facility (GC3F).

Sequence Preparation

Raw Illumina sequence reads were demultiplexed, cleaned, and processed primarily using the Stacks pipeline (Catchen, Hohenlohe et al. 2013). Paired-end reads were demultiplexed with **`process_shortreads`** and cleaned using **`process_radtags`** using default criteria (throughout this document, names of scripts, programs, functions, and command-line arguments will appear in **`fixed-width font`**).

Overlapping read pairs were then merged with **`fastq-join`** (Aronesty 2011) (Fig. S1). Pairs that failed to merge were removed from further analysis. In order to retain the majority of the sequence data for analysis in Stacks and still maintain adequate contig lengths, merged contigs were trimmed to 350 bp and all contigs shorter than 350 bp were discarded. We aligned these contigs to the stickleback reference genome (Jones, Grabherr et al. 2012, Glazer, Killingbeck et al. 2015) using **`bbmap`** with the most sensitive alignment settings ('**`vslow=t`**'; http://jgi.doe.gov/data-and-tools/bbtools/) and used the **`pstacks`**, **`cstacks`**, and **`sstacks`** components of the Stacks pipeline to create stacks and call SNPs and haplotypes, create a catalog of RAD tags across individuals, and match tags across individuals. All data were then passed through the Stacks error correction module **`rxstacks`** to prune unlikely haplotypes . We ran the Stacks component program populations on the final dataset to filter loci genotyped in fewer than four individuals in each population and to create output files for sequence analysis. Below, we use the naming conventions of Baird et

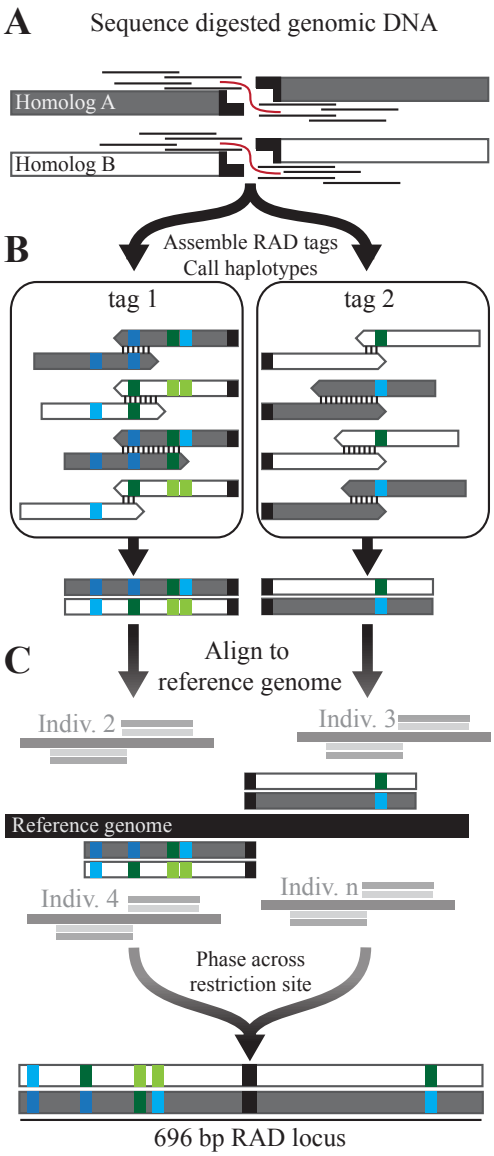

Figure 2.1. hRAD-seq generates local haplotypes from next-generation sequencing technologies. (A) Overlapping paired-end reads are anchored to each end of a palindromic restriction enzyme cut site. (B) Overlapping reads are merged into contigs and haplotypes called at each locus. (C) If a genomic reference is available, the two loci representing a single cut site are phased into a single locus. Haplotypes shown here correspond to RAD data shown in Figure 2.5A. Heterozygous SNPs are shown: A, G, T, C are dark blue, light blue, dark green, light green, respectively.

al. (2008): A "RAD tag" refers to sequence generated from a single end of a restriction site and the pair of RAD tags sequenced at a restriction site comprises a "RAD locus" (Figure 2.1).

We used the program **`phase`** (Stephens, Smith et al. 2001) to phase pairs of RAD tags originating from the same restriction site. We coded haplotypes present at each RAD tag, which often contain multiple SNPs, into multiallelic genotypes. This both simplified and reduced computing time for the phasing process. Custom Python scripts automated this process. We required that each individual had at least one sequenced haplotype at each tag for phasing to be attempted. If a sample had called genotypes at only one tag in the pair, the sample was removed from further analysis of that locus. The resultant phased haplotypes were used to generate sequence alignments for import into BEAST.

Sequence Diversity and Divergence

We quantified sequence diversity within and among populations and sequence divergence between populations using R (R Core Team 2016). We used the R package 'ape' (Paradis, Claude et al. 2004) to compute pairwise distance matrices for all alleles at each RAD locus and used these matrices to calculate the average pairwise nucleotide distances, π, within and among populations along with d_{xy}, the average pairwise distance between two sequences using only across-population comparisons (Nei and Li 1979). We also calculated haplotype-based F_{ST} from Hudson et al. (1992) implemented in the R package 'PopGenome' (Pfeifer, Wittelsburger et al. 2014).

To test for correlations between relative (F_{ST}) and absolute divergence (d_{xy}) and nucleotide diversity (π), we used type-II linear models implemented in the R package 'lmodel2' (https://CRAN.R-project.org/package=lmodel2). We binned all RAD loci into 100 kb non-overlapping genomic windows and then ran linear models using the average estimate of each statistic within that window. To identify patterns of diversity and divergence in a genomic context, we kernel-smoothed population genetic statistics using the base R function ksmooth() with a bandwidth of 100kb and a normal kernel density.

Analysis scripts along with our R function to calculate π, πwithin, and d_{xy} are provided as supplementary material.

Lineage Sorting at RAD Loci

Allelic divergence can occur by multiple modes of lineage sorting during adaptation (see Figure 2.5). To identify patterns of lineage sorting associated with freshwater colonization, we analyzed gene tree topologies at all RAD loci using BEAST v. 1.7 (Drummond and Rambaut 2007, Drummond, Suchard et al. 2012). We used blanket parameters and priors for BEAST analyses across all RAD loci. Markov chain Monte Carlo (MCMC) runs of 1,000,000 states were specified, and trees logged every 100 states. We used a coalescent tree prior and the GTR+Γ substitution model with four rate categories and uniform priors for all substitution rates. We identified evidence of lineage sorting by analyzing trees logged from the BEAST analysis in R. We used the R package 'ape' (Paradis, Claude et al. 2004) to parse trees and identify monophyly of groups of alleles. We determined for each sampled tree whether haplotypes originating from Boot Lake (BT) or Rabbit Slough (RS) fish formed monophyletic clades. Trees therefore fell into one of four categories: (1) neither BT nor RS are monophyletic, (2) RS is monophyletic, (3) BT is monophyletic, or (4) BT and RS are reciprocally monophyletic. To call a given locus as one of these four categories, 90% of sampled trees must have exhibited the given pattern.

Eda *Gene Tree and Comparison to Other Loci*

To compare results from hRAD sequencing to known patterns of adaptation from standing genetic variation, we identified RAD loci adjacent to coding regions of the *Ectodysplasin A* locus (*Eda*). We chose a reciprocally monophyletic RAD locus 10 kb downstream of the last exon of *Eda* to highlight local within- and among-population haplotype variation. We analyzed the gene tree with BEAST as described above. We then used the program **`treeannotator`** (Drummond, Suchard et al. 2012) to summarize BEAST log file and generate a maximum clade credibility tree using median node heights. The resulting gene tree was visualized in **`figtree`** (http://tree.bio.ed.ac.uk/

software/figtree/). We also used the function **`haploNet()`** in the R package 'pegas' (Paradis 2010) to generate a haplotype network under the infinite sites model. To ensure fair comparison of RAD loci adjacent to *Eda* with other anonymous RAD loci, we manually inspected alignments for all reciprocally monophyletic RAD loci with high levels of oceanic-to-freshwater sequence divergence (above the 1.5 interquartile range; $d_{xy} > 0.038$) to identify those with spuriously high estimates resulting from repetitive sequence or otherwise unclear homology between oceanic and freshwater haplotypes.

RESULTS

Raw Data are of High Quality

RAD sequencing resulted in 95,302,848 raw read pairs, 94,065,236 of which (98.7%) passed quality filtering. Across samples, 84.4 percent of the paired reads successfully merged into single contigs ranging in length from 251 bp to 490 bp. Contigs averaged 379 bp in length, with 99.2% of contigs ≥350 bp in length (Figure 2.2A). 359,300 RAD tags were present and genotyped in at least four of the five samples from each population and were passed to the phasing step. We successfully recovered and phased RAD loci from 102,823 PstI restriction sites (205,646 RAD tags), resulting in 696 bp haplotypes with 690 possible variable sites per locus (6 bp PstI recognition motif was invariant) for a total of 71.9 Mb of aligned genomic sequence, or 15.5% of the total stickleback genome. Average RAD locus coverage across the dataset after final filtering was 11.6X; coverage of individual samples ranged from 9.8X to 13.7X (Figure 2.2B). Based on a genome size of 463 Mb (Jones, Grabherr et al. 2012), RAD loci averaged 4.50 kb apart. Using alignments to the stickleback reference genome (Jones, Grabherr et al. 2012, Glazer, Killingbeck et al. 2015), RAD loci averaged 4.24 kb apart; the median distance between RAD loci was 2.79 kb (Figure 2.2C).

Genetic Differentiation is Correlated with Sequence Divergence

Average genome-wide nucleotide diversity, π, was 0.00496 (range: 0.00014 — 0.11077). Genetic diversity was reduced slightly within each population, with genome-

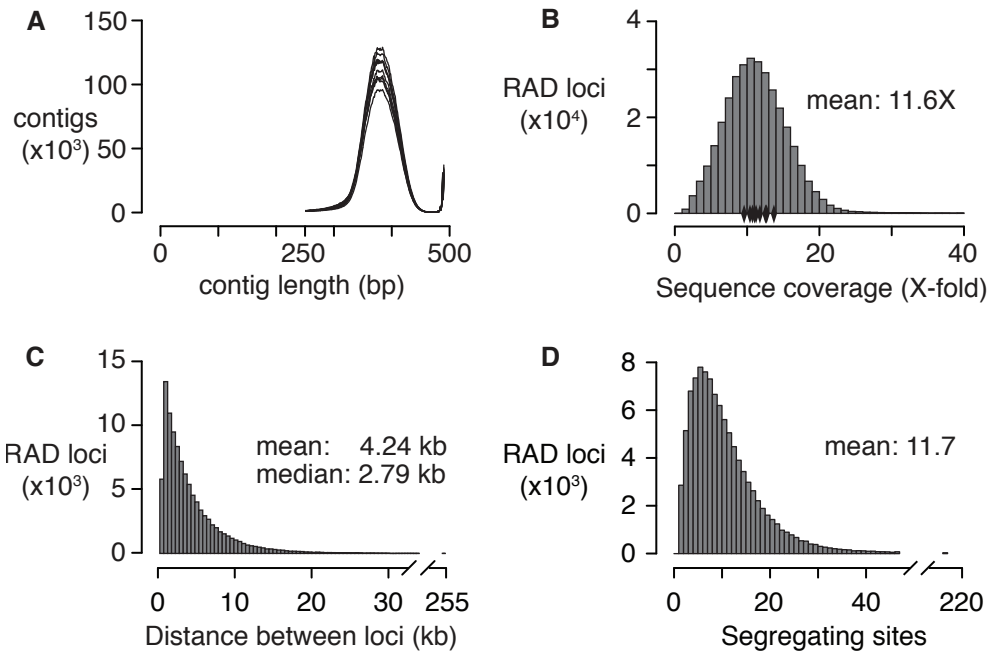

Figure 2.2. **hRAD-seq densely samples genomic diversity**. (A) Distributions of merged contig lengths for each individual in the dataset. (B) Histogram of per-locus coverage across the dataset. The histogram averages all individuals, diamonds represent mean coverage for each individual. (C) Distances between adjacent RAD across the stickleback reference genome. All phased loci are included. (D) Number of segregating sites per RAD locus.

wide averages in Rabbit Slough and Boot Lake at 0.00423 (range: [0, 0.10594]) and 0.00392 (range: [0, 0.12589]), respectively. This level of variation translated to an average of 11.7 variable sites per RAD locus (Figure 2.2D). We estimated average haplotype F_{ST} between Rabbit Slough and Boot Lake to be 0.157, which accords well with previous studies using microsatellite and RAD-seq loci (Hohenlohe, Bassham et al. 2010).

Genetic differentiation and sequence divergence between the populations were strongly and positively correlated (Figure 2.3). Absolute sequence divergence, d_{xy}, was highest in F_{ST} outlier regions, especially in genomic windows where F_{ST} exceeded 0.4 ($r^2 = 0.2215$, $p < 9 \times 10^{-240}$; Figure 2.3A). Nucleotide diversity was also correlated with F_{ST} ($r^2 = 0.0400$, $p < 2 \times 10^{-40}$; Figure 2.3B). But this relationship was much weaker and high nucleotide diversity was also common in windows of little differentiation. In contrast to among-population sequence divergence, within-population diversity was strongly and

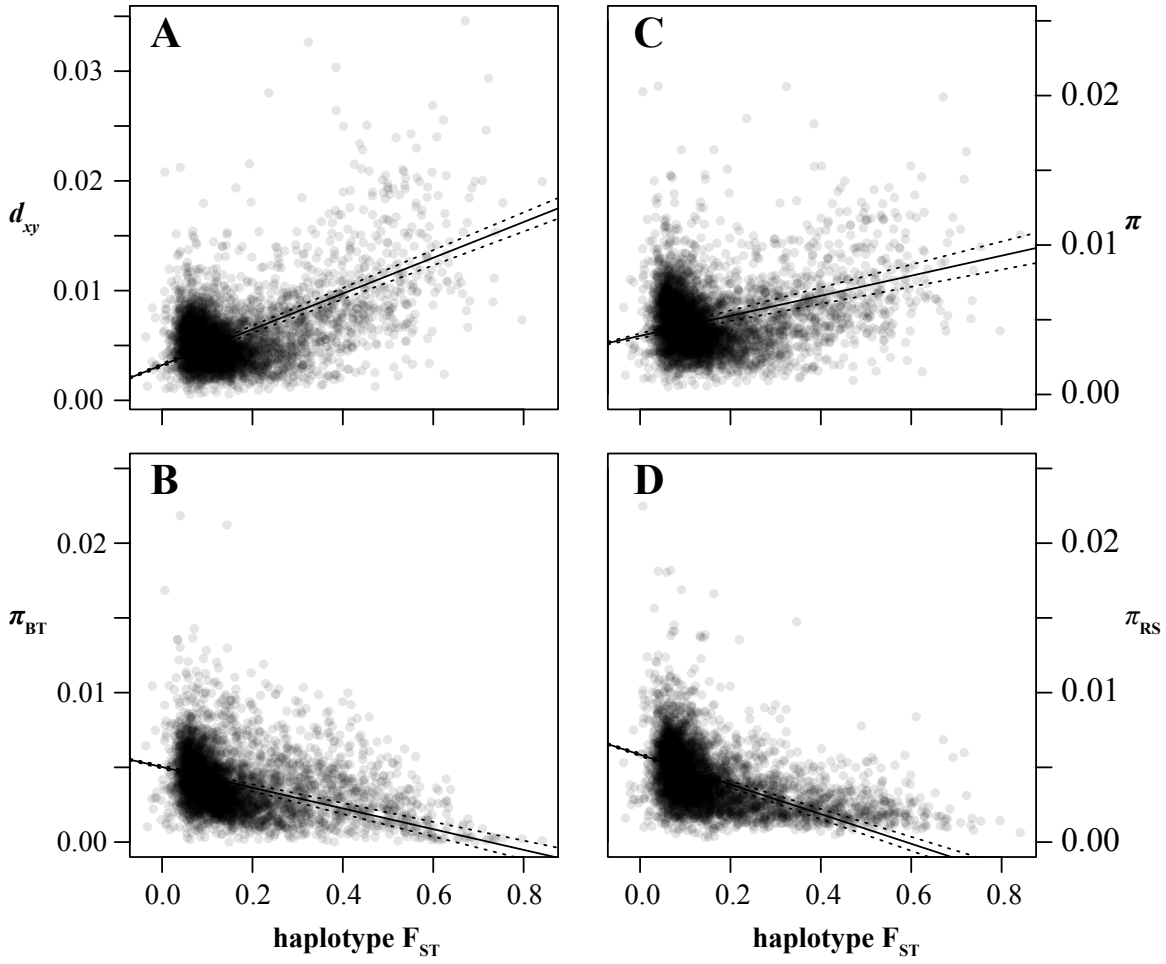

Figure 2.3. Genome-wide correlations between F_{ST} and polymorphism levels. (A) d_{xy} and (B) π are positively correlated with F_{ST} between Boot Lake and Rabbit Slough. Solid lines are fits from type-II linear models. Dashed lines represent 95% confidence intervals from 99 permutations. Within-population π in both (C) Boot Lake and (D) Rabbit Slough is negatively correlated with F_{ST} between the populations. All points represent 100 kb non-overlapping genomic windows.

negatively correlated with F_{ST} (π_{BT}: $r^2 = 0.0920$, $p < 2 \times 10^{-93}$; π_{RS}: $r^2 = 0.1870$, $p < 2 \times 10^{-198}$; Figure 2.3C,D). This effect was stronger in the oceanic population than the freshwater population (two-tailed t-test on log-transformed π: $t_{461} = -2.714$, $p < 0.007$). Nucleotide diversity in Rabbit Slough averaged 0.00239 in windows with $F_{ST} > 0.4$, while diversity in Boot Lake averaged 0.00253.

A Chromosomal Inversion is Associated With High Sequence Divergence

To illustrate the impact of genome architectural features on genetic divergence, we performed a genome scan along stickleback chromosome 1, which contains a polymorphic 400 kb inversion (Figure 2.4). This chromosome contains multiple F_{ST} peaks, with the highest smoothed F_{ST} values in the region between 8 and 10 Mb. However, the highest level of sequence divergence (d_{xy}) between these two populations occurred within the chromosomal inversion (highlighted in orange in Figure 2.4). Nucleotide diversity within both populations drops sharply within the inversion (Figure 2.4B), matching the genome-wide patterns we observed (Figure 2.3C,D).

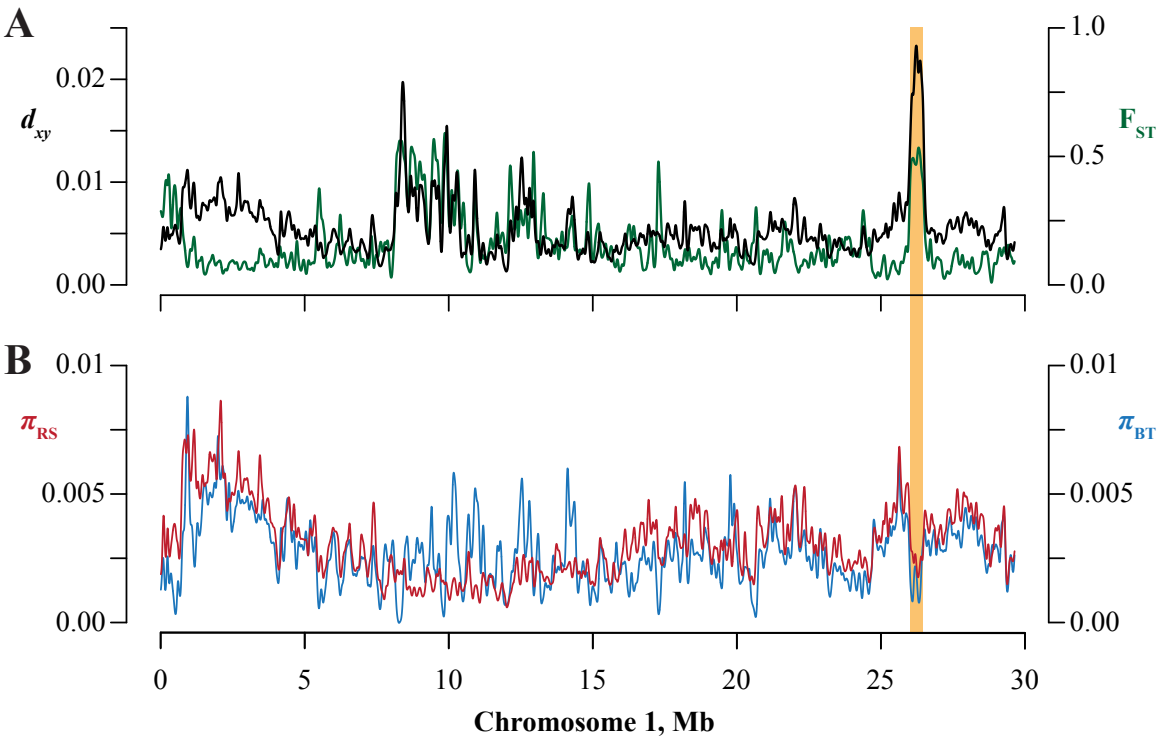

Figure 2.4. **Divergence and diversity across stickleback chromosome 1**. (A) d_{xy} and haplotype F_{ST} in 100 kb overlapping windows. (B) Within-population polymorphism (π) for Rabbit Slough, red, and Boot Lake, blue. The orange box covers a chromosomal inversion known to be divergent between freshwater and marine stickleback populations.

Adaptive Divergence is Associated with Distinct Haplogroups

To investigate genomic regions at the extreme of freshwater-oceanic divergence, we used BEAST to identify 5382 RAD loci with evidence of complete lineage sorting in either Boot Lake or Rabbit Slough, accounting for 5.2% of all loci (Figure 2.5). All three patterns of lineage sorting we examined were associated with increased levels of polymorphism over unsorted RAD loci (permutation tests, all p-values < 1 x 10-4). Loci with single clades of Boot Lake haplotypes were on average the least polymorphic of these three groups and showed very low levels of polymorphism within Boot Lake (mean $\pi = 0.00640$, mean $\pi_{BT} = 0.00046$). In fact, in over half of these loci (1019 of 1715), we recovered only a single haplotype from the Boot Lake sample ($\pi_{BT} = 0$). When we observed monophyly of Rabbit Slough haplotypes, on the other hand, overall polymorphism levels increased to almost twice the genome-wide average despite a three-

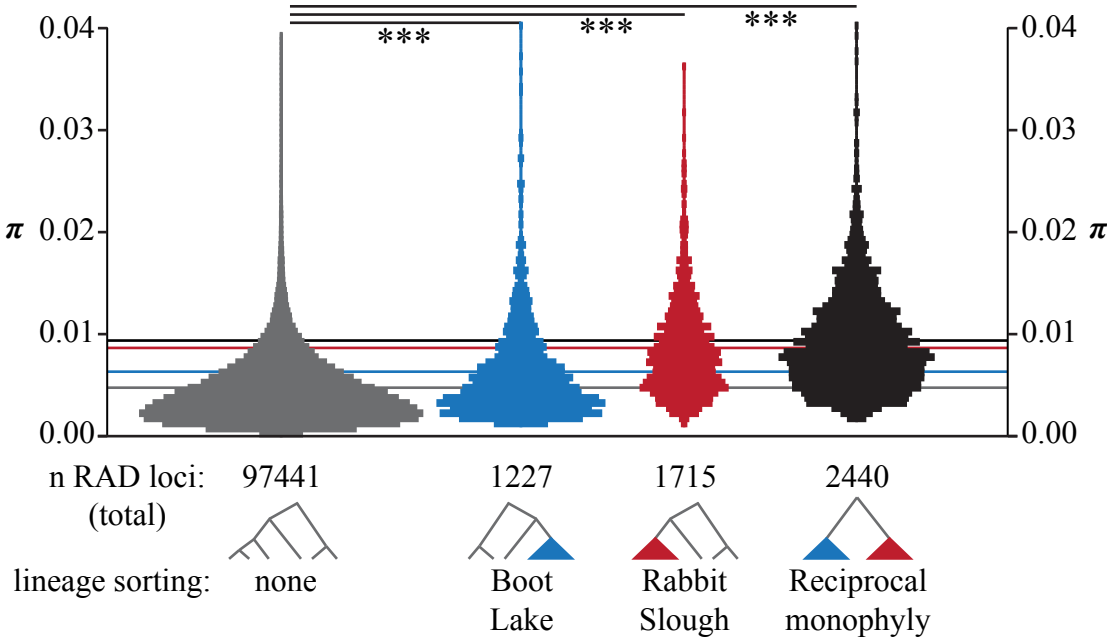

Figure 2.5. RAD loci with evidence of lineage sorting are more polymorphic. Histograms show distributions of π for all loci with evidence of a given lineage sorting pattern: gray, no sorting; blue, Boot Lake haplotypes form a clade to the exclusion of Rabbit Slough haplotypes; red, Rabbit Slough haplotypes form a clade to the exclusion of Boot Lake haplotypes; black, haplotypes from Rabbit Slough and Boot Lake are reciprocally monophyletic. Cartoon examples of each pattern are shown below the distributions, along with the number of loci in each category. *** $p < 1 \times 10^{-4}$ (permutation test).

fold reduction in diversity in Rabbit Slough compared to the genomic average (mean π: 0.00919, mean π_{RS} = 0.00142).

Loci showing reciprocal monophyly between Rabbit Slough and Boot lake haplotypes were the single most common and most polymorphic pattern we observed (n = 2440, mean π = 0.00949, black histogram in Figure 2.5). This class of loci included those adjacent to the *Ectodysplasin A* (*Eda*) gene (Figure 2.6A, gold lines in Figure 2.6B), which is known from targeted sequencing studies to have distinct oceanic and freshwater haplogroups (Colosimo, Hosemann et al. 2005). Almost universally,

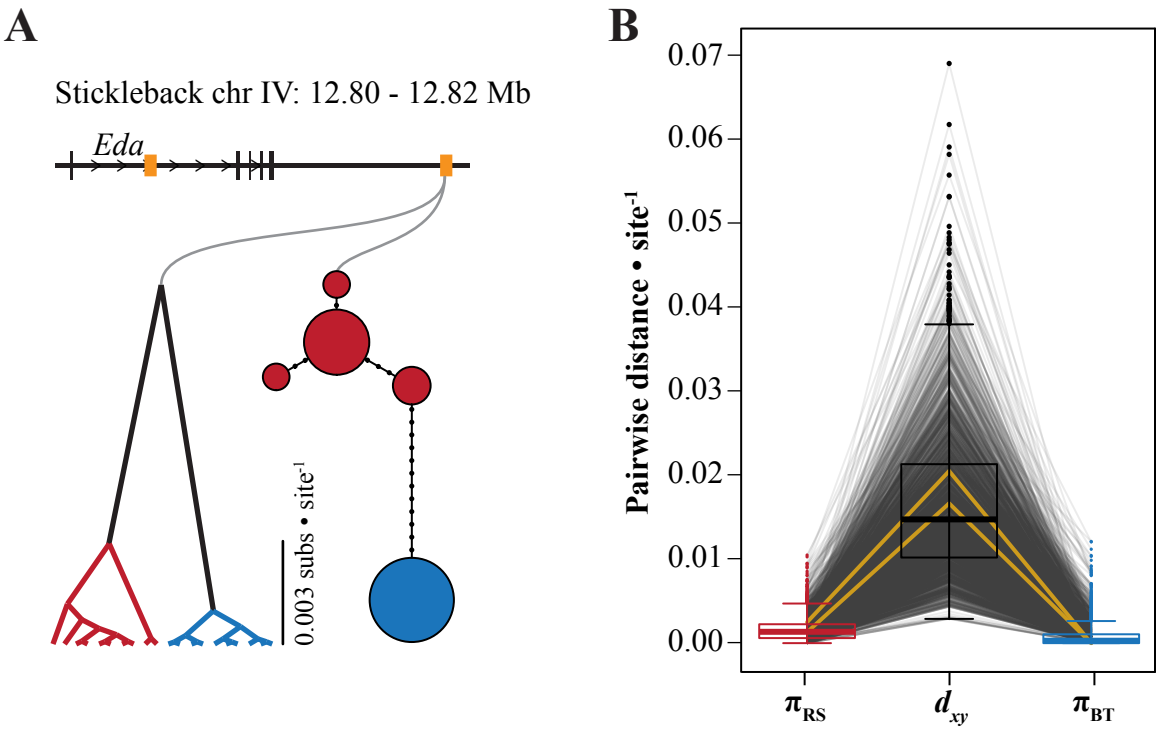

Figure 2.6. Reciprocal monophyly at RAD loci is associated with deep divergence between and low polymorphism within haplotypes from Rabbit Slough and Boot Lake. (A) A RAD locus 10 kb downstream of the *Eda* coding region shows a stereotypical pattern reciprocal monophyly, corroborating previous work on this gene (Colosimo *et al.*, 2005). The same locus is shown as a haplotype network at right. Circles are proportional to the number of sampled haplotypes. Red, Rabbit Slough; blue, Boot Lake. Black vertical bars represent exons of *Eda*; arrows indicate the direction of transcription. Orange boxes represent RAD loci identified in out dataset. (B) Within-population π and d_{xy} for all reciprocally monophyletic RAD loci. Within-population boxplots are colored by population. Gray lines connect divergence estimates for individual loci. The y-axis is scaled to highlight divergence among the majority of RAD loci.

reciprocally monophyletic loci showed high levels of sequence divergence among populations (mean d_{xy} = 0.01695) and reduced levels of within-population polymorphism (mean π_{BT} = 0.000798, mean π_{RS} = 0.001618). Notably, 28% (n = 691) of reciprocally monophyletic loci had levels of sequence divergence between marine and freshwater haplotypes that exceeded those we observed within and around *Eda*.

DISCUSSION

Anciently Derived Variation Drives Recent Adaptive Divergence

Adaptive evolution of stickleback to the freshwater environment of Boot Lake involved selection on standing genetic variants with ancient ancestry. Genome-wide genetic differentiation between Rabbit Slough and Boot Lake was moderate and comparable to previous estimates of these same populations (Hohenlohe, Bassham et al. 2010). We found that peaks of genetic differentiation between these populations were largely associated with increased sequence divergence between marine and freshwater allelic variants as measured by d_{xy} and sharp reductions in sequence variation in both populations (Figure 2.3). The extent of between-population sequence divergence is notable because Boot Lake formed no earlier than the glacial retreat at the end of the Pleistocene 9,000 to 14,000 years ago (Reger and Pinney 1996), or approximately 10,000 stickleback generations ago. Despite this young age, average d_{xy} exceeded 0.02 in some genomic regions, an order of magnitude greater than that observed in other taxa showing similar levels of F_{ST} genome-wide (Burri, Nater et al. 2015) or comparable peaks of F_{ST} in divergent genomic regions (Malinsky, Challis et al. 2015). The extent of sequence divergence we observed across many of these regions cannot, therefore, be explained by the fixation of new mutations. Rather, those adaptive mutations were pre-existing and, importantly, occurred long prior to the colonization of Boot Lake.

High levels of sequence divergence between freshwater and oceanic alleles was expected in some regions of the genome from previous targeted sequencing studies. Work on the *Eda* locus in stickleback showed clearly that independently derived freshwater populations had fixed alleles at that locus that were identical-by-descent and only

distantly related to the sequence common in marine populations (Colosimo, Hosemann et al. 2005). We found a very similar gene tree structure using hRAD-seq, confirming that the freshwater population in Boot Lake is fixed for an anciently derived *Eda* variant. We also identified over 2000 other reciprocally monophyletic RAD loci like those surrounding *Eda*. Our genome-wide data suggest that *Eda* is not an outlier from the standpoint of freshwater adaptation, but is in fact part of a larger suite of loci associated with adaptive divergence in stickleback. Moreover, a number of RAD loci showed sequence divergence between Boot Lake and Rabbit Slough that exceeded that observed at RAD loci adjacent to *Eda*, suggesting that many alleles currently contributing to freshwater adaptation evolved long before the major effect allele contributing to the phenotype that is now a defining feature of most freshwater stickleback populations.

Echoing a growing number of empirical studies, our data also support an important role for chromosomal inversions in maintaining adaptive differentiation (Kirkpatrick and Barton 2006, Lowry and Willis 2010, Joron, Frezal et al. 2011). The inversion on stickleback chromosome 1 is one of three known to be involved in marine-freshwater divergence (Jones, Grabherr et al. 2012, Roesti, Gavrilets et al. 2014, Roesti, Kueng et al. 2015). Allele frequency divergence within this region was comparable to other F_{ST} peaks along chromosome 1, reflecting its adaptive role in the Boot Lake population. Importantly, however, absolute sequence divergence within the inversion was higher than in any other window along the chromosome and was in the top 1% of windows genome-wide. The ocean and freshwater associated forms of the inversion are therefore very distinct in sequence space from one another as would occur if both were maintained in the global stickleback population through differential local selection in the two different habitats. Coupled with strong reductions in within-population polymorphism, especially in Boot Lake, these data together point to the maintenance of anciently derived, alternatively adaptive haplotypes within this species.

Divergent Selection Maintains Alternative Marine and Freshwater Haplogroups

Patterns of sequence variation at a number of RAD loci suggest that the same genomic regions are under selection in both the freshwater and marine stickleback populations. When we searched for loci with strong evidence of lineage sorting, we commonly found monophyletic clades of marine and freshwater alleles separated by extensive sequence divergence. Low polymorphism within Boot Lake at these loci is likely the result of selective sweeps during colonization of the new habitat, and the observed divergence from marine alleles could point to the marine stickleback population as a long-term reservoir of variation. However, we observed strong reductions in sequence diversity in the marine population as well, which cannot be explained by adaptation to freshwater alone. We hypothesize that these loci are under continuous selection toward an alternate fitness optimum in marine stickleback populations. Relative to the freshwater forms, marine stickleback are phenotypically homogenous across large geographic distances, especially among phenotypes that evolve rapidly in freshwater populations (Bell and Foster 1994, Catchen, Bassham et al. 2013, Feulner, Chain et al. 2013). Linkage disequilibrium is also surprisingly high in marine populations given their large census population size (Hohenlohe, Bassham et al. 2012). Marine populations may therefore be adapting toward a relatively stable fitness optimum. Rather than being reservoirs of standing genetic variation, selection in marine stickleback populations is likely acting to constantly remove variation introduced through hybridization with adjacent freshwater populations.

Our results suggest that genomic divergence in threespine stickleback is driven primarily by genetic and genomic variants that evolved anciently and have been retained within the species for long periods of time, clearly supporting a metapopulation structure for the species as a whole (Schluter and Conte 2009, Lescak, Bassham et al. 2015). A leading hypothesis for the rapidity and repeatability of freshwater divergence in stickleback is that allelic variants are shuttled among freshwater populations through the marine population via hybridization (Schluter and Conte 2009). While our data do not provide direct support for gene flow among freshwater populations, they highlight

standing genetic variation as the primary driver of adaptation in this system. Combined with evidence of selection in the marine population favoring alternative alleles at the same loci, these data suggest that the reservoirs of genetic variation in the threespine stickleback are the many freshwater populations connected to each other through the ocean. Moreover, the extensive sequence evolution we observed between marine and freshwater allelic variants is evidence for this being an ancient process, with recent and rapid bouts of phenotypic evolution underlain by millions of years of genome evolution.

BRIDGE

Adaptive evolution in the recently founded Boot Lake population drew heavily from standing genetic variation that arose long ago, having been maintained by divergent natural selection. I focused this analysis on regions of the genome with strong signatures of selection, looking for common patterns in their molecular evolutionary histories, but these results have implications for patterns of adaptation across geography and patterns of variation across the genome. The maintenance of adaptive variation suggests that separate populations adapting to similar environments may be more likely to do so using the same ancestral pool of genetic variation, increasing identity-by-descent among otherwise unrelated populations. The effects of such long-term selection on patterns of linked genomic variation are also relatively unexplored in natural systems. In Chapter III, I expand these analyses by including a second, geographically isolated freshwater population and by explicitly estimating the time to the most recent common ancestor of allelic variation within and among populations in adaptive, linked, and unlinked genomic regions. These results show the extent to which natural selection has shaped genomic variation over the course of threespine stickleback evolution.

CHAPTER III

A LINK TO THE PAST: NATURAL SELECTION AND POPULATION STRUCTURE INTERACT TO MAINTAIN GENOMIC DIVERSITY IN STICKLEBACK

INTRODUCTION

Understanding how genetic variation is maintained and partitioned within and among natural populations is central to our understanding of evolution and adaptation. Genetic variation is partitioned across geography, often among multiple interconnected populations (Hohenlohe, Bassham et al. 2010, Corbett-Detig, Zhou et al. 2013, Stankowski, Sobel et al. 2015), and distributed non-randomly across the physical expanse of the genome (Begun, Holloway et al. 2007). When selection pressures vary across geography, local adaptation can maintain variation at loci under divergent selection and in linked genomic regions (Charlesworth, Nordborg et al. 1997). The impacts of local adaptation and population structure on levels of genetic variation within and among populations can vary drastically across the genome (Nosil and Feder 2012), but how these processes interact to produce observed patterns of variation remains poorly understood.

Divergent selection effectively partitions variation among selective environments, but we know less about its ability to maintain variation in natural systems over longer evolutionary timescales (Charlesworth, Nordborg et al. 1997). Theoretical models predict that divergent natural selection can maintain polymorphism essentially indefinitely, as long as alternative habitats exist and selection is strong enough to resist the homogenizing effects of gene flow (Charlesworth, Nordborg et al. 1997, Lenormand 2002, Guerrero, Rousset et al. 2012). A growing number of examples of divergent selection in the wild (Yatabe, Kane et al. 2007, Lowry, Rockwood et al. 2008, Jones, Grabherr et al. 2012, Nadeau, Whibley et al. 2012, Nosil and Feder 2012, Roesti, Gavrilets et al. 2014) argue that studies of adaptation from standing genetic variation should incorporate this potential deep history of adaptive genetic and genomic variants. These studies often examine the effects of divergent selection on differentiation at single nucleotide polymorphism (SNP) loci between populations, which, although useful in

quantifying genomic divergence among populations or across geography, provide little information about the evolutionary history of genomic variation affected by selection.

The maintenance of adaptive polymorphism within a species may also increase the probability of parallel evolution in selectively similar environments. In contrast to adaptive divergence, in which populations respond to variable selective pressures across geography, isolated populations can also experience common environments and common selective pressures. While parallelism at the phenotypic level may be driven by entirely distinct genetic mechanisms (Arendt and Reznick 2008), evidence from multiple natural systems suggests that the independent evolution of similar phenotypes among populations of the same species, or among pairs of closely related species, involves the same genes (Shapiro, Marks et al. 2004, Chan, Marks et al. 2010, Hohenlohe, Bassham et al. 2010, Jones, Grabherr et al. 2012, Roda, Liu et al. 2013, Roesti, Gavrilets et al. 2014, Soria-Carrasco, Gompert et al. 2014) and even the same alleles (Colosimo, Hosemann et al. 2005, Stankowski and Streisfeld 2015, Van Belleghem, Rastas et al. 2017). By maintaining genetic variation for fitness-related phenotypic traits, ongoing selection has the potential to promote parallelism *via* alleles that are identical-by-descent (IBD). In a recent example of parallelism in the butterfly genus *Heliconius*, wing pattern variation shared among multiple species was found to be associated with common haplotypes in enhancer regions downstream of the *optix* gene (Wallbank, Baxter et al. 2016). Haplotypes originating approximately two million years ago have been shared *via* gene flow between multiple species in the genus, some of which diverged around four million years ago.

The effects of natural selection extend beyond loci directly under selection and into linked genomic regions (Maynard Smith and Haigh 1974, Nosil and Feder 2012). Within a population, positive selection reduces or eliminates variation at the locus under selection (Maynard Smith and Haigh 1974). This effect extends into adjacent genomic regions as a function of the strength of selection and the recombination rate between the selected and linked sites. Among populations in alternative selective environments, divergent selection counteracts the homogenizing effects of gene flow and recombination

to maintain divergence in linked genomic regions. However, the ability of divergent selection alone to maintain linked variation is uncertain. Single-locus models can result in rapid erosion of divergence away from the locus under selection and little contribution to maintenance of linked variation (Charlesworth, Nordborg et al. 1997). Multilocus models, on the other hand, often result in extensive divergence at linked loci (Feder and Nosil 2010, Feder, Gejji et al. 2012, Via 2012), which may provide a platform for longer-term maintenance of variation among populations. Finally, the effects of parallel adaptation on patterns of linked variation are largely uncharacterized. Maintenance of variation among populations could result from the fixation of different *de novo* mutations or adaptation from a diverse pool of standing variation (Hermisson and Pennings 2005, Chan, Marks et al. 2010). Parallel adaptation may also increase identity-by-descent among populations if shared adaptive alleles exist on one or a few genetic backgrounds, as has been demonstrated in instances of adaptive introgression (Huerta-Sánchez, Jin et al. 2014, Stankowski and Streisfeld 2015).

Here, we investigate how divergent natural selection and population structure affect coalescence times throughout the genome of the threespine stickleback fish, *Gasterosteus aculeatus*. We examine patterns of genomic variation and estimate the time to the most recent common ancestor (T_{MRCA}) of alleles within and among locally adapted populations to answer the following questions: *(1) How long have neutral and adaptive genomic regions been maintained? (2) Does parallel adaptation involve alleles that are identical-by-descent (IBD) among independently derived populations? (3) How does natural selection affect the maintenance of linked genomic variation within and among populations experiencing similar and divergent selective pressures?*

The threespine stickleback is a holarctically distributed species inhabiting coastal oceanic habitats and freshwater lake and river systems (Bell and Foster 1994). Oceanic and freshwater ecotypes are locally adapted, with stereotypical differences in anti-predator defenses, trophic morphologies, behavior, and immunology (McKinnon and Rundle 2002), which together reflect a high degree of genomic differentiation between oceanic and freshwater forms (Hohenlohe, Bassham et al. 2010, Jones, Grabherr et al.

2012). Oceanic and freshwater ecotypes also experience differing population structures: stickleback in the oceanic habitat approximate a large, panmictic population, with little genetic differentiation across thousands of kilometers, while freshwater habitats are smaller and populations experience greater physical and genetic isolation (Hendry, Taylor et al. 2002, Catchen, Bassham et al. 2013).

In Cook Inlet in south central Alaska, oceanic stickleback colonized hundreds of freshwater lakes after the glacial retreat at the end of the Pleistocene glaciation approximately 12,000 years ago (Cresko, Amores et al. 2004). Populations in these lakes adapted in parallel phenotypically and genomically to a striking degree, likely from a common pool of standing genetic variation present in the ancestral oceanic population. Previously, we demonstrated that adaptive variation differentiating a freshwater population in Boot Lake from the neighboring oceanic population in Rabbit Slough is not only older than the Boot Lake population, but is older than the majority of the segregating variation throughout the rest of the stickleback genome (Chapter II of this work). These data suggested that oceanic-freshwater genomic differentiation has been maintained over long periods of time, but the questions of how it has been maintained, and for how long, remain mostly unanswered.

In the present study, we describe the genealogical patterns associated with divergent and parallel adaptation by comparing genomic divergence in the Boot Lake population to a second, independently derived freshwater population in Bear Paw Lake (see Figure 1.2). We use restriction site-associated DNA sequencing to construct local haplotypes genome-wide (hRAD-seq) and collect DNA sequence information within and among populations at tens of thousands of RAD loci (Chapter II of this work). We also performed a *de novo* genome assembly of a single ninespine stickleback, *Pungitius pungitius*, to estimate sequence divergence between threespine and ninespine stickleback and, based on a split time of 15 million years (Aldenhoven, Miller et al. 2010), estimate T_{MRCA}, in years, of alleles sampled at RAD loci. We use these data to address the questions stated above and provide the first genome-wide view of the timescale of marine-freshwater divergence in the stickleback.

METHODS

Study Populations and Sample Collection

Wild threespine stickleback were collected from Rabbit Slough (N 61.5595, W 149.2583), Boot Lake (N 61.7167, W 149.1167), and Bear Paw Lake (N 61.6139, W 149.7539). Rabbit Slough is an offshoot of the Knik Arm of Cook Inlet and is known to be populated by anadromous populations of stickleback that are stereotypically oceanic in phenotype and genotype (Cresko, Amores et al. 2004). Boot Lake and Bear Paw Lake are both shallow lakes formed during the end-Pleistocene glacial retreat. Fish were collected in the summers of 2009 (Rabbit Slough), 2010 (Bear Paw Lake), and 2014 (Boot Lake) using wire minnow traps and euthanized *in situ* with Tricaine solution. Euthanized fish were immediately fixed in 95% ethanol and shipped to the Cresko Laboratory at the University of Oregon (Eugene, OR, USA).

Threespine Stickleback RAD Library Preparation and Sequencing

We generated restriction site-associated DNA (RAD) libraries of five fish each from Rabbit Slough and Boot Lake and four fish from Bear Paw Lake. Genomic DNA was isolated from ethanol-preserved fin clips by proteinase K digestion followed by DNA extraction with Ampure magnetic beads. We created RAD libraries using the single digest and shearing method of Baird *et al* (2008) with the modifications of Nelson and Cresko (Chapter II of this work). High molecular weight genomic DNA from each fish was digested with PstI-HF (New England Biolabs) and ligated to Illumina P1 adaptors with 6 bp inline barcodes. All barcodes differed by at least two positions, allowing for recovery of sequence reads with single errors in the barcode sequence. Ligated samples were then multiplexed at approximately equimolar concentrations and mechanically sheared via sonication to a fragment range of ~200-800 bp. Sheared DNA was size selected by extraction from a 1.25% agarose gel to generate a narrow insert size range of 425 bp to 475 bp. This size range allowed consistent overlap of paired-end Illumina reads for the construction of local contigs surrounding restriction enzyme cut sites. We then ligated Illumina P2 adaptors to the size-selected libraries and and amplified P1/P2-adapted

fragments with 12 cycles of PCR using Phusion-HF polymerase (New England Biolabs). RAD libraries were then sequenced in a single lane on an Illumina HiSeq 2500 to generate paired-end 250-bp sequence reads. All libraries generated for this study were sequenced at the University of Oregon's Genomics and Cell Characterization Core Facility (GC3F: http://gc3f.uoregon.edu).

Ninespine Stickleback Genome Assembly

In order to estimate the T_{MRCA} of threespine stickleback RAD alleles, we used the ninespine stickleback (*Pungitius pungitius*) as an outgroup (Figure 3.1, see Figure 1.2). RAD sequence analysis, however, relies on the presence of homologous restriction sites among sampled individuals and results in null alleles when mutations occur within a restriction site (Arnold, Corbett-Detig et al. 2013). Because this probability increases with greater evolutionary distance among sampled sequences, we elected to use RAD-seq to only estimate sequence variation within the threespine stickleback. We then generated a contig-level *de novo* ninespine stickleback genome assembly from a single ninespine stickleback individual from St. Lawrence Island, Alaska (collected by J. Postlethwait) using DISCOVAR *de novo* (https://software.broadinstitute.org/software/discovar). We used this single ninespine stickleback haplotype to estimate threespine-ninespine sequence divergence and time calibrate coalescence times within the threespine stickleback. DISCOVAR de novo requires a single shotgun library of

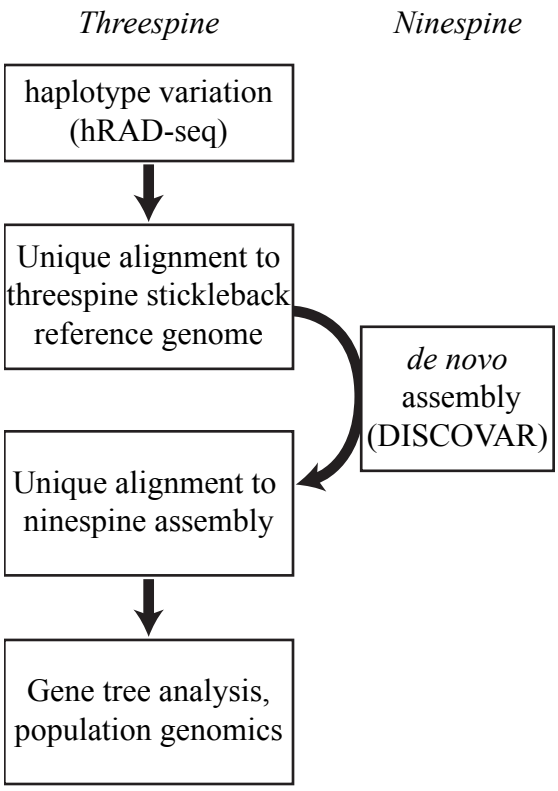

Figure 3.1. Flowchart for threespine stickleback RAD locus assembly and alignment of threespine stickleback consensus loci to the ninespine stickleback genome assembly.

paired-end 250-bp sequence reads from short-insert-length DNA fragments. High molecular weight genomic DNA was extracted from an ethanol-preserved fin clip by proteinase K digestion followed by DNA extraction with Ampure magnetic beads. Purified genomic DNA was mechanically sheared by sonication and size selected to a range of 200-800 bp by gel electrophoresis and extraction. We selected this fragment range to agree with the recommendations for *de novo* assembly using the DISCOVAR *de novo* (https://software.broadinstitute.org/software/discovar/blog). This library was sequenced on a single lane of an Illumina HiSeq2500 at the University of Oregon's Genomics and Cell Characterization Core Facility (GC3F: https://gc3f.uoregon.edu/).

We assembled the draft ninespine stickleback genome using DISCOVAR de novo. Raw sequence read pairs were first quality filtered and adaptor sequence contamination removed using the program **`process_shortreads`**, which is included in the Stacks analysis pipeline (Catchen, Hohenlohe et al. 2013). Because DISCOVAR requires a single fastq file with interleaved paired-end reads, we used a custom Python script to interleave the filtered first- and second-end sequence files. We ran the genome assembly on the University of Oregon's Applied Computational Instrument for Scientific Synthesis (ACISS: http://aciss-computing.uoregon.edu).

Sequence Processing and SNP Discovery

We used the Stacks analysis pipeline to process RAD sequence read pairs and call SNPs (Catchen, Hohenlohe et al. 2013). Raw reads were first demultiplexed without quality filtering using **`process_radtags`**, and then quality filtered using **`process_shortreads`**. This allowed for read trimming, rather than strict removal, if quality decreased toward the end of the first-end read. Overlapping read pairs were then merged using **`fastq-join`** (Aronesty 2011), allowing for up to 25% of bases in the overlapping region to mismatch, and the resulting contigs were trimmed to 350 bp. Any read pairs that failed to merge, or were shorter than 350 bp, were removed from further analysis. This step was required for processing reads through the Stacks pipeline. Below, we use the naming conventions of Baird et al. (2008): A "RAD tag" refers to sequence

generated from a single end of a restriction site and the pair of RAD tags sequenced at a restriction site comprises a "RAD locus".

All polymorphisms were called relative to the threespine stickleback reference genome v1.0 (Jones, Grabherr et al. 2012), using the updated scaffolding of Glazer, et al. (2015). Trimmed contigs were aligned to the reference using **`bbmap`** with the most sensitive settings ('**`vslow=t`**'; http://jgi.doe.gov/data-and-tools/bbtools/). We then used the Stacks core pipeline to identify read stacks, call SNPs, and identify alleles and haplotypes based on genomic alignment (**`pstacks`** and **`cstacks`**); find homologous RAD tags across individuals (**`sstacks`**); and catalog biologically plausible haplotypes based on within- and among-individual haplotype variation (**`populations`**). We required that a RAD tag be present in all three populations and in at least four fish in each population.

We used the program PHASE to generate phased haplotypes at each RAD locus that include sequence information from both RAD tags. We wrote custom Python scripts to identify all unique haplotypes at each of a pair RAD tags and code them as alleles at a single, multiallelic locus. We required that each individual included in this analysis was genotyped at both RAD tags. Loci containing individuals only genotyped at a single RAD tag were removed from further analysis.

Once complete RAD loci were identified, we incorporated the ninespine genome sequence by aligning the consensus sequence at each threespine stickleback RAD locus to the ninespine assembly. We aligned sequences with **`bbmap`**, using the most sensitive alignment settings ('**`vslow=t`**'), and output the alignments in SAM format. We then used alignment specifications in the SAM file to mutate all positions in the stickleback consensus sequence that differ from the ninespine reference to those present in the ninespine reference.

Population Genetics and Time to the Most Recent Common Ancestor

The scripting language R (R Core Team 2016), and packages written for it, was used to compute population genetic statistics. We estimated differentiation among

threespine stickleback populations (all pairwise combinations) and among ecotypes (combined freshwater ponds versus Rabbit Slough) using a haplotype-based F_{ST} (equation 3 in Hudson, Slatkin et al. 1992) implemented in the R package 'PopGenome' (Pfeifer, Wittelsburger et al. 2014). We calculated π per site within and among populations at each RAD locus by calculating pairwise distances among all RAD haplotypes with the R package 'ape' (Paradis, Claude et al. 2004).

We used BEAST v. 1.7 (Drummond and Rambaut 2007, Drummond, Suchard et al. 2012) to identify evidence of lineage sorting within and among populations and to estimate divergence times of alleles at each RAD locus. BEAST analyses were performed on each RAD locus independently and with blanket parameters and priors across all loci. We specified Markov Chain Monte Carlo (MCMC) runs of 1,000,000 states and logged trees and parameter estimates every 100 states. In all following analyses, we regarded the first 10% of the MCMC chain as burn-in. We specified a coalescent tree prior, the GTR +Γ substitution model with four rate categories, and uniform priors for all substitution rates. We identified evidence of lineage sorting by analyzing all logged trees using R and the package 'ape' (Paradis, Claude et al. 2004). For each tree, we used the function `is.monophyletic()` to determine monophyly of alleles at the population (Boot Lake, Bear Paw Lake, Rabbit Slough), habitat (oceanic, freshwater), and species (threespine) levels and calculated the proportion of sampled gene trees with evidence of each pattern. In the data that follow, we used a cutoff of 50% of sampled gene trees as evidence of a particular pattern of lineage sorting, but genome-wide data were qualitatively similar using more stringent criteria.

To convert node ages estimated in BEAST into divergence times, in years, we assumed a 15 million-year divergence time between threespine and ninespine stickleback at each RAD locus (Aldenhoven, Miller et al. 2010). The T_{MRCA} of all alleles in each gene tree was set at 15 Mya at each node age of interest was converted into years relative to the total height of the tree. Because the facts of the genealogical process mean that the true T_{MRCA} at any locus likely differs from the 15 My estimate (Kingman 1982, Kingman 1982, Tajima 1983), in the data that follow we do not rely heavily on T_{MRCA} estimates at

individual RAD loci. Rather, we use these estimates to understand patterns of broad patterns of ancestry throughout the threespine stickleback genome — along and among chromosomes and genome-wide.

Threespine-Ninespine Haplotype Sharing and Gene Flow

We identified evidence of haplotype sharing between threespine and ninespine stickleback in two ways. We first screened for RAD loci with gene trees that did not support monophyly of threespine stickleback alleles. We used the R package 'ape' to analyze all gene trees sampled from BEAST MCMC runs from each RAD locus, using the is.monophyletic() function to identify monophyly of threespine stickleback alleles. In the data we present here, we considered it evidence of haplotype sharing if greater than 50% of gene trees sampled lacked threespine stickleback monophyly, although results across the entire dataset were quantitatively and qualitatively similar when we increased stringency to 90% of sampled gene trees.

To differentiate among the potential causes of haplotype sharing — incomplete lineage sorting versus post-speciation gene flow — we compared polymorphism among threespine stickleback RAD alleles to the level of sequence divergence to the ninespine haplotype. Incomplete lineage sorting would result from at least one coalescence event among threespine stickleback alleles predating the threespine-ninespine divergence, leading to elevated sequence diversity within threespine stickleback and elevated threespine-ninespine sequence divergence. Recent gene flow would lead to a contrasting pattern, with low threespine-ninespine sequence divergence irrespective of levels of polymorphism within threespine stickleback. BEAST analyses may also have lacked evidence for threespine stickleback monophyly due to a complete lack of polymorphism, and thus a lack of phylogenetic signal, across all sampled alleles.

RESULTS

Ninespine Stickleback Genome Assembly and Threespine-Ninespine Alignments

The assembled ninespine genome was 484.8 Mb in length with 99.97% of bases called (0.03% gaps [167.6 kb]). This is comparable to, but slightly larger than, the published 463-Mb threespine stickleback genome assembly (Jones, Grabherr et al. 2012). Because we used a single library with insert sizes ranging from ~200-800 bp, we expected very few gaps in the assembly but a relatively large number of short scaffolds. The assembly consisted of 239264 scaffolds (240940 contigs) greater than 250 bp in length. We obtained a scaffold N50 of 11.6 kb (contig N50: 10.6 kb), a maximum scaffold length of 193.1 kb (contig max: 160.2 kb), and mean scaffold and contig lengths of 2.0 kb.

We successfully recovered 58,087 threespine stickleback RAD loci that aligned uniquely to both the threespine and ninespine genome assemblies. These loci comprised 68.4% of a total of 84,975 RAD loci uniquely aligned to the threespine stickleback reference and genotyped in all threespine stickleback populations. The average distance between adjacent RAD loci on the threespine stickleback genome was 8.0 kb.

Adaptive Divergence was Associated with Ancient Ancestry Across the Stickleback Genome

Ancestry across the stickleback genome varied widely within and among populations (Figure 3.2). The genome-wide average T_{MRCA} across all three stickleback populations was just over 3 million years ago (Mya) (3.79 Mya, median = 3.33 Mya) with 90% of RAD loci estimated to be between 1.89 and 7.49 million years old (range: 1.03 Mya - 12.88 Mya). Individual populations had somewhat more recent ancestry genome-wide. When we partitioned our dataset based on sampling location, variation within the oceanic Rabbit Slough population had the highest average T_{MRCA} at 3.34 million years old followed by the freshwater populations, Boot Lake and Bear Paw Lake, which averaged 3.14 and 2.77 Mya, respectively (ANOVA on $\log_2$-transformed T_{MRCA}:

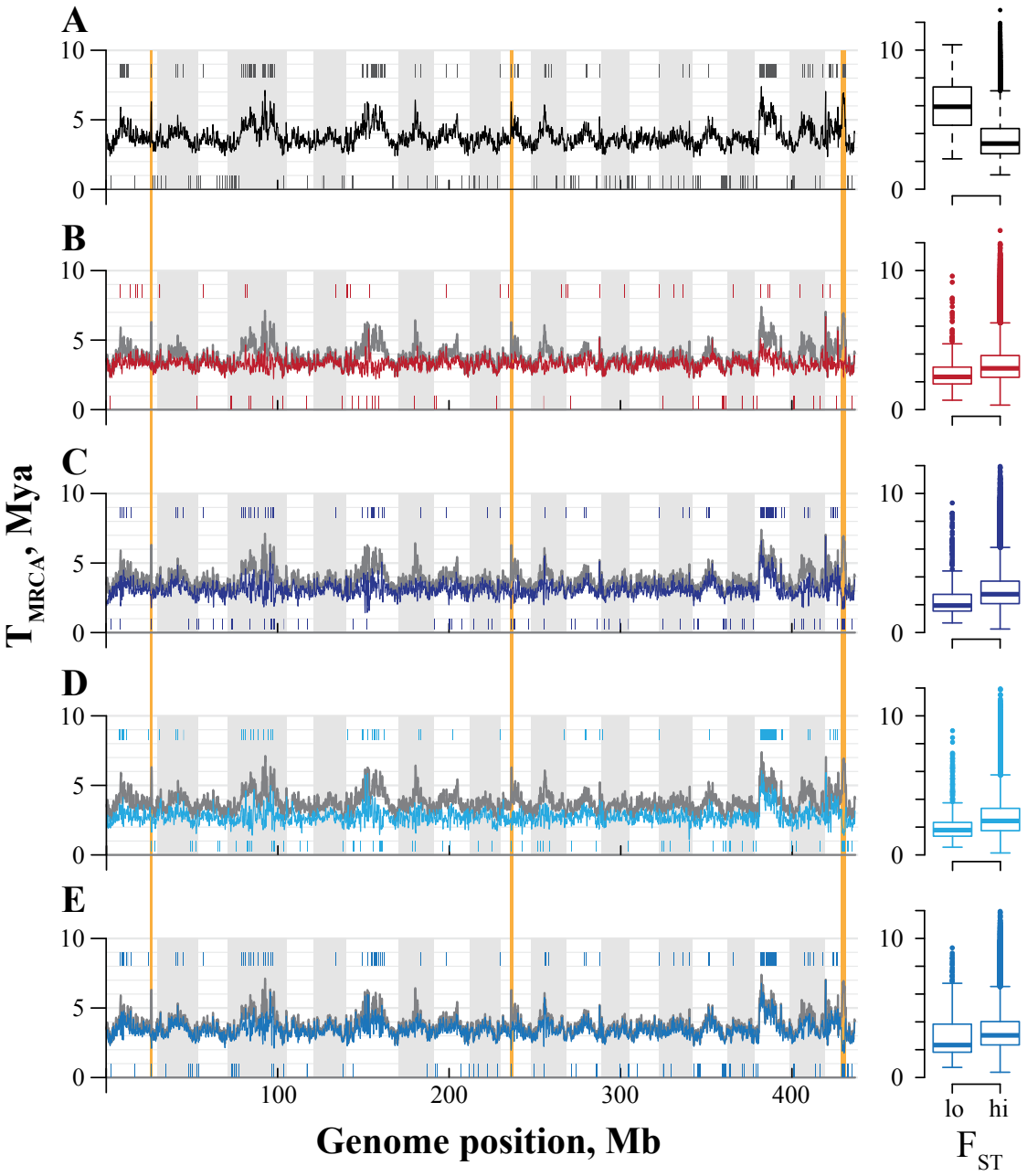

Figure 3.2. Time to the most recent common ancestor (T_{MRCA}) of alleles across the threespine stickleback genome. Genome scans show 21 chromosomes are shown in order. Chromosome 19 is the heteromorphic sex chromosome. Boxplots show T_{MRCA} of RAD loci in the top 1% of marine-freshwater F_{ST}. (A) All populations combined; (B) Rabbit Slough (marine); (C) Boot Lake (freshwater); (D) Bear Paw Lake (freshwater); (E) combined freshwater populations. Bars above and below scans are intervals where smoothed T_{MRCA} is significantly high or low, respectively, based on permuted 99.9% confidence bands. Yellow vertical bars highlight three known chromosomal inversions.

$F_{2,142800} = 3289$, $p < 10^{-10}$, Tukey's HSD: $p < 10^{-10}$ for all comparisons). In contrast, when we partitioned our dataset based on habitat, the freshwater populations combined had greater average T_{MRCA} (3.48 Mya) than the oceanic population (t-test on $\log_2$-transformed T_{MRCA}: $t_{94652} = 12.8$, $p < 10^{-10}$). In all populations, we observed a long tail RAD loci with deep coalescence times extending to over 12 Mya, or approximately 4X older than the genome-wide average.

Regions of the stickleback genome associated with adaptive divergence to freshwater ponds were also, on average, the oldest genomic regions apart from the sex chromosome (Figure 3.2A). Genomic regions with elevated F_{ST} in oceanic-to-freshwater comparisons were associated with clear increases in T_{MRCA}, which often extended over many megabases. The extent of this relationship, however, varied among chromosomes and even among regions within a chromosome. Chromosomes 4 and 7, which are commonly involved in adaptive divergence globally, had increased T_{MRCA} along much of their lengths (Figures 3.2 and 3.3B). While chromosome 7 had a single, broad peak across its center, T_{MRCA} along chromosome 4 was elevated across three large regions that aligned with regions of high F_{ST} (Figure 3.3B). The broad first region of chromosome 4 runs approximately 10 Mb and peaks locally at nearly 6 Mya in the region surrounding the gene *Eda*. The two regions distal from *Eda*, however, had peak T_{MRCA} estimates up to a million years earlier, including local maxima of 6.5 and 7.2 Mya.

Genomic Structural Variation is Associated with Deep Divergence Times

Three known inversions on chromosomes 1, 11, and 21 were all associated with local peaks in T_{MRCA} (Figure 3.2). The inversion polymorphisms on chromosomes 1 and 21 both had the highest smoothed T_{MRCA} to be found on either chromosome. This pattern also holds true for the chromosome 21 inversion, with the exception of a small region on the very end of the chromosome that was unassociated with freshwater divergence. The small sizes of the inversions on chromosomes 1 and 11 resulted in single peaks of T_{MRCA} at the window sizes we used, but the chromosome 21 inversion contained three distinct peaks. We identified a single peak at 11.4 Mb near the distal breakpoint of the inversion.

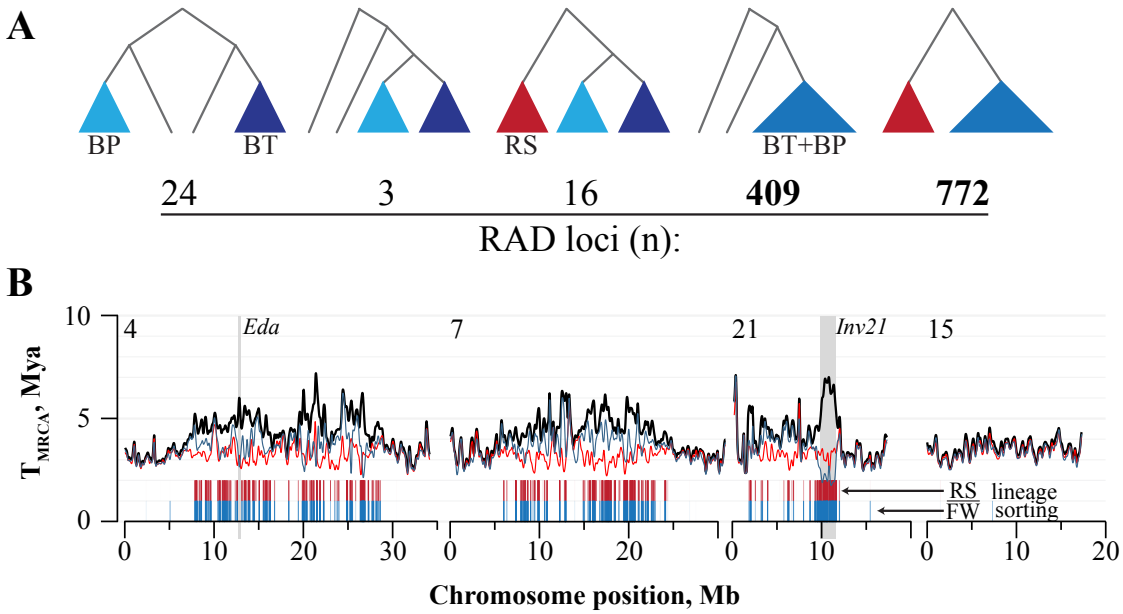

Figure 3.3. **Marine-freshwater divergence is associated with identity-by-descent among freshwater populations**. Panel A: Gene trees at RAD loci showing complete lineage sorting in both freshwater populations were enriched for trees containing a common clade containing haplotypes from both freshwater populations. Panel B: marine-freshwater lineage sorting is associated with increased coalescence times.

Two peaks closer to the center of the inversion, at 10.4 Mb and 10.8 Mb, were the tallest peaks in this region, reaching a maximum of 7 Mya.

Ocean-to-freshwater divergence in stickleback populations has been associated with sharp reductions in polymorphism due to recent ancestry surrounding adaptive alleles (Roesti, Kueng et al. 2015, Chapter II of this work). However, we did not observe a consistent, concomitant decrease in smoothed within-population T_{MRCA} in regions of high F_{ST} at a genomic scale (Figure 3.2B,C,D). Within-population T_{MRCA} was reduced in both freshwater populations. In addition, when they were combined, T_{MRCA} was also reduced within the three chromosomal inversions, but this pattern was less evident elsewhere. T_{MRCA} within each population across the large, divergent regions on chromosomes 4 and 7 was highly variable but was neither strongly nor consistently reduced. In fact, with the exception of the inversions, T_{MRCA} among the combined

freshwater populations hewed closely to the smoothed T_{MRCA} of all three populations combined (Figure 3.2E).

To further investigate the relationship between ocean-to-freshwater divergence and within- and among-population T_{MRCA}, we identified the most highly differentiated RAD loci by estimating F_{ST} between Rabbit Slough and the combined freshwater populations (RS vs [BT+BP]) and identified loci in the top 1% of this distribution (F_{ST} > ~0.767, n = 456). This set of loci averaged nearly twice as old as the remainder of the genetic variation genome-wide (5.97 Mya vs 3.75 Mya; t-test on log-transformed values: $t_{482} = 27.39$, $p \leq 10^{-10}$; Figure 3.2A). Consistent with previous results, T_{MRCA} within all three populations, and among the combined freshwater populations, was reduced at these RAD loci relative to the lower 99% (Figure 3.2).

Parallel Freshwater Divergence Involved Alleles that are Identical-by-Descent

Parallel divergence of stickleback populations to freshwater habitats commonly involves the same genomic regions, but it is mostly unknown whether it involves the same or different alleles within those regions (Hohenlohe, Bassham et al. 2010, Jones, Grabherr et al. 2012). To identify whether ocean-to-freshwater divergence involved a common set of haplotypes among the Boot Lake and Bear Paw Lake populations, we screened all RAD loci for genealogical sorting patterns that indicate the fixation of haplotypes that are either of independent origin or are IBD (Figure 3.3A). We identified a total of 1223 RAD loci with evidence of complete lineage sorting in both freshwater populations, finding examples of each pattern we screened for, including alleles with independent origins. But by far the most common pattern among these, at 1181 loci (96.6%), was the formation of a common clade of freshwater haplotypes composed of those found in both Boot Lake and Bear Paw Lake populations. This pattern could result either from soft sweeps of the same group of haplotypes in either or both populations, which would retain some polymorphism within either population, or from independent hard sweeps of the same haplotypes in both populations, resulting in the complete reduction of within- and among-population polymorphism. Of the 1181 loci with a

common clade of freshwater haplotypes, we observed polymorphism ($\pi > 0$) in both populations at 362 loci (30.7%), while we observed no polymorphism in either population at 423 loci (35.8%). A total of 396 loci (33.5%) retained polymorphism in a single population ($\pi > 0$ in Boot Lake: 270; $\pi > 0$ in Bear Paw Lake: 126). Despite residual polymorphism at many of these loci, π within freshwater populations was substantially reduced in all of the above groups of loci.

Freshwater Chromosomes Retain Anciently Derived Linked Variation, Marine Chromosomes do Not

We investigated the impact of divergent selection on linked variation within and among stickleback populations by partitioning RAD loci into those with evidence of complete oceanic-freshwater lineage sorting (hereafter 'divergent'), those on the same chromosome as a divergent locus ('linked'), and those on chromosomes without divergent loci ('unlinked'). T_{MRCA} of alleles at linked loci was highly correlated with proximity to a divergent locus across all population partitions (Spearman's rho: all *p*-values $\leq 10^{-10}$) but these relationships varied qualitatively between the marine and freshwater populations. When all populations were combined, the average T_{MRCA} at linked loci increased sharply when loci were within approximately 250 kb from a divergent locus (Figure 3.4A). This proximity effect, however, extended out to around a megabase. A much weaker, albeit still significant, relationship between proximity to a divergent locus and T_{MRCA} at linked loci existed in the oceanic Rabbit Slough population. Freshwater populations, however, showed a more complex proximity effect distinct from either Rabbit Slough or the combined populations (Figure 3.4C,D). In both Boot Lake and Bear Paw Lake populations, as well as when these populations were combined, we observed an increase in T_{MRCA} in proximity to a divergent locus which peaked at approximately 200 kb before reversing direction. In freshwater populations, therefore, the loci with the highest T_{MRCA} genome-wide were those that were approximately 200 kb from a divergent locus.

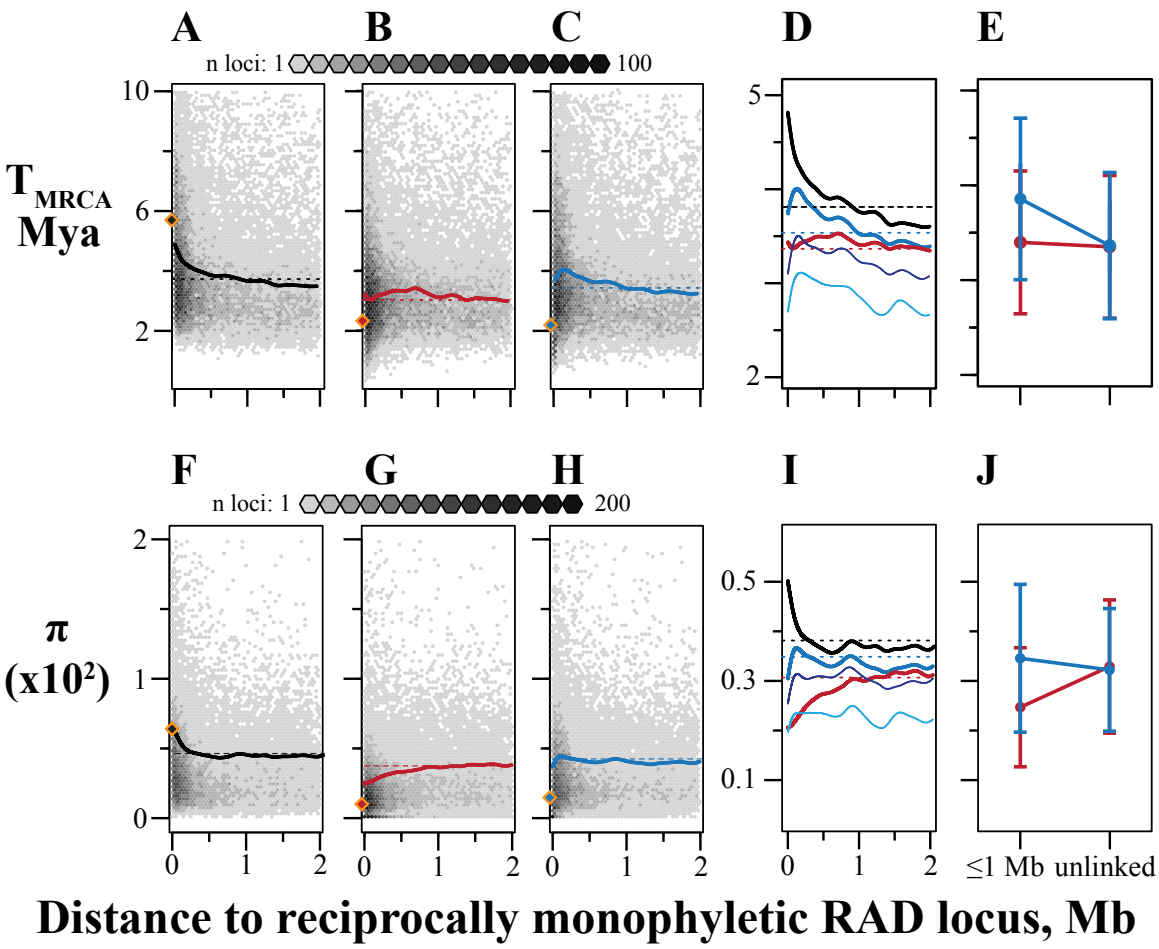

Figure 3.4. **The effects of divergent selection on linked genomic variation**. Panels A-C: The x-axis is distance from each RAD locus to the nearest reciprocally monophyletic RAD locus (see figure 3.3A, middle and far-right trees). In gray is a 2D histogram showing the density of RAD loci in a given hexagonal bin. Lines are smoothed splines, excluding reciprocally monophyletic loci. Diamonds at x=0 are average T_{MRCA} at reciprocally monophyletic RAD loci. A: All populations; B: Rabbit Slough; C: Combined freshwater populations. Panel D: Expanded view of splines for each population and grouping. Boot Lake and Bear Paw Lake populations are shown with thin dark and light blue lines, respectively. Panel E: Interaction plot of T_{MRCA}-by-distance for Rabbit Slough (red) and the combined freshwater populations (blue). "Unlinked" includes only loci on chromosomes carrying no reciprocally monophyletic RAD loci. The y-axis is the same as (D). F-J: Same as A-E, but with π.

Sequence diversity, π, showed qualitatively similar patterns to T_{MRCA} among populations, but contrasted notably in the oceanic population (Figure 3.4F-J). Among all three populations, we observed a similar increase in nucleotide diversity at linked loci in proximity to divergent loci as we saw in our estimates of T_{MRCA}. Among the two

freshwater populations, nucleotide diversity, like T_{MRCA}, peaked among linked loci around 200 kb from a divergent locus, although the correlation was nonsignificant among all linked loci (Spearman's rank order correlation: $\rho = 0.004996$, $p > 0.25$). Within-population nucleotide diversity decreased within ~250 kb of divergent loci, as observed with T_{MRCA}. This trend was evident in both freshwater populations despite the differences in overall nucleotide diversities between the two, but it was in the oceanic population that we observed the most drastic reduction in π in proximity to divergent loci relative to background levels of genetic variation. Moreover, this effect extended out beyond a megabase, on average. This had the effect of reducing overall nucleotide diversity in the oceanic population to levels at or below those seen in the smaller, recently derived freshwater populations even though background levels of π were consistently higher in the oceanic population away from divergent loci.

We thus observed a proximity-by-population interaction in levels of sequence variation at linked versus unlinked loci in oceanic or freshwater populations (Figure 3.4). While estimates of genetic diversity were similar between the Rabbit Slough population and the combined freshwater populations on chromosomes without evidence of divergent natural selection, both T_{MRCA} and π were higher among freshwater populations than in Rabbit Slough within 1 Mb of a divergent locus (Figure 3.4E; ANOVA: population*proximity interaction $p \leq 10^{-10}$). On chromosomes with no evidence of divergent natural selection, we found no evidence of differences in T_{MRCA} between Rabbit Slough and the combined freshwater populations (RS: mean = 3.35 Mya, BT+BP: mean = 3.36 Mya; Tukey's HSD, $p > 0.99$). However, among linked loci within 1 Mb of a divergent locus, T_{MRCA} in the combined freshwater populations was greater than in Rabbit Slough (Tukey's HSD, $p < 10^{-5}$), and this difference was due to an increase in T_{MRCA} among the freshwater populations (Tukey's HSD, $p < 10^{-5}$). Patterns of nucleotide diversity showed a similar population-by-proximity interaction (Figure 3.4J; ANOVA: population*proximity interaction $p \leq 10^{-10}$). But, unlike our estimates of T_{MRCA}, we observed the highest levels of π among linked loci in freshwater populations due to both

an increase in π in freshwater populations in proximity to a divergent locus and a decrease in π among these same loci in Rabbit Slough relative to unlinked loci.

Haplotype Sharing is Common Between Threespine and Ninespine Sticklebacks

Despite ninespine and threespine stickleback having diverged approximately 15 million years ago (Aldenhoven, Miller et al. 2010), threespine haplotypes were not universally monophyletic to the exclusion of the single ninespine haplotype among RAD loci in our dataset. Out of a total of 55425 RAD loci included in our dataset, we identified 9620 loci (17.36%) in which threespine monophyly was not supported by at least 50% of trees sampled from the posterior distribution (median across the entire dataset: 87.2% of sampled trees supported monophyly). Because we only sampled a single ninespine haplotype, this was likely due to insufficient sequence divergence between sampled threespine haplotypes and the ninespine haplotype, resulting either from (1) lack of polymorphism across all haplotypes or (2) polymorphism among threespine haplotypes comparable to the divergence to the ninespine haplotype. To differentiate between these two possibilities, we compared polymorphism (π) within threespine stickleback to the minimum sequence distance between a threespine haplotype and the ninespine haplotype (Figure 3.5). Across all loci, the minimum distance to the ninespine haplotype averaged 0.059 substitutions/site, although this distribution was distinctly trimodal, with modes at approximately 0, 0.005, and 0.075 substitutions/site. Sequence divergence to the ninespine haplotype was therefore, either nonexistent, comparable to what was common among threespine haplotypes, or approximately an order of magnitude greater than the average pairwise distance between threespine haplotypes. Many RAD loci at which we observed little or no divergence between threespine and ninespine haplotypes had average ($\pi \sim 0.005$) or above-average levels of polymorphism within threespine stickleback. Notably, loci in which we detected no threespine-ninespine sequence divergence (n = 1389) did not, on average, have drastically reduced levels of within-threespine polymorphism (mean $\pi = 0.00489$) and we observed at least some polymorphism within threespine haplotypes in all cases (min $\pi = 0.00015$).

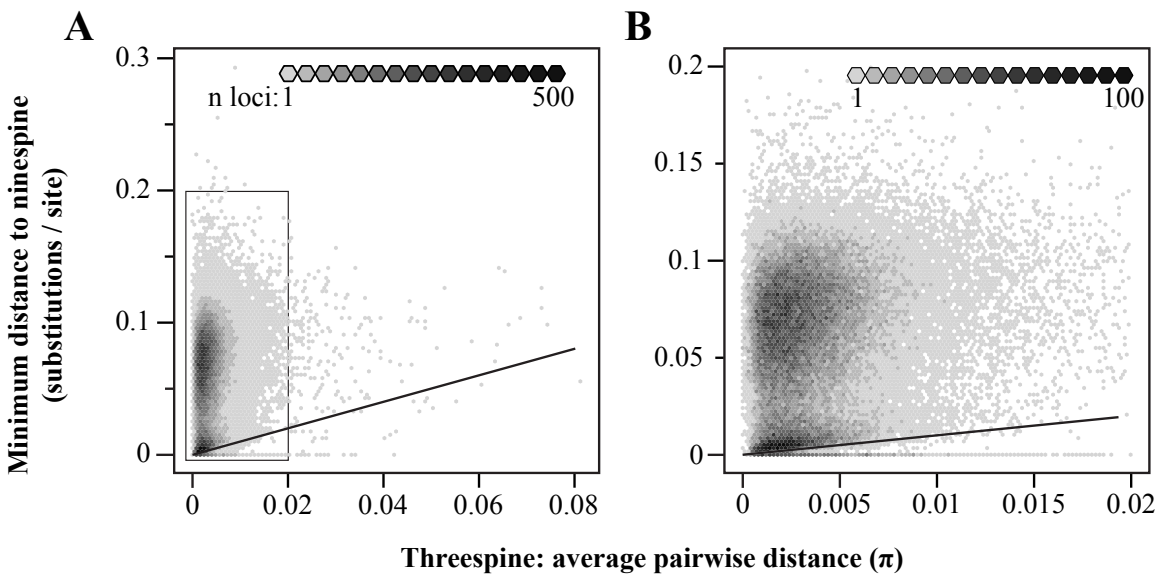

Figure 3.5. **Sequence divergence between threespine and ninespine sticklebacks compared to sequence diversity within threespine stickleback**. Diagonal lines have a slope of 1. The area below the 1:1 line includes RAD loci where average sequence distance between threespine haplotypes is less than the minimum sequence distance to the ninespine haplotype.

DISCUSSION

A primary goal of evolutionary biology is to understand how evolutionary forces act and interact to maintain genetic variation (Wright 1932, Lewontin 1974, Kimura 1983, Bernatchez 2016). Using the tools of population genomics, we are beginning to understand how direct and indirect (linked) natural selection, through the genomic architecture, affect the distribution of genetic variation across the genome (Begun, Holloway et al. 2007, Hahn 2008, Burri, Nater et al. 2015, Martin, Most et al. 2016). In this study, we have shown not only that divergent natural selection is essential to the long-term maintenance of genetic variation in the threespine stickleback, but also that the impact of selection on the maintenance of linked variation is asymmetric between chromosomes carrying marine- and freshwater-adaptive alleles.

Across the stickleback genome, divergent selection has maintained alternative allelic forms in freshwater and oceanic habitats for millions of years longer than in the majority of other genomic regions. The maintenance of one or a few freshwater-adaptive

haplotypes at many genomic regions was evident in the distribution of variation among two recently derived freshwater populations, where we observed a higher degree of identity-by-descent at divergently selected loci than elsewhere in the genome. The maintenance of variation at linked loci, however, differed greatly among habitat types. In the marine Rabbit Slough population, we observed strong reductions in genetic diversity in the broad vicinity of loci under divergent selection, suggesting persistent purifying selection. In contrast, and perhaps counterintuitively, the small, recently founded Boot Lake and Bear Paw Lake populations retained anciently derived variation in these same genomic regions. Below, we discuss the implications of these findings on our understanding of adaptation in nature and suggest a model of demographic and adaptive history to explain the contrasting patterns we observed.

Divergent Selection Maintains Adaptive Variation for Millions of Years

Spatially varying selection pressures create opportunities for the maintenance of adaptive variation (Charlesworth, Nordborg et al. 1997, Lenormand 2002). In the threespine stickleback, adaptive divergence is pervasive and occurs at local (Hohenlohe, Bassham et al. 2010) and global scales (McKinnon and Rundle 2002, Jones, Grabherr et al. 2012). Studies of freshwater divergence at the local scale have found remarkable levels of genomic parallelism even among very recently derived freshwater populations, suggesting a common pool of adaptive variation already present in the founding populations (Hohenlohe, Bassham et al. 2012, Lescak, Bassham et al. 2015, Roesti, Kueng et al. 2015). The presence of SNP variation that differentiates freshwater and oceanic stickleback across the Northern Hemisphere further indicates that adaptive variation could be transported globally. The large spatial scale and immense census size of the species as a whole does, however, leave open the possibility of parallel mutational input in different parts of the species range rather than shared variation with a single evolutionary origin (Shapiro, Marks et al. 2004, Chan, Marks et al. 2010, Ralph and Coop 2015).

In this study, adaptive divergence in the Boot Lake and Bear Paw Lake populations proceeded primarily through a common pool of standing genetic variation that had been maintained for millions of years and was identical-by-descent among the two populations. Allelic divergence at those loci most associated with recent ocean-to-freshwater divergence dated back ~6 Mya on average, nearly twice as old as the average T_{MRCA} across the rest of the genome (Figure 3.2A). This suggests that divergent natural selection is the primary force maintaining genomic variation in the threespine stickleback.

Despite the ancient divergence of oceanic and freshwater alleles, we observed the maintenance of one or a few core haplotypes at divergently selected loci, which manifested in recent identity-by-descent among freshwater populations (Figure 3.3). Of the RAD loci that were most divergent between the oceanic and freshwater populations, the majority contained a common clade of freshwater alleles in which haplotypes isolated from Boot Lake and Bear Paw Lake were indistinguishable. At nearly two-thirds of these loci, we observed residual polymorphism either within or among freshwater populations, indicating soft sweeps on standing genetic variation. The remainder contained no polymorphism either within or among the freshwater populations. While it is possible that variation at these loci represents the stochastic loss of haplotypes after initially soft sweeps ("hardened" soft sweeps: Wilson, Petrov et al. 2014), both populations often fixed identical haplotypes. A more plausible explanation may therefore be that the founding of these populations involved a single adaptive haplotype in both cases, resulting in a true hard sweep. Moreover, given that adaptive genetic variation among these two populations was often identical-by-descent despite being millions of years diverged from marine alleles, we hypothesize that the shared allelic states among freshwater stickleback populations across the Northern Hemisphere are due to shared haplotypes derived from the same original mutations. This mode of parallel divergence has been suggested previously (Schluter and Conte 2009), notably with regard to the *Eda* locus (Colosimo, Hosemann et al. 2005). Our data support these suspicions, demonstrating that parallel

evolution leading to IBD among independently derived freshwater populations is a genomic phenomenon.

Asymmetries in Selection and Population Structure Maintain Asymmetries in Linked Variation

The patterns of linked variation that we found highlight not only the importance of selection in maintaining genetic variation but also of the structure of the threespine stickleback metapopulation itself. Throughout the species range, the marine population is remarkably uniform phenotypically, a uniformity reflected in the observation of minimal isolation-by-distance over thousands of kilometers (Bell and Foster 1994, Catchen, Bassham et al. 2013). In contrast, stickleback adapted to freshwater environments can be highly variable within and among populations (McKinnon and Rundle 2002, Leaver and Reimchen 2012), which suggests either reduced selective constraint or diversifying selection on a number of traits. In addition to selective asymmetries, marine and freshwater stickleback ecotypes experience divergent demographic histories, with the large panmictic marine population abutting against and giving rise to thousands of freshwater lake and stream populations. Many freshwater populations, like those studied here, do not directly connect with their neighboring oceanic populations, but many do, and gene flow among divergently adapted populations occurs across much of the species range. And while the influence of freshwater populations on the evolution of the species as a whole has been questioned (Bell and Foster 1994), gene flow between freshwater and oceanic stickleback populations is now known to be common and may facilitate adaptation through the indirect sharing of alleles among freshwater populations (Schluter and Conte 2009).

Contrary to previous models of stickleback evolution, which emphasized the immense marine population the stable store of variation and envisioned freshwater populations as small, ephemeral, and often dead-end (Bell and Foster 1994), our findings suggest that the combined actions of asymmetric selection pressures in marine and freshwater habitats and (indirect) gene flow among freshwater populations has led to

freshwater populations being the primary reservoirs of genetic variation in the species. Loci under divergent selection were associated with some of the most anciently derived alleles that we observed genome-wide, stretching back perhaps to the original split with the ninespine stickleback lineage 15 million years ago. Within-population and within-ecotype diversity, however, was lowest at these loci in all three populations we studied (Figure 3.2, Figure 3.4). In genomic regions far from those under divergent selection, chromosomes isolated from marine and freshwater habitats were essentially identical (Figure 3.3, Figure 3.4). We observed no difference in coalescence times among ecotypes and only marginal differences in allele frequencies, indicative of mostly neutral sorting of standing genetic variation since the postglacial isolation of these populations.

At intermediate distances from loci under divergent selection, extending out approximately one megabase on average, chromosomes carrying marine and freshwater alleles had strikingly different patterns of ancestry — the genomic consequences of asymmetric selection pressures and asymmetric demography among stickleback ecotypes (Figure 3.6). Diversity on marine chromosomes appeared to be the result not just of selection at divergently adaptive loci but also of selection acting on additional loci contributing to the marine phenotype. Average sequence diversity on marine chromosomes increased steadily away from loci under divergent selection (the smoothed spline in Figure 3.4), but there was also a high density of RAD loci within approximately 250 kb of a locus under divergent selection for which we found only a single haplotype ($\pi = 0$). This additional contribution of persistent selection in the marine habitat also helps to explain pervasive linkage disequilibrium in marine populations, a pattern identified earlier but with unknown causation (Hohenlohe, Bassham et al. 2012).

In stark contrast, these same regions were the oldest and most diverse genomic regions we observed on freshwater chromosomes (Figure 3.4E,J). We hypothesize this to be the result of the combined effects of extensive population structure among freshwater populations and protection against recombination with marine chromosomes. Established theory shows that population structure itself will increase inter-population coalescence times relative to panmixia (Charlesworth, Charlesworth et al. 2003), but natural selection

can easily muddy the signature of population structure in genomic data. The increased coalescence times on freshwater chromosomes adjacent to divergently adapted loci may be the result of the geographic structuring of variation among semi-isolated freshwater populations (Figure 3.6). This signature is retained only in those genomic regions under additional selection in the marine environment because gene flow among freshwater populations often takes place exclusively through the marine population. Were recombination between freshwater and marine chromosomes not selected against in these regions, the among-freshwater population ancestry would be erased similar to chromosomes not affected by divergent selection (Figure 3.3B).

Our findings demonstrate that the results of Roesti *et al*. (2014) have important implications for the long-term evolution of the threespine stickleback. Those authors found that adaptive evolution in independently derived freshwater stickleback populations occurred via sweeps of common sets of SNPs in each population, but that

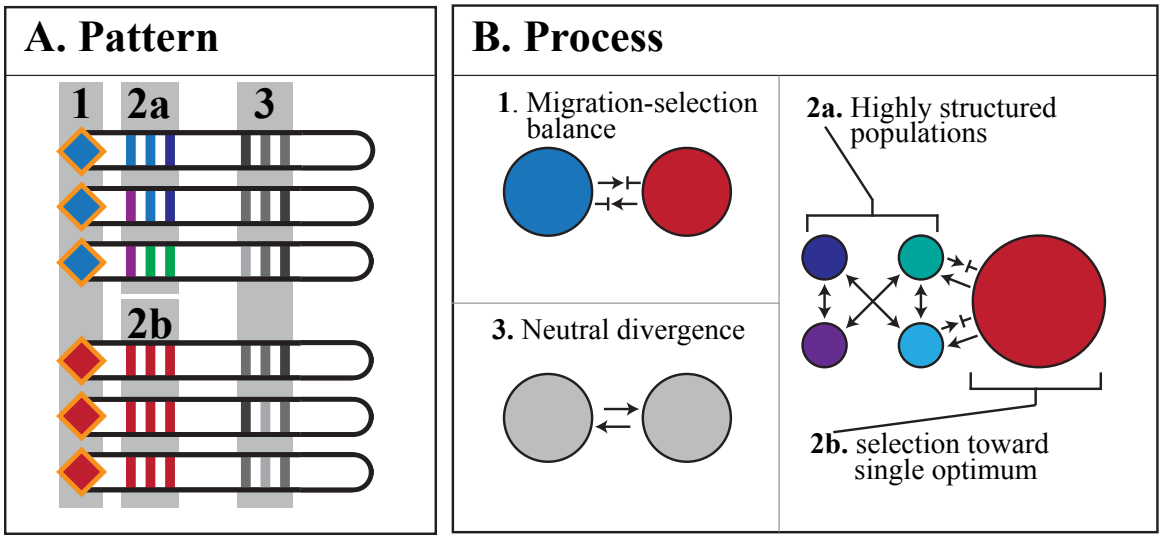

Figure 3.6. Hypothesized evolutionary processes which have led to observed patterns of genomic variation. Panel A: Chromosomes carrying freshwater- (blue diamonds) and marine-adaptive (red diamonds) alleles under divergent selection (labeled as region 1). Vertical bars are variable sites, with different colors representing different nucleotide identities. Patterns of variation on chromosomes are matched with depictions of population-level processes in panel B. Circles represent populations, with sizes representing relative census sizes. Arrows represent gene flow. "T" arrows indicate selection against migrant alleles. Colors indicate genetic differentiation of populations.

those SNPs resided on different haplotype backgrounds in each instance. This led to freshwater populations being most differentiated from one another in genomic regions nearby those loci under selection. By adding measures of sequence diversity and coalescence times, we have shown these linked genomic regions to be the most diverse in the genome. Alleles in these regions share common ancestry that is nearly a million years older, on average, than on chromosomes uninvolved in marine-freshwater divergence (Figure 3.4). We find it likely that selection and gene flow happening over shorter evolutionary timescales (tens to thousands of years), like those investigated by Roesti *et al.* (Roesti, Gavrilets et al. 2014), have occurred throughout the evolutionary history of this species and have long-lasting, collective effects on patterns of genomic variation over the course of millions of years.

Gene Flow Introduces Variation Across Species Boundaries

The field of evolutionary genomics has yielded many surprises in recent years, chief among these is the porosity of species boundaries (Kane, King et al. 2009, Fontaine, Pease et al. 2015). Shared genetic variation among distinct species was long thought to be exceptionally rare due to the sorting of residual polymorphism over the long timescales of species divergence and the prevention of gene flow by the presence of intrinsic and extrinsic reproductive isolating barriers (Coyne and Orr 2004). Population genomic analyses have, however, demonstrated not only that shared variation across species boundaries is a common phenomenon in many sexually reproducing taxa but that these patterns often result from or contribute to adaptive evolution (Stankowski and Streisfeld 2015, Pease, Haak et al. 2016, Wallbank, Baxter et al. 2016). In addition to the influences of selection and population structure among threespine stickleback populations, we found that gene flow between threespine and ninespine stickleback species also contributes to the standing pool of genetic variation (Figure 3.5). Despite an estimated 15 million-year divergence time between the two species (Aldenhoven, Miller et al. 2010), we found that over 17% of RAD loci lacked clear evidence of monophyly of threespine stickleback alleles. Most of these loci showed typical levels of polymorphism within threespine

stickleback, discounting a primary contribution from incomplete lineage and maintenance of polymorphism predating the species split. In fact, we commonly observed very low levels of sequence divergence between threespine and ninespine stickleback haplotypes, suggesting recent gene flow. In many cases, the ninespine haplotype was identical to a threespine haplotype despite typical levels of polymorphism among threespine haplotypes.

Evidence for gene flow between stickleback species was pervasive but showed no clear genomic pattern. Lack of threespine monophyly occurred genome-wide and without clear relationship with existing habitat-associated genetic or genomic variation. Nonetheless, the extent to which we identified patterns of polymorphism consistent with recent gene flow between these long-diverged species highlights the potential for gene flow to introduce variation across species boundaries and contribute to high levels of standing genetic variation within species.

Conclusions

The standing pool of genomic variation in the threespine stickleback is the combined result of an interaction of natural selection and population structure mediated by the physical structure of the genome. Natural selection has structured variation across diverse habitats and has maintained adaptive alleles that first arose millions of years ago. Highly structured freshwater populations, connected via gene flow that often occurs through the marine population, maintain additional diversity in regions of the genome where selection guards freshwater chromosomes against recombination with marine forms. These combined results emphasize the importance of linkage in structuring genomic variation in this species, and additional work focusing on the recombination landscape will be critical for our understanding of stickleback evolution. Lastly, interspecific gene flow with the ninespine stickleback has infused much of the genome with additional genetic variation. Given clear evidence for gene flow across species boundaries and the ability of natural selection to maintain genetic variation, our results highlight the diverse potential sources for phenotypically and adaptively significant

genetic variation and the need for broad taxonomic sampling across long evolutionary timescales to explain even recent events in the history of a species.

BRIDGE

Marine-freshwater divergence has been maintained throughout much of threespine stickleback evolutionary history. Despite this long history, newly established freshwater populations adapt rapidly and do so with remarkable parallelism, drawing on the same core haplotypes across much of the genome. These results, along with previous work showing high levels of linkage disequilibrium in marine populations (Hohenlohe, Bassham et al. 2012), suggest that adaptation to freshwater begins with large, pre-structured haplotypes, each containing many individual adaptive genetic variants held together in linkage disequilibrium. It is possible that strong selection for multilocus genotypic combinations is solely responsible for this pattern, but the recombination landscape is what provides the raw genomic variation available to selection. Indeed, links between the recombination rate and genetic diversity are evident in studies in diverse taxa (Lowry and Willis 2010, Andrew and Rieseberg 2013, Küpper, Stocks et al. 2016, Martin, Most et al. 2016, Ortiz-Barrientos, Engelstadter et al. 2016), including stickleback (Jones, Grabherr et al. 2012, Roesti, Moser et al. 2013, Roesti, Gavrilets et al. 2014). In chapter IV, I add genome-wide estimates of recombination rate variation to my investigation of the forces governing genomic variation. I find that genomic divergence is almost exclusive to chromosomal regions of low recombination. The recombination landscape in multiple instances turns large genomic regions into discrete Mendelian loci, contributing to the selective maintenance of long, adaptive haplotypes.

CHAPTER IV

E PLURIBUS UNUM: THE RECOMBINATION LANDSCAPE SIMPLIFIES THE GENETIC ARCHITECTURE OF ADAPTIVE DIVERGENCE

INTRODUCTION

Divergent selection is a major driver of biological diversity at all levels, from local adaptation at micro-geographic scales to the formation of new species (Darwin 1859, Hendry, Taylor et al. 2002, Lekberg, Roskilly et al. 2012, Nosil 2012). Especially among closely related populations, adaptive divergence often takes place in the context of gene flow, when diverging populations exist in close geographical proximity and reproductive isolating barriers are incomplete (Lenormand 2002, Andrew and Rieseberg 2013). The combined action of selection and gene flow shape patterns of genomic variation and the genetic architecture of divergence (Turner, Hahn et al. 2005, Yeaman and Whitlock 2011, Via 2012, Yeaman 2015). Understanding the mechanisms that permit and maintain divergence in the face of gene flow is necessary for a complete understanding of the evolution of biodiversity.

The combined action of selection and gene flow leads to heterogenous genomic divergence among diverging populations (Harr 2006, Nadeau, Whibley et al. 2012). In many taxa, genomic divergence is clustered within the genome into relatively few regions, which stand out as "islands" against a sea of genomic homogeneity (Turner, Hahn et al. 2005, Nosil, Funk et al. 2009). Among taxa, the size and number of genomic islands range widely, demonstrating that the evolution of heterogenous genomic divergence is itself heterogenous. In some studies of closely related populations or species, including crows (Poelstra, Vijay et al. 2014), warblers (Toews, Taylor et al. 2016), and butterflies (Van Belleghem, Rastas et al. 2017), distinct phenotypic morphs are distinguished by few, narrow islands, often containing one or a few genes each. In other systems, large genomic regions differentiate closely related taxa, as is the case in *Mimulus* monkeyflowers (Lowry and Willis 2010, Twyford and Friedman 2015), *Helianthus* sunflowers (Andrew and Rieseberg 2013), and threespine stickleback fish

(Hohenlohe, Bassham et al. 2010, Jones, Grabherr et al. 2012, Chapter III of this work). While evidence from these and other systems demonstrate that genomic divergence can take many forms, we still understand little about how multiple evolutionary forces give rise to specific patterns of heterogenous genomic divergence.

The recombinational landscape — which we define as the distribution of observed recombination events across the genome and among individuals — is a crucial factor in the evolution of genomic heterogeneity, but the precise role it plays is often unclear and likely context-dependent. Recombination breaks physical linkage between alleles under selection and selectively neutral variation in adjacent genomic regions (Smith and Haigh 1974, Charlesworth, Nordborg et al. 1997). Given sufficient time, recombination will erode genetic divergence between hybridizing populations except in those genomic regions that are under divergent selection and thereby resistant to gene flow. Under this model, gene flow will homogenize genomic variation except at loci under sufficiently strong selection to resist gene flow (Yeaman and Whitlock 2011), which may be the case in closely related taxa distinguished by narrow divergence peaks (e.g. butterfly wing pattern morphs: Van Belleghem, Rastas et al. 2017).

The existence in some taxa of broad genomic regions of divergence, sometimes encompassing millions of base pairs and hundreds of coding genes, necessitates more complex models of evolution which are difficult to distinguish in part because of uncertainty in the evolutionary histories of the studied populations. Adaptive introgression from an unobserved population, for instance, will initially involve a single, large haplotype tract, even if only a single variant within that haplotype is selected for. Large islands of divergence have also been argued to be the result of secondary contact between formerly allopatric populations (Durrett, Buttel et al. 2000). In this case, recombination erodes neutral divergence built up in allopatry. Genomic regions surrounding those under divergent selection erode more slowly, leading to larger islands than would be expected at equilibrium (Charlesworth, Nordborg et al. 1997). A common theme among these models is that large islands are likely transient and recombination will narrow them into peaks with sufficient time and continued gene flow.

The recombinational landscape is also hypothesized to play a role in the accumulation of divergence within existing islands and in the expansion of islands into neighboring genomic regions (Maynard Smith and Haigh 1974, Charlesworth, Nordborg et al. 1997, Via 2012). This is because the establishment and persistence of a new allele depends both on its effect on fitness and the combined fitness effects of linked variation. Alleles with weak effects on fitness, that may by themselves be unable to resist the homogenizing effects of gene flow, may still establish if they are in linkage disequilibrium with larger-effect alleles already resistant to gene flow. Supporting the importance of maintaining linkage disequilibrium, islands of divergence often coincide with chromosomal inversions. Inversions almost completely suppress recombination in heterozygotes while maintaining free recombination within homozygotes. Inversions are common in nature, especially among hybridizing taxa where the potential is high for recombination among divergently adapted haplotypes (Kirkpatrick and Barton 2006, Lowry and Willis 2010, Jones, Grabherr et al. 2012, Fontaine, Pease et al. 2015, Lee, Fishman et al. 2016, Lee, Wang et al. 2017).

A full understanding of adaptive divergence requires descriptions of both the genomic and recombinational landscapes of divergence. The genomic extent of divergence provides clues to the number of genes involved and the extent of molecular divergence between populations (see Chapter III). The recombinational landscape provides complementary information, describing how molecular variation is inherited during adaptive divergence. Moreover, because the recombinational landscape itself can vary across individuals and populations, as is the case with chromosomal inversions, measurements of recombination rate should include individual-level variation from the populations being studied.

In this study, we investigate how variation in the recombination rate across the genome and in different genomic backgrounds influences the inheritance and maintenance of heterogenous genomic divergence in the threespine stickleback (*Gasterosteus aculeatus*). The stickleback is an ecologically diverse species, with locally adapted populations existing throughout the Northern Hemisphere (Bell and Foster 1994,

McKinnon and Rundle 2002). Ancestrally marine populations have repeatedly colonized and adapted to freshwater habitats. These multiple instances of adaptive divergence involve parallel phenotypic evolution including changes to body shape, defensive and trophic morphologies, and the immune system. Parallel phenotypic evolution reflects parallel genomic evolution, with stereotypical patterns of genomic divergence across multiple chromosomes (Hohenlohe, Bassham et al. 2010, Jones, Grabherr et al. 2012). Adaptive genomic variation is recruited from standing genetic variation in the ancestral population (Colosimo, Hosemann et al. 2005, Schluter and Conte 2009, Roesti, Gavrilets et al. 2014, Chapter II of this work). Importantly, although many freshwater populations are young — some less than a century old (Lescak, Bassham et al. 2015) — the regions of genomic divergence between marine and freshwater stickleback are large and old, having been maintained not only in the context of gene flow among divergently adapted populations but also as segregating variation in the marine population.

We combine the results of genome scans from a marine-freshwater population pair with genome-wide genetic maps from stickleback lines derived from both populations, including and an F_1 hybrid. These data allow us to directly compare the genomic extent of divergence between these populations with patterns of recombination within and among individual fish. We hypothesize that large islands of divergence will be associated with low recombination rates. We further predict that recombination rates will be further suppressed when divergent genetic backgrounds are brought together in F_1 and advanced hybrids. We find that broad genomic islands of divergence often, but not always, coincide with regions of low recombination and that these were conserved across all three genetic maps. Where recombination rates did vary among individuals, they were often lowest in the hybrid. In effect, large, highly divergent genomic regions segregated as Mendelian loci. Our results add to a growing body of evidence that recombination rate is a key factor influencing genomic divergence and suggest that variation in recombination rate *among* individuals and populations impacts the maintenance of genomic divergence.

METHODS

Study Populations

We used a well-known instance of adaptive marine-to-freshwater divergence in a population pair in Cook Inlet, Alaska. The Boot Lake population is a stereotypically freshwater stickleback population in phenotype and genotype (Cresko, Amores et al. 2004, Hohenlohe, Bassham et al. 2010), which was founded when Boot Lake formed and was colonized by an ancestrally marine population at the end of the last glacial maximum approximately 12,000 years ago (Francis, Baumgartner et al. 1986, Reger and Pinney 1996). The Rabbit Slough population is phenotypically and genotypically representative of the globally-distributed marine stickleback population (Cresko, Amores et al. 2004, Hohenlohe, Bassham et al. 2010). In this study, we use wild-caught individuals to estimate genome-wide polymorphism and divergence as well as separate laboratory lines derived from both sampling locations to estimate recombination rates across the genome.

Polymorphism in Natural Populations

To describe natural genomic variation, we used the hRAD sequencing dataset described in chapter II. Briefly, we sampled five individuals each from Boot Lake and Rabbit Slough. These fish were genotyped using hRAD-seq to estimate sequence diversity and haplotype structure using 692 bp haplotypes at 102,823 *PstI* restriction sites across the genome. Paired-end sequences were quality filtered using **`process_radtags`** (Catchen, Hohenlohe et al. 2013). Paired-end reads were then merged using **`fastq-join`** (Aronesty 2011) and aligned to the stickleback reference genome (Jones, Grabherr et al. 2012, Glazer, Killingbeck et al. 2015) using **`bbmap`** (http://jgi.doe.gov/data-and-tools/bbtools) set to '**`vslow=t`**', the parameter set most sensitive to complex alignments. Aligned sequences were processed and genotypes called using the Stacks pipeline (Catchen, Hohenlohe et al. 2013).

To estimate sequence diversity and levels of relative and absolute sequence divergence between Boot Lake and Rabbit Slough populations, we analyzed all RAD loci

using packages and custom scripts written the R language (R Core Team 2016). We estimated F_{ST} using equation (3) in Hudson, Slatkin, and Maddison (1992) and implemented in the package 'PopGenome' (Pfeifer, Wittelsburger et al. 2014). Sequence diversity within and among populations (π_{BT}, π_{RS}, and π) and among-population sequence divergence (d_{xy}) were calculated as described in Nei (1987) using matrices of raw pairwise sequence distances calculated with the R package 'ape' (Paradis, Claude et al. 2004).

Laboratory Crosses and Genetic Mapping

To compare how heterogenous genomic divergence on the physical map is distributed across the genetic map, we generated mapping families from laboratory lines of fish derived from the Boot Lake and Rabbit Slough populations, as well as an F_1 hybrid female produced from a cross between a Boot Lake female and a Rabbit Slough male. These crosses allowed us to examine variation in the recombinational landscape within and among chromosomes and also across multiple genetic backgrounds. All maps were constructed using a pseudo-testcross design. The pseudo-testcross takes advantage of existing heterozygosity in wild-caught or outbred individuals without the need to generate inbred lines or F_1 mapping parents. To generate mapping families from Boot Lake and Rabbit Slough lines, we manually crossed unrelated individuals from each line. We also mapped the F_1 hybrid female by backcrossing it to a Boot Lake male. All progeny were raised to 14 days post-fertilization, euthanized with MS-222 (Sigma Aldrich), and fixed in 95% ethanol. We extracted genomic DNA from whole progeny and from pectoral and caudal fins from all parents using proteinase K digestion (Qiagen) followed by DNA purification with Ampure magnetic beads.

RAD genotyping of progeny and parents was used to identify segregating haplotypes for use as genetic markers. We used the original single-digest RAD-seq protocol (Baird, Etter et al. 2008) with the restriction enzyme *SbfI*, which cuts genomic DNA at approximately 22,000 sites in the stickleback genome. All RAD libraries were multiplexed by family at equimolar ratios except for mapping parents, which were

multiplexed at twice the concentration of the progeny to increase sequencing depth. The Boot Lake mapping family was sequenced in two lanes on an Illumina HiSeq2000. The Rabbit Slough family was sequenced on a single HiSeq2000 lane. The F_1 family was sequenced in one lane on an Illumina HiSeq2500.

RAD-seq data from all mapping crosses were processed with the Stacks analysis pipeline (Catchen, Hohenlohe et al. 2013). We demultiplexed and quality filtered sequences with process_shortreads followed by alignment to the stickleback reference genome with GSNAP (Wu and Nacu 2010). We used **`ref_map.pl`** to identify RAD tags and call genotypes. The Stacks component program **`genotypes`** was used identify segregating markers for export to genetic mapping software. We specified a minimum coverage of 3x to call individual genotypes and required that a marker be genotyped in at least 50% of progeny.

Below, we present data for a single parent from each mapping cross (Table 4.1). By conducting pseudo-testcrosses, we identified polymorphic RAD markers segregating in all six mapping parents. However, to investigate the genome-wide recombinational landscape, as well as relationships between recombination rate and natural levels of polymorphism and divergence, we restricted our analysis to mapping parents for which we observed segregating markers on all 21 chromosomes with no gaps of more than 1 megabase between adjacent markers. One parent from each cross fit these criteria. Thus, we present data for female parents from the Boot Lake and F_1 crosses and the male parent

Table 4.1. Genetic map statistics for the three genetic maps used in this study.
*map length given in centiMorgans (cM). †Genomic coverage as a percentage of the reference genome assembled onto chromosomes (436.6 Mb total).

Family	progeny	markers	map length*	genomic coverage†	cM/Mb
Boot Lake	189	6041	1694	0.986	3.93
Rabbit Slough	94	7049	1047	0.987	2.43
F_1 hybrid	102	6669	1406	0.988	3.26

from the Rabbit Slough cross. In our results, we focus on the effects of genotype on relative rates of recombination within and among mapped individuals, rather than absolute rates, because population-of-origin was confounded with sex.

We estimated genome-wide recombination rates between RAD markers under the assumption of collinearity between all genetic maps and the stickleback reference genome (Glazer, Killingbeck et al. 2015) (with some exceptions, *see below*). We used the mapping software Lep-MAP2 (Rastas, Calboli et al. 2016) to estimate map positions of RAD markers with the marker order fixed to the aligned positions on the reference genome. Known marker orders increased throughput of mapping iterations and reduced the impact of genotyping errors on recombination rate estimation. While fixed marker orders do not explicitly allow the detection of structural variants in the genome — such as chromosomal inversions that are known to exist among stickleback populations — discrepancies in the estimated map do provide indirect, and correctable, evidence of changes to map order.

Chromosomal Inversion Detection and Correction

We identified genomic structural variation among genetic maps relative to the reference genome by manually inspecting each genetic map for evidence of recombination rate distortion. Specifically, chromosomal inversions are known to segregate in stickleback populations and show large frequency differences between marine and freshwater populations (Jones, Grabherr et al. 2012, Roesti, Kueng et al. 2015). Because we used reference-ordered marker positions to estimate recombination rates, any markers present within inversions relative to the reference genome were at first forced into an incorrect order (Figure 4.1). However, because this method placed genomically distant markers in close proximity, estimated map distances between adjacent markers were artificially and detectably inflated (Figure 4.1C). Reversal of the marker order, in the case of true inversions, should reduce overall map length and eliminate large map distances between adjacent markers (Figure 4.1D).

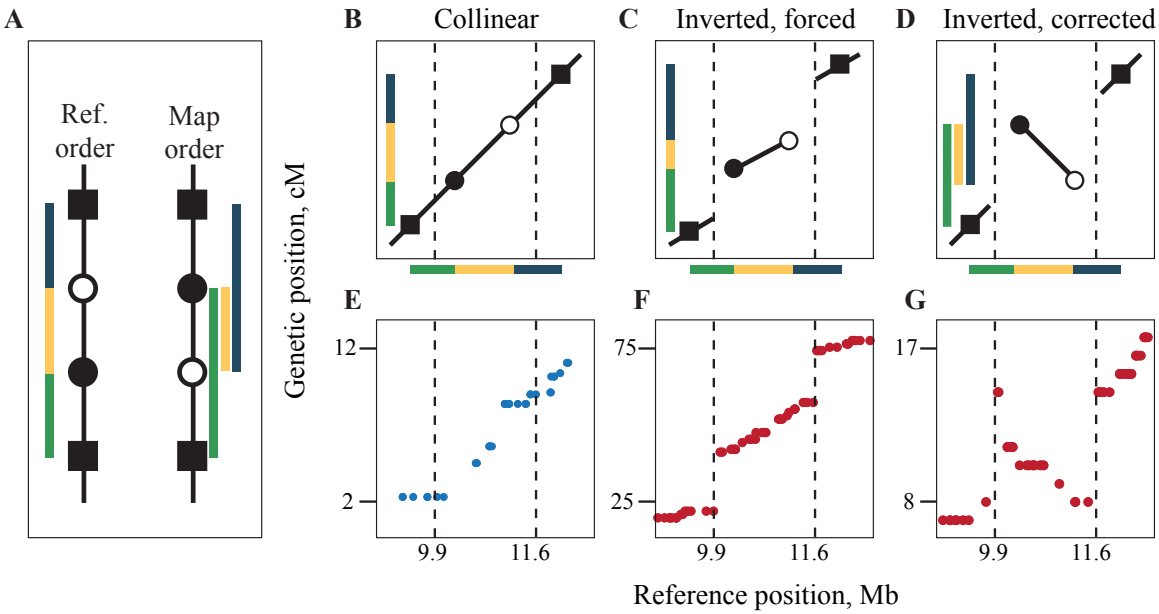

Figure 4.1. Detection of chromosomal inversions in fixed-order genetic maps. (A) Chromosomal inversions change the physical order of homologous genetic markers. Distances between markers inside the inversion remain the same (yellow bars) while their distances with markers outside the inverted region change (blue and green bars). (B) Inferred map distance increases monotonically when the genetic order agrees with the reference genome. (C) Forcing an inverted genetic map into the reference genome order exaggerates inferred distances among markers flanking the inversion breakpoint. (D) Correcting marker orders within the inversion restores the correct map order and eliminates exaggerated map distances flanking the breakpoints. (E-G) Fixed-order genetic maps from this work, including Boot Lake (E: collinear with the reference genome) and the Rabbit Slough map forced into the reference order (F) and corrected (G).

We tested this manual correction method using a known inversion on stickleback chromosome 21 that is highly differentiated among natural freshwater and marine populations (Jones, Grabherr et al. 2012). This inversion is fixed between our laboratory lines derived from Boot Lake and Rabbit Slough. In our fixed-order genetic maps we saw the expected pattern of recombination rate inflation at the breakpoints of the inversion in the Rabbit Slough map (Figure 4.1F), which carries the inversion orientation opposite the reference genome order. When we corrected the map order, the total map distance in that region of chromosome 21 shrank from over 50 cM to approximately 10 cM and

recombination rates between markers adjacent to inversion breakpoints were comparable to those elsewhere on the map. In curating genetic maps from the three crosses we performed, we found no further evidence of structural variation based on recombination distortion aside from the known chromosomal inversions on chromosomes 1, 11, and 21.

Recombination-Polymorphism Correlations

We employed two methods to investigate the relationships between the recombinational landscape and patterns of polymorphism within and among natural populations. To quantify genome-wide correlations among recombination rate and population genetic statistics, we broke the stickleback genome into 250-kb non-overlapping windows and calculated average recombination rates (in centiMorgans per megabase, cM/Mb), sequence polymorphism (π_{BT}, π_{RS}, and π), and sequence divergence (F_{ST}, d_{xy}) in each window. We used nonparametric correlations to test for correlations between variables because the distributions of most variables lacked normality even using standard data transformations. Below we present Spearman's rank order correlations. Kendall's tau and parametric linear models gave qualitatively similar results.

Genomic heterogeneity exists not only in the proportion of the genome affected by divergent selection but also in how genetic variation and divergence are clustered within the genome. Marine-freshwater genomic divergence in stickleback is clustered into few, large regions that can encompass most of the length of a chromosome. We sought to directly compare the genomic distributions of population genetic statistics along the physical genome and on the three genetic maps we constructed. We used a windowing approach that allowed direct comparisons across maps despite differing numbers and distributions of markers among genetic maps. Using the R package 'ksmooth', we binned each chromosome into equally sized intervals, the number of which we set equal to the number of segregating RAD markers on the genetic map with fewest markers. For each interval, we calculated genetic position from each laboratory cross and F_{ST} using high-density hRAD sampling of wild individuals within a 250-kb normally distributed kernel. Using this windowing approach, we were able to make direct

comparisons between recombination rates within and among genetic maps from different genetic backgrounds and patterns of polymorphism in the populations from which the laboratory lines were derived.

RESULTS

Mapping Crosses

RAD sequencing of mapping progeny from the three maps generated 740,331,821 raw single-end sequence reads distributed among 385 individuals. Of these, 520,990,827 (70.4%) passed quality filtering and 370,490,044 (71.1%) of those were unambiguously aligned to the reference genome. Sequencing results for each map are summarized in Table 4.2.

All mapping crosses resulted in over 6000 segregating RAD markers distributed among all 21 stickleback chromosomes (Table 4.1, Figure 4.2). The Boot Lake female, mapped using the largest family of 189 progeny, had the longest overall map length of 1694 cM among 6041 segregating markers and a genome-wide average recombination rate of 3.93 cM/Mb. The Rabbit Slough male, mapped using the fewest number of offspring (n = 94) but the largest number of segregating markers (7049), had the shortest estimated map length of 1047, resulting in an average recombination rate of 2.43 cM/Mb. The F_1 hybrid genetic map was intermediate in length, and also in the number of progeny and markers used in mapping. The total length of the hybrid map was 1406 cM (mean

Table 4.2. Sequencing statistics of libraries for each mapping family. Filtered and aligned reads are shown as percentages of the total number of raw reads for each library. Coverage is the mean coverage per RAD tag across all progeny.

Family	raw reads	filtered reads	aligned reads (unique)	mean coverage
Boot Lake	322871980	229263868	196451426	27.77
Rabbit Slough	188240315	110264451	95689374	25.21
F_1 hybrid	229219526	181462508	78349244	18.75
Total	*740331821*	*520990827*	*370490044*	*24.72*

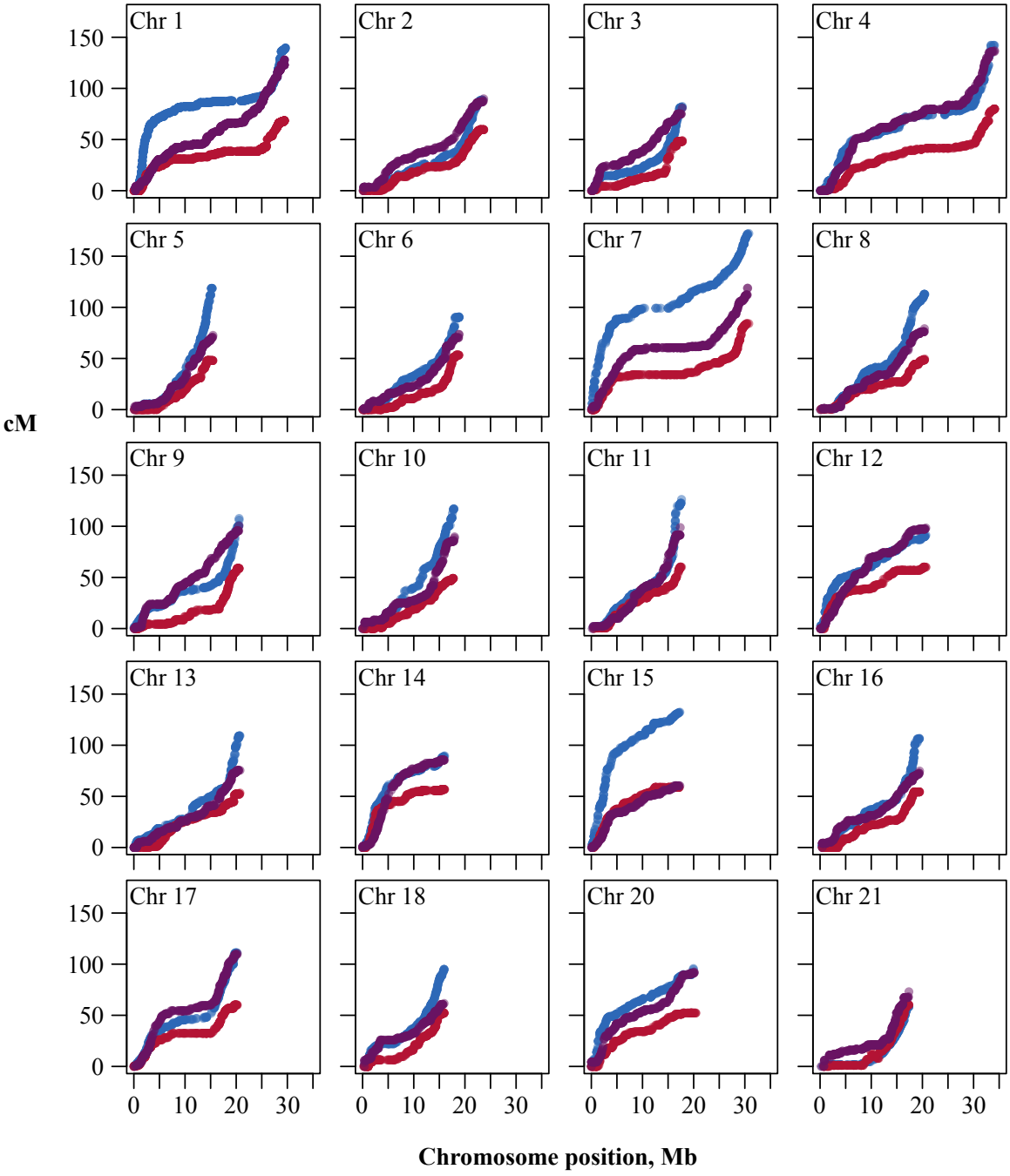

Figure 4.2. Genetic maps from Boot Lake, Rabbit Slough, and an F_1 hybrid have broadly similar recombinational landscapes. Points represent segregating RAD markers plotted by genetic position in centiMorgans (y-axis) against the reference position, in megabases, on each chromosome (x-axis). Rabbit Slough genetic map orders in known chromosomal inversions are shown corrected as described in Figure 1 and *Methods*. Blue: Boot Lake; Red: Rabbit Slough; Purple: Hybrid.

recombination rate = 3.26 cM/Mb) across 6669 markers in 102 progeny.

Conservation of the Recombinational Landscape Among Populations and Crosses

Patterns of recombination were broadly conserved across the three maps. Recombination rates were generally higher toward one or both ends of a chromosome. This pattern was particularly clear along the longest chromosomes (chromosomes 4, 7, and 1), in which markers within 5 to 10 Mb of either chromosome end comprised the majority of the length of each genetic map. Among the remaining chromosomes, recombination rates increased toward a single chromosome end. This pattern was especially evident on chromosomes 5, 14, and 15.

Qualitative differences among the three genetic maps, were most often observed in the hybrid. As expected of an inversion heterozygote, we observed no recombination events within the breakpoints of a chromosomal inversion on chromosome 21 (Figure 4.3). The homologous regions of the Boot Lake and Rabbit Slough maps extend for 6.8 cM and 6.5 cM, respectively. While recombination was suppressed within the inversion, we observed elevated recombination rates outside of the inversion in the hybrid. This resulted in roughly equivalent chromosome 21 map lengths in all three crosses. Recombination rates along chromosomes 1 and 9 were more uniform in the hybrid map than in either Boot Lake or Rabbit Slough maps. In both cases, the hybrid map was comparable in length to the Boot Lake map but the distribution of recombination events was very different between the two maps: the majority of the genetic distance in both Boot Lake maps was concentrated into

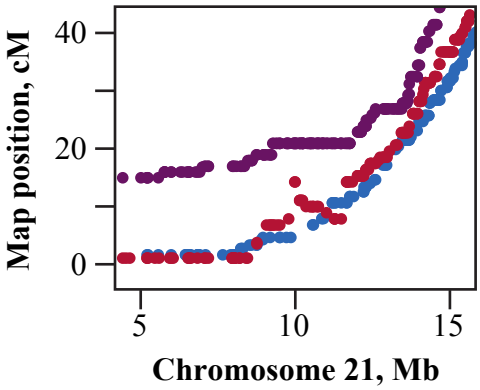

Figure 4.3. A chromosomal inversion on stickleback chromosome 21 suppresses recombination in an F_1 hybrid. Laboratory fish derived from Boot Lake (blue) and Rabbit Slough (red) populations harbor inverted chromosomal forms. The F_1 hybrid is heterozygous for the inversion. The y-axis starts at 0 cM for all maps; the hybrid map appears elevated due to higher recombination rate across the first 5 Mb of the chromosome.

less than 25% of the physical length of the chromosome while steadier recombination was observed across the entire length of both chromosomes in the hybrid.

Divergence is Associated with Regions of Low Recombination

Genetic diversity within and divergence between wild stickleback populations in Boot Lake and Rabbit Slough were strongly correlated with recombination rates estimated from experimental crosses (Table 4.3, Figure 4.4). Sequence diversity among and within both populations (π, π_{BT} and π_{RS}) was positively correlated with recombination rate. This correlation was largely driven by the lack of low-diversity genomic windows with high recombination rates, as the entire range of polymorphism levels observed genome-wide occurred in windows of average or below-average recombination rate.

Conversely, relative and absolute sequence divergence among Boot Lake and Rabbit Slough (F_{ST} and d_{xy}, respectively) were both negatively correlated with recombination rate (Figure 4.4C,D). The highest values of F_{ST} and d_{xy} were both associated with genomic windows of lowest recombination. Similar to estimates of sequence diversity, the correlation between divergence and recombination rate was not due to an overall shift in the distribution of divergence estimates in low- versus high-recombination windows, but rather a lack of extreme values of these statistics in windows of high recombination.

Genomic Regions of Highest Divergence Take Up a Small Fraction of the Genetic Map

To further explore the relationship between recombination rate and sequence divergence, we estimated the proportion of each map (physical genome, Boot genetic, Rabbit Slough genetic, F_1 genetic) with F_{ST} elevated above background levels. Using a resampling method that takes into account the distribution of RAD loci in the stickleback genome, and thus the amount of information per genomic window, we computed 99.9% confidence bands for F_{ST} values in 250-kb genomic windows. The average upper confidence limit was $F_{ST} = 0.23$, although outlier cutoffs varied among windows based on the number of RAD loci per window. We estimated that 19.7% of genomic windows,

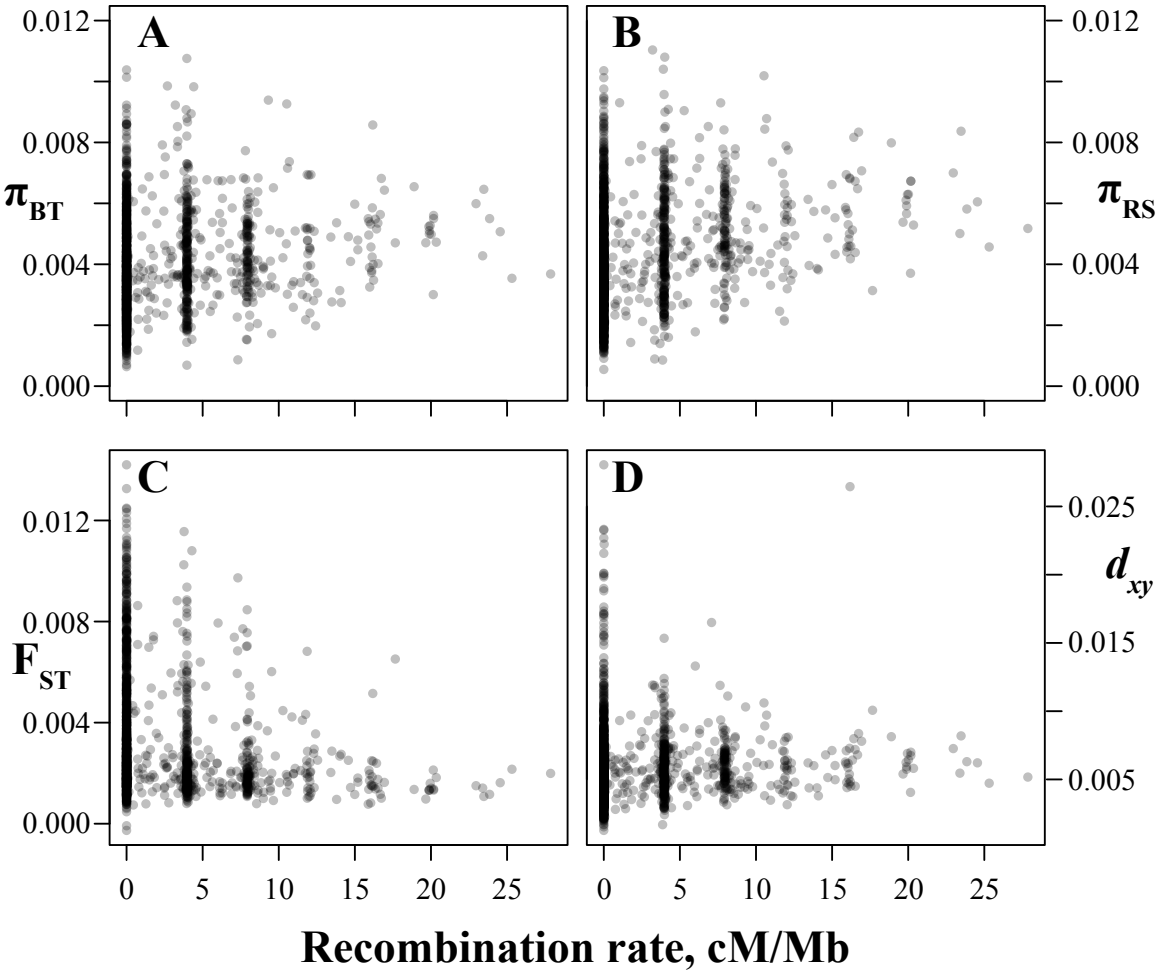

Figure 4.4. Correlations between recombination rate and genetic diversity within, and divergence between, Boot Lake and Rabbit Slough populations. Recombination rate is estimated from an F_1 hybrid between Boot Lake and Rabbit Slough laboratory lines. Each point represents a 100 kb non-overlapping genomic window. BT: Boot Lake; RS: Rabbit Slough. All Spearman's ρ correlations are significant at $p \leq 10^{-5}$. Coefficients are given in Table 4.3.

Table 4.3. Spearman's rank order correlations (ρ) between recombination rate (rows) and population genetic statistics. All correlations are significant at $p \leq 10^{-5}$.

Genetic map	π	π_{BT}	π_{RS}	d_{xy}	F_{ST}
Boot Lake	0.298	0.365	0.430	0.219	-0.386
Rabbit Slough	0.229	0.223	0.333	0.175	-0.268
F_1 hybrid	0.164	0.218	0.304	0.109	-0.250

comprising 86.0 Mb of the physical genome, were F_{ST} outliers. These outlier windows shrank to less than 10% of the overall map length when placed onto any of the three genetic maps (Figure 4.5; Boot Lake: 8.0%; Rabbit Slough: 9.1%; F_1: 8.1%). Genomic regions of highest divergence were further compressed in the genetic maps (Figure 4.5D, inset, Figure 4.6, Figure 4.7, Figure 4.8). For instance, some of the largest regions of divergence we observed — including those surrounding *Eda*, the major effect locus for lateral plate number (Colosimo, Hosemann et al. 2005), and the chromosome 21

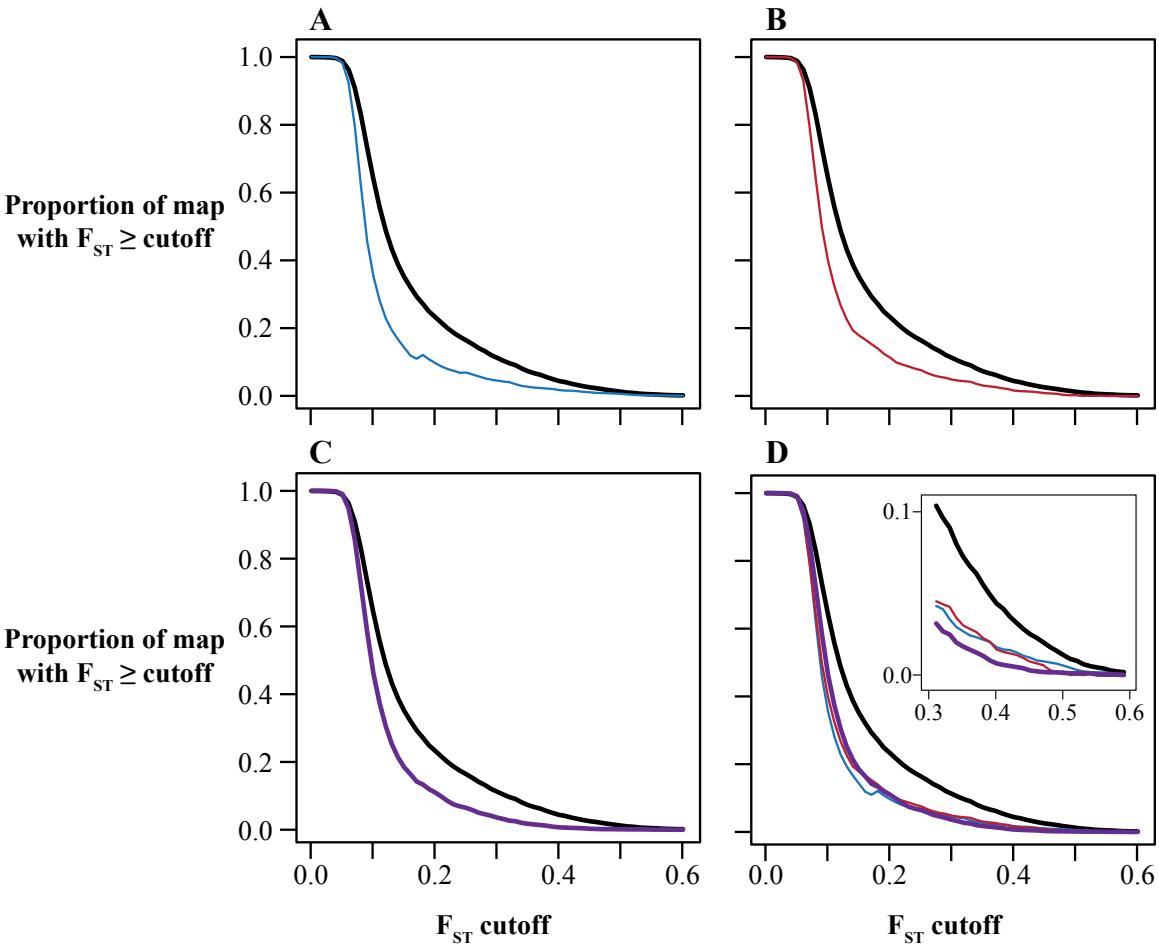

Figure 4.5. Marine-freshwater divergence affects a smaller proportion of the genetic map than the physical genome. Each curve represents the proportion of a map (y-axis), physical or genetic, in which average F_{ST} values exceed a given cutoff (x-axis). The reference genome is in black in all panels. (A) Boot Lake. (B) Rabbit Slough. (C) F_1 hybrid. (D) Three genetic maps compared to the reference genome. Inset: Expansion of curves representing the top 10% of the reference genome.

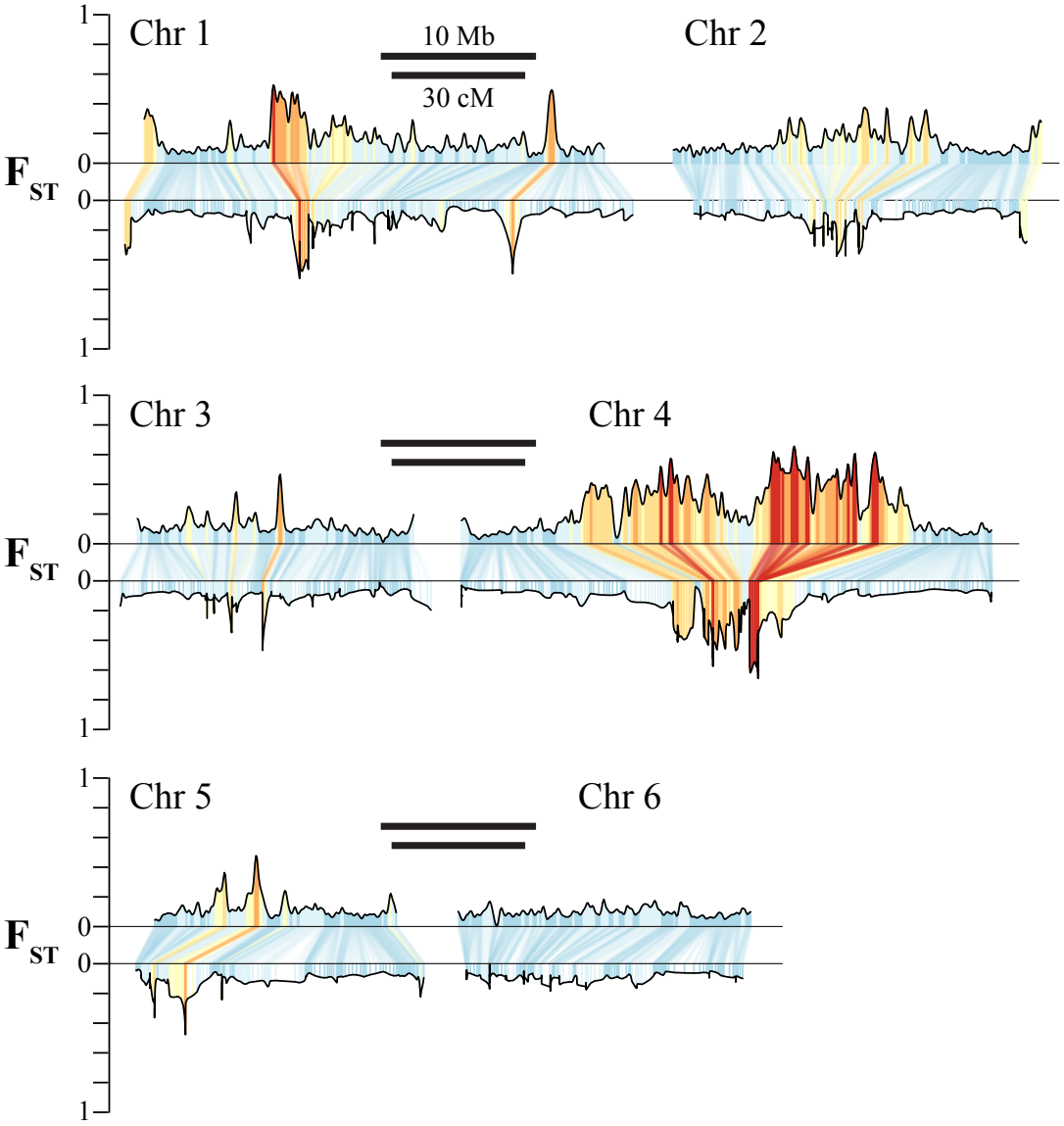

Figure 4.6. Genomic islands of divergence span narrow intervals on the F_1 hybrid genetic map — chromosomes 1 - 6. All maps are shown on the same scale. Panels are scaled to the length of the longest chromosome (chromosome 4) or the longest linkage group (in this instance linkage group 4). Lines between the physical (top of each plot) and genetic maps (bottom, y-axis reversed) show centers of homologous regions between the two maps (see Methods) and are colored by average F_{ST} within the window.

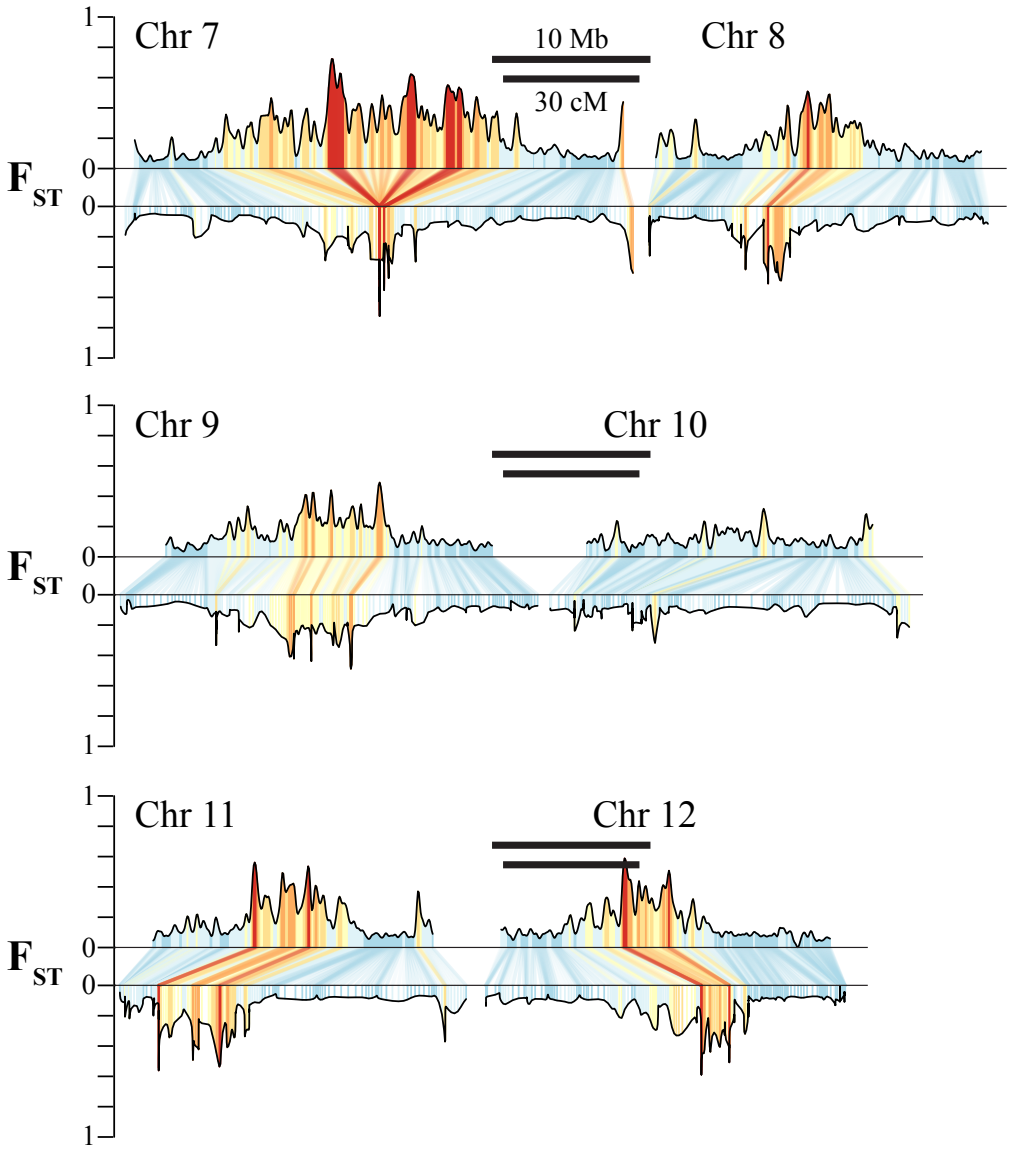

Figure 4.7. Genomic islands of divergence span narrow intervals on the F_1 hybrid genetic map — chromosomes 7 - 12. All maps are shown on the same scale. Panels are scaled to the length of the longest chromosome (chromosome 4) or the longest linkage group (in this instance linkage group 4). Lines between the physical (top of each plot) and genetic maps (bottom, y-axis reversed) show centers of homologous regions between the two maps (see Methods) and are colored by average F_{ST} within the window.

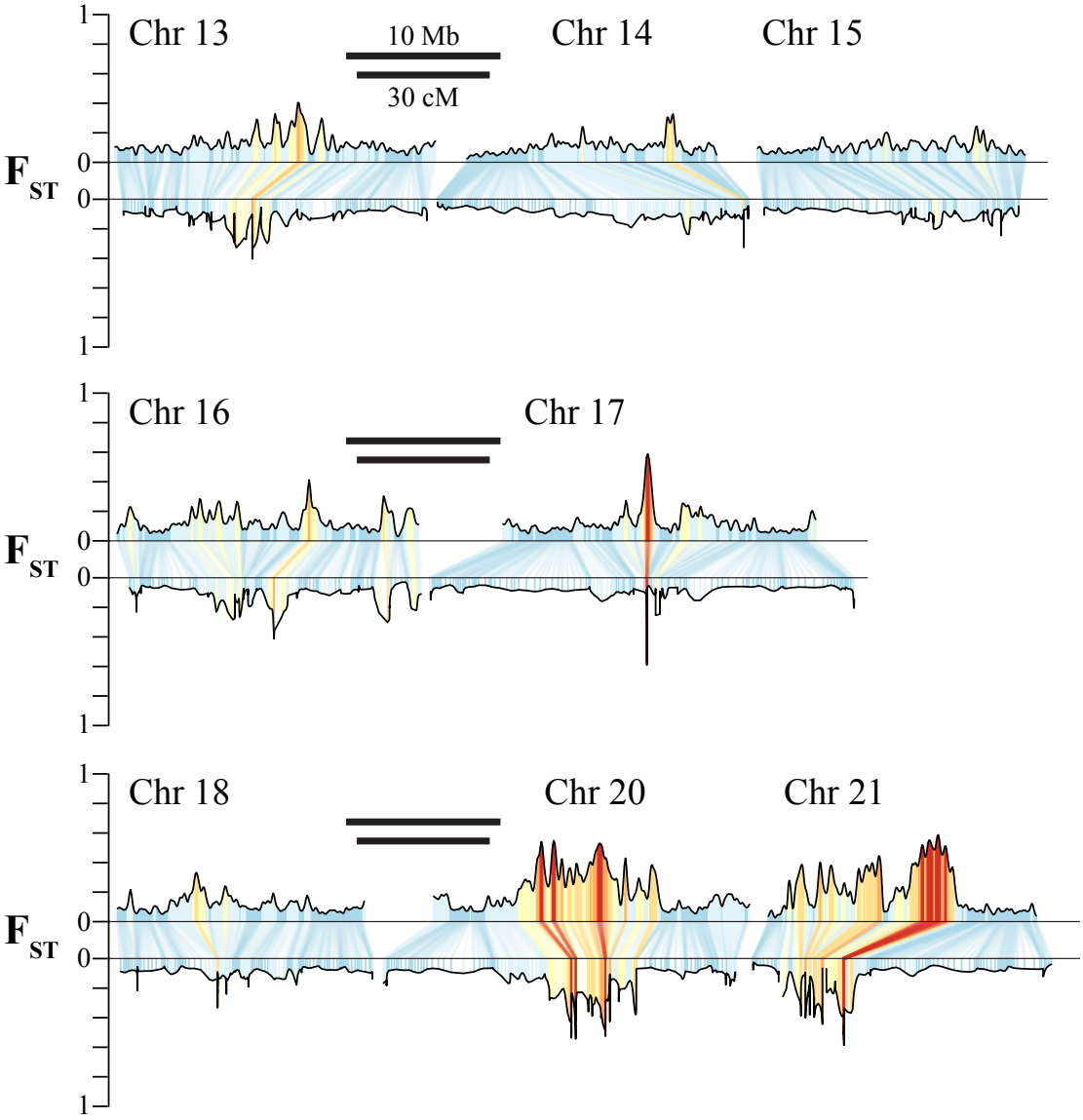

Figure 4.8. Genomic islands of divergence span narrow intervals on the F_1 hybrid genetic map — chromosomes 13 - 21. All maps are shown on the same scale. Panels are scaled to the length of the longest chromosome (chromosome 4) or the longest linkage group (in this instance linkage group 4). Lines between the physical (top of each plot) and genetic maps (bottom, y-axis reversed) show centers of homologous regions between the two maps (see Methods) and are colored by average F_{ST} within the window.

inversion — had average F_{ST} values in excess of $F_{ST} = 0.4$. Regions above this threshold spanned 4.4% of the physical genome (19.2 Mb) but only 0.7% of the F_1 genetic map (10 cM). This represents a six-fold decrease in the proportion of the map that is highly divergent between these two populations.

Genomic Divergence is Compressed on the Genetic Map and to a Greater Extent in the Hybrid

While the recombinational landscape of the stickleback genome is broadly conserved across the three genetic maps, evidence from natural stickleback populations suggests that reductions in recombination rate that are specific to F_1 hybrids are important for adaptive divergence with gene flow (Hohenlohe, Bassham et al. 2012, Jones, Grabherr et al. 2012, Roesti, Kueng et al. 2015). Specifically, chromosomal inversions are thought to be targets of selection because they suppress recombination between divergently adaptive haplotypes (Kirkpatrick 2010). We examined local recombination rate variation in genomic regions known to be under divergent selection to identify the extent to which reductions in recombination rate around these peaks of divergence are shared among genetic maps or unique to a specific map.

Low recombination rates across much of chromosome 4 were shared across genetic maps and condensed large regions on divergence on the reference genome into narrow intervals on the genetic map (Figure 4.9). Chromosome 4 consists of multiple peaks of divergence distributed across its length, with outlier regions comprising nearly 50% of the chromosome (46.8%, 16.0 Mb of 34.15 Mb). This entire region collapses into the centers of all three maps, spanning 10.9%, 21.8%, and 15.3% in the Boot Lake, Rabbit Slough, and F_1 hybrid maps, respectively. The most extreme F_{ST} outlier regions ($F_{ST} \geq 0.5$) appear as two or three narrow intervals, spanning less than 4 percent of all maps (Rabbit Slough: 2.3% [1.76 cM], Boot Lake: 3.2% [4.2 cM], Hybrid: 3.1% [3.7 cM]).

Marine-freshwater divergence on two other chromosomes was associated with variable recombination rates across genotypes, the lowest of which were observed in the

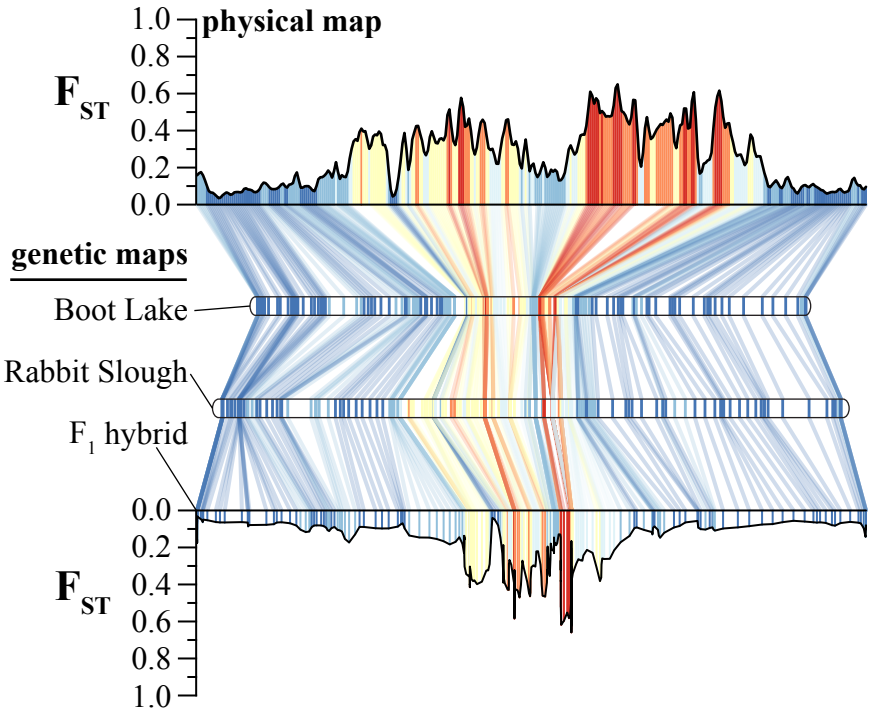

Figure 4.9. Genomic divergence on stickleback chromosome 4 is associated with globally low recombination rates. Vertical lines show centers of evenly-spaced windows on the reference genome (top line, see Methods) and are colored by average F_{ST} within the window. Lines connect to genetic positions of each window on Boot Lake (BT), Rabbit Slough (RS), and F_1 hybrid (F_1) genetic maps.

F_1 hybrid map (Figure 4.10, Figure 4.11). On chromosome 21, a large peak of marine-freshwater divergence residing within a chromosomal inversion collapsed to a single point on the hybrid map and segregated as a single mendelian locus. Unlike on chromosome 4, the divergence peak on chromosome 21 was not associated with reduced recombination in either the Boot Lake or Rabbit Slough maps relative to flanking genomic regions. Chromosome 7, which contained three distinct peaks of high divergence ($F_{ST} \geq 0.4$: 3.85 Mb of 30.85 Mb) and was moderately differentiated across much of its length (permuted confidence bands: 13.29 Mb, 43.1% of windows), displayed more complex patterns of inheritance among the crosses despite showing no evidence of structural variation. All three divergence peaks segregated as distinct loci in the Boot Lake cross and comprised 4.3% of the genetic map (7.0 cM). In the Rabbit Slough cross, the first two peaks collapse to a single locus, but we still observed recombination within

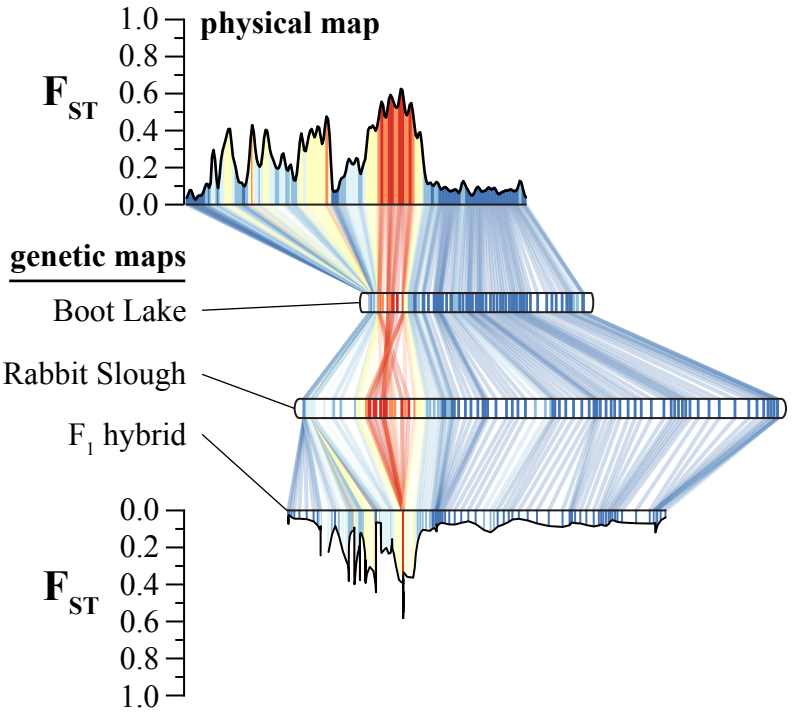

Figure 4.10. Genomic divergence on stickleback chromosome 21 is associated with a chromosomal inversion and suppression of recombination in an F_1 hybrid. Vertical lines show centers of evenly-spaced windows on the reference genome (top line, see Methods) and are colored by average F_{ST} within the window. Lines connect to genetic positions of each window on Boot Lake (BT), Rabbit Slough (RS), and F_1 hybrid (F_1) genetic maps.

the third peak (2.0% of the map [1.7 cM]) and between that peak and the first two. The F_1 map collapsed this third peak into a single mendelian locus, the overall effect being to reduce the three broad genomic regions of divergence separated by multiple megabases into two loci within a narrow interval (0.97 cM) on the genetic map.

DISCUSSION

In this study, we investigated the connections between the recombinational landscape and adaptive genomic divergence in the wild. Genomic divergence between locally adapted threespine stickleback populations was concentrated in regions of low recombination, supporting our core hypothesis. This correlation had the effect of condensing large regions of divergence — regions that span many megabases on the physical genome — into Mendelian loci on the genetic map. While the recombinational

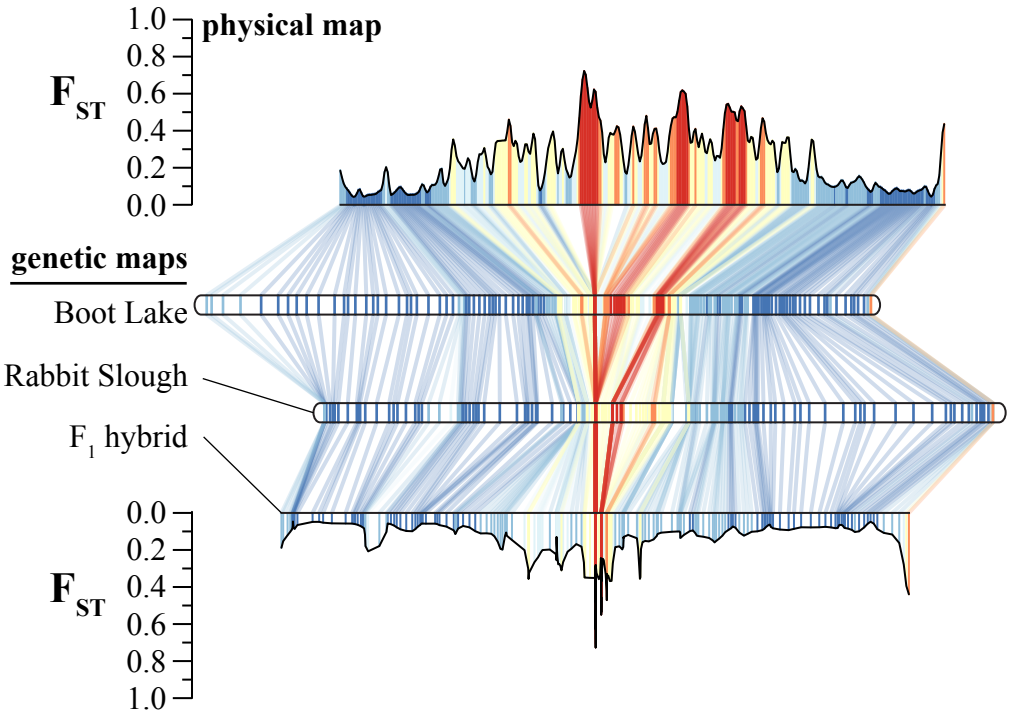

Figure 4.11. Genomic divergence on stickleback chromosome 7 is associated with variable inheritance patterns. Vertical lines show centers of evenly-spaced windows on the reference genome (top line, see Methods) and are colored by average F_{ST} within the window. Lines connect to genetic positions of each window on Boot Lake (BT), Rabbit Slough (RS), and F_1 hybrid (F_1) genetic maps.

landscape was broadly conserved between genetic maps, genomic divergence was further condensed in the F_1 hybrid in regions harboring chromosomal inversions as well as regions with no evidence of structural variation. Our results add to a growing body of literature highlighting the role of recombination in shaping adaptive divergence in the wild and suggest that variation in recombination rate among different genetic backgrounds has important consequences for the evolution of genomic divergence.

The Recombinational Landscape: Conservation and Variation

Variation in recombination rate occurs at all biological levels, from within a chromosome to between species (Smukowski and Noor 2011). The extent to which individuals within a species share a common recombinational landscape appears to be taxon-specific and is unknown for most taxa (Nachman 2002, Coop, Wen et al. 2008, Smukowski and Noor 2011, Comeron, Ratnappan et al. 2012). The patterns and

frequency of recombination events we observed here were largely conserved among the three mapping crosses and similar to previous reports in stickleback (Roesti, Moser et al. 2013, Glazer, Killingbeck et al. 2015). The overall map lengths in our crosses varied from approximately 1000 cM to nearly 1700 cM (Table 4.1), which are comparable to genetic maps generated from threespine stickleback populations from western North America (Glazer, Killingbeck et al. 2015) and central Europe (Roesti, Moser et al. 2013). The distributions of recombination events along chromosomes were also generally consistent between crosses in this study and between this study and other reports (Roesti, Moser et al. 2013, Glazer, Killingbeck et al. 2015). Recombination was typically more frequent toward one or both chromosome ends, and larger chromosomes had more symmetrical patterns of recombination, perhaps indicating obligate crossovers on both arms of meta- and acrocentric chromosomes (compare chromosome 5 and chromosome 7 in Figure 4.2).

While conservation was the norm among our mapping crosses, especially between the Boot Lake and Rabbit Slough crosses, recombination rates for some chromosomes varied substantially in the F_1 hybrid (Figure 4.2, chromosomes 1 and 21). The individual used for the genetic mapping was heterozygous for at least one chromosomal inversion, on chromosome 21 (Figure 4.3), and we saw the expected suppression of recombination throughout this region. Intriguingly, inversion heterozygosity is known to elevate recombination rates elsewhere on the chromosome on which it resides (Schultz and Redfield 1951), and we indeed found elevated recombination rate along the first ~5 Mb of chromosome 21 in the hybrid relative to either the Boot Lake or Rabbit Slough maps. Chromosomal inversions that differentiate marine and freshwater stickleback populations are also present on chromosomes 1 and 11 (Jones, Grabherr et al. 2012). The resolution of our genetic maps, which was limited by mapping family sizes and the distribution of *SbfI* restriction sites in the stickleback genome, was not high enough to confirm inversion genotype in either case. Despite this shortcoming, the distribution of recombination events on chromosome 1 in the hybrid was qualitatively different from the other maps: recombination occurred steadily across the length of chromosome 1 in the hybrid, but was biased toward the ends of the chromosome in both parental maps.

Continents of Divergence Exist Only on the Physical Map

By breaking and establishing physical linkage among adaptive and neutral genetic variants, recombination mediates the interactions between selection and gene flow when adaptation occurs across an ecologically diverse landscape. The importance of these three forces is now recognized in a range of taxa (Haasl and Payseur 2016, Martin, Most et al. 2016) but how they interact to generate observed patterns of heterogenous genomic divergence remains, for the most part, mysterious (Nosil and Feder 2012). In the threespine stickleback, genomic divergence during adaptation to freshwater habitats is highly heterogenous, involving large regions of multiple chromosomes and, in some instances, multiple regions on the same chromosome (Hohenlohe, Bassham et al. 2010, Jones, Grabherr et al. 2012, Chapter II of this work). In addition, adaptive divergence primarily involves standing genetic variation, much of which evolved anciently and is shared among many freshwater populations (Colosimo, Hosemann et al. 2005, Roesti, Gavrilets et al. 2014, Chapters II and III of this work).

Our results show that, while adaptive divergence can affect allele frequency trajectories across large proportions of the length of a chromosome, these islands — or continents (Nosil and Feder 2012) — of divergence collapse into narrow peaks on the genetic map (Figure 4.6). Some of the largest continents of divergence span small fractions of the genetic map, leaving the vast majority of the map length of these chromosomes nearly untouched by divergent selection. These same genomic regions show consistent parallelism among independently derived freshwater populations (Hohenlohe, Bassham et al. 2010), suggesting that the recombinational landscape helped to structure haplotypic variation in the marine population that was directly ancestral to the extant freshwater populations. On chromosome 4, we found that a region of divergence that comprises over 25% of the physical length of the chromosome (~10 Mb of 34 Mb) spans only ~2.5% of the genetic map in the F_1 hybrid and showed similar patterns of inheritance in both the Boot Lake and Rabbit Slough maps (Figure 4.11). Genomic divergence became even more punctuated on chromosome 7. In the F_1 hybrid map, three strongly differentiated genomic regions, each of around 1 Mb in length,

segregated as two distinct loci on a span of the genetic map totaling no more than 4 cM (Figure 4.11). Due to their low frequency in the marine populations, freshwater-adaptive haplotypes are likely found almost exclusively in the heterozygous form. The recombinational landscape influencing their persistence in that habitat is probably most similar to the landscape we observed in the F_1 hybrid genetic map. The additional reduction in recombination rates in divergence islands may help prevent their breakup onto high-frequency marine genetic backgrounds in that environment.

The transmission of genomic islands was not entirely uniform between genetic maps, but showed variation with evolutionarily significant implications. In an exception to the rule of discrete inheritance across maps, the divergence island on chromosome 21 spanned 6.8 cM on the Boot Lake genetic map and 6.5 cM on the Rabbit Slough map (Figure 4.10). This represents approximately 10% of the total length in both cases and, despite its smaller physical size (~1.7 Mb), spans over twice the genetic distance of the largest chromosome 4 island. This region, however, harbors a chromosomal inversion that entirely suppresses recombination in the F_1 hybrid, discretizing inheritance in a genotype-specific manner. This inversion is consistently differentiated between marine and freshwater populations throughout the species range. Chromosome 21 also harbors highly pleiotropic quantitative trait loci (Albert, Sawaya et al. 2008, Miller, Glazer et al. 2014), which may in reality be due to multiple variants with distinct effects on phenotype that are genomically clustered and locked into a single genetic unit by the inversion.

Genotype-specific collapses also occurred in entirely collinear genomic regions. On chromosome 7, in which islands collapsed entirely to Mendelian loci in the F_1, patterns of inheritance were different in all three maps. The three highest peaks on this chromosome formed three regions of the Boot Lake map, with only the first peak occupying a single genetic position. In the Rabbit Slough map, the first two peaks collapsed into a single genetic locus, though we still observed recombination events within the third. This third peak then further contracted in the hybrid map, segregating as a single locus as mentioned above.

The Recombination Landscape Crystallizes Genomic Divergence and Facilitates Phenotypic Divergence

We suggest that the the recombinational landscape plays two important roles in the maintenance of adaptive divergence in the threespine stickleback. First, lack of recombination prevents the breakup of freshwater-adaptive haplotypes when they are present at low frequency in the marine environment. In the threespine stickleback, unlike in models where genomic divergence accumulates as a function of the timescale of population divergence (Feder, Gejji et al. 2012, Nosil and Feder 2012, Wolf and Ellegren 2016), genomic divergence predates population divergence (Chapters II and III of this work). Even in relatively young freshwater populations, including the ~12,000 year old Boot Lake population studied here, much of the genomic variation driving adaptive divergence evolved millions of years ago. This variation persists at low frequency in marine populations, presumably as the result of gene flow with existing freshwater populations (Schluter and Conte 2009). While the average residence time of a freshwater haplotype in the marine environment is unknown, freshwater variation will exist almost exclusively in the heterozygous form and recombination with marine haplotypes should quickly homogenize variation. Genomic regions with generally low levels of recombination are then hospitable environments for the establishment of new adaptive variation and the maintenance of existing haplotypes. Where the recombination rate is higher and thus less favorable to the maintenance of divergent haplotypes, suppression of recombination in heterozygotes may evolve either through genomic structural variation (*e.g.* chromosomal inversions) or as-yet-unidentified recombination modifiers.

Second, the recombinational landscape simplifies the genetic architecture of divergence, increasing the rate of adaptive evolution to freshwater habitats. Rapid evolution appears to be a common phenomenon in this species, with phenotypically freshwater populations evolving from marine ancestors within decades (Lescak, Bassham et al. 2015). The large contribution of standing genetic variation to divergence in stickleback accelerates adaptive evolution compared to the waiting time required for

adaptation from new mutation (Barrett and Schluter 2008, Schluter and Conte 2009, Chapter II of this work). The nature of the standing genetic variation may still impede this process because adaptation likely involves complex genetic architectures and traits with polygenic inheritance (Barton and Keightley 2002, Burke and Long 2012). If adaptive variation exists on many genetic backgrounds, selection would need to act independently on each locus, the likely outcome being the loss of smaller-effect alleles and interference between multiple haplotypes (Hill and Robertson 1966). The recombinational landscapes we observed here discretize the inheritance of large genomic regions, each of which may contain many divergently adaptive alleles (Chapters II and III of this work). This effectively creates a small number of loci with large effects on fitness, simplifying the genetic architecture of adaptation.

When is the Recombinational Landscape a Mediator of Adaptive Divergence, and When is it a Result?

The recombinational landscape evolves (Kirkpatrick 2010, Smukowski and Noor 2011). There is plentiful evidence that recombination rates not only vary throughout the genome but that there is heritable variation for recombination rate heterogeneity among individuals in a population (Coop, Wen et al. 2008, Dumont, Broman et al. 2009, Kong, Thorleifsson et al. 2010). But models of genome evolution and divergence often treat the recombinational landscape as static (Charlesworth, Nordborg et al. 1997, Feder and Nosil 2010, Feder, Gejji et al. 2012). This is often a necessary simplifying assumption for certain evolutionary models but nonetheless provides an incomplete picture of potential evolutionary trajectories. While our genetic mapping data do not provide direct evidence of the evolution of the recombination rate, we observed variation among genomic islands of divergence that suggests distinct interactions between selection and recombination, including evidence against and for the adaptive evolution of the recombination rate.

Genomic divergence and recombination rate along chromosome 4 suggest that divergence has accumulated in a region of intrinsically low recombination rate (Figure 4.11). We observed a striking correspondence between marine-freshwater divergence in

wild-caught fish with low recombination rates in our lab crosses. This reduction in recombination rate relative to weakly differentiated regions at the ends of chromosome 4 was conserved across the three genetic maps, indicating little influence of genetic background on recombination rate. Both of these patterns are also shared with other stickleback populations and genetic maps: divergence on chromosome 4 is shared among freshwater populations throughout the species range (Jones, Grabherr et al. 2012) and multiple genetic maps made from geographically distant stickleback populations have identified similar distributions of recombination (Roesti, Moser et al. 2013, Glazer, Killingbeck et al. 2015).

In contrast, we identified several islands of divergence with more complex patterns of recombination between genetic crosses, which may be the footprint of an adaptively evolving recombinational landscape. Our genetic maps of chromosome 21 highlight (1) the recombination-suppressing effects of a chromosomal inversion and (2) the implications for genomic divergence had the inversion not evolved (Figure 4.10). Unlike chromosome 4, we observed free recombination across the divergence island in the center of this chromosome in both the Boot Lake or Rabbit Slough maps which was comparable to or greater than recombination rates in other regions of the chromosome. Only in the F_1 hybrid, which was heterozygous for the inversion that isolates marine and freshwater chromosomal forms (Jones, Grabherr et al. 2012), did we observe a complete suppression of recombination which isolated marine and freshwater haplotypes. Given the strong association between genomic divergence and recombination rate in the rest of the genome, it is safe to suggest that, in the absence of a chromosomal inversion, the recombinational landscape on chromosome 21 would not have been conducive to the extensive marine-freshwater divergence we and other have observed.

Genotype-specific changes in recombination were not limited to locations of genomic structural variation, however. Chromosome 7 contained three peaks of divergence on the reference genome, spanning nearly a third the length of the chromosome (Figure 4.11). These peaks collapsed uniquely onto each genetic map. Three genomic peaks corresponded to three distinct regions on the Boot Lake map but only two

even narrower intervals on the Rabbit Slough map. The F_1 hybrid map was not intermediate between these two, but showed the sharpest reduction of these genomic islands into two loci less than 3 cM apart. The mechanism behind these altered landscapes is unknown in stickleback and may include sequence-specific recombination modifiers (Jensen-Seaman, Furey et al. 2004). Given the often extensive sequence divergence between marine and freshwater chromosomes (Chapters II and III of this work), variation in sequence homology may also impact the distribution of resolved crossovers in a hybrid genome.

Conclusions

The recombinational landscape of the threespine stickleback condenses widespread genomic divergence into discrete loci, which are largely inherited as single units. During adaptive divergence, divergently adapted haplotypes are bundled into a relatively few number of Mendelian loci, each with a putatively large effect on fitness, which facilitates the assembly of a common freshwater genome from low-frequency variation present in the marine ancestor. This both simplifies the genetic architecture of adaptation and increases the rate of divergence. On the longer evolutionary timescales over which genomic divergence has accumulated, the recombinational landscape appears to have largely dictated where in the genome adaptive variation may establish. Divergently adaptive alleles which evolved in regions of higher recombination rate subsequently failed to establish or are geographically restricted to the populations in which they evolved. Alternatively, the recombinational landscape itself may have evolved, as indicated by the presence of three chromosomal inversions isolating freshwater and marine haplotypes. The demographic and evolutionary history of the threespine stickleback have led to unique patterns of polymorphism, divergence, and recombination, but we expect the forces shaping patterns to be representative of a wide range of taxa experiencing divergence with gene flow, providing insight into genome evolution along the continuum of adaptive divergence.

CHAPTER V

CONCLUSIONS

The variation we observe in nature is the result of an intricate interplay between neutral and selective evolutionary processes taking place across a diverse and ever-changing geography (Darwin 1859, Lewontin 1974, Endler 1977, Endler 1986, Lescak, Bassham et al. 2015). Over the past century, continual advancements in the technologies used to detect variation (Morgan, Sturtevant et al. 1920, Dobzhansky and Sturtevant 1938, Hubby and Lewontin 1966, Kreitman 1983, Begun, Holloway et al. 2007), and in conceptual and analytical models used to interpret it (Fisher 1918, Lewontin 1974, Kingman 1982, Kimura 1983, Tajima 1983, Crow 2008, Flaxman, Wacholder et al. 2014), have revealed ever more abundant genetic variation and more complex interactions between geography, selection, and genome structure needed to explain its abundance. Lately, the field of population genomics has taken advantage of next-generation sequencing technologies — a quantum leap in the ability to assay genetic variation — to describe genome-wide variation in a variety of natural systems (Brawand, Wagner et al. 2014, Yeaman, Hodgins et al. 2016). However, except in the limited number of model systems with existing high-quality genomic resources (Linnen, Kingsley et al. 2009, Huerta-Sánchez, Jin et al. 2014), these studies are limited to examining the sorting of SNP variation due to geography and selection that has likely occurred in the relatively recent history of the species. The deeper patterns of evolutionary history influencing this variation, therefore, remain unknown. In this work, I advanced population genomic inference in non-model systems by extending restriction site-associated DNA sequencing to efficiently detect haplotype structure at thousands of anonymous genomic loci. I used this technique to describe patterns of genomic variation and reconstruct the evolutionary histories behind that variation in threespine stickleback, demonstrating that the genomic variation involved in recent adaptive events and differentiating young populations in this species has a history that goes back millions of

years and is maintained by natural selection that varies across geography and across the genome.

These results confirm the suspicions of many recent papers on threespine stickleback, namely that the extensive genomic parallelism associated with marine-to-freshwater transitions has, at its core, a suite of alleles that are identical by descent (Colosimo, Hosemann et al. 2005, Schluter and Conte 2009, Jones, Grabherr et al. 2012, Roesti, Gavrilets et al. 2014, Lescak, Bassham et al. 2015). My results then extend this model, demonstrating a history of molecular divergence extending throughout the species' history. I also show for the first time that freshwater populations harbor the majority of the genetic variation present in the species and provide evidence of recent gene flow between threespine and ninespine stickleback. Below, I discuss the relevance of these results in the context of three dimensions of evolutionary genetics: genomic space, geographic space, and time.

The Genome: Integration and Variation

The haploid genome is the core unit of selection (Chapter 6 in Lewontin 1974). This is because (1) the physical linkage of genetic variants onto chromosomes results in their coinheritance (Morgan, Sturtevant et al. 1920), and (2) the effect of an allele on fitness depends on the genetic background on which it resides (Fisher 1930, Phillips 2008). Under equilibrium conditions — given free recombination, the absence of selection, and negligible input of new variation — the organization of genetic variation in a genome would need not be considered. We now know that adaptation in the wild, however, can often violate these assumptions. In the stickleback, non-equilibrium selective and demographic processes are rampant, and patterns of genetic variation can only be understood in their genomic context. The body of my work demonstrates how variation in recombination forms cohesive evolutionary units from large genomic regions, maintaining genetic and genomic variation linked to adaptive alleles. My findings place stickleback evolutionary genomics in a new light.

The stickleback's recombination landscape provides a proximate mechanism for the maintenance of, and concerted selection on, large regions of adaptive divergence (Chapter IV and figures therein). The consolidation of adaptive variation in regions of low recombination promotes the coinheritance of adaptive alleles at multiple, genomically distinct loci, creating a single, integrated unit of inheritance. Selection then acts on the sum total of the fitness effects of all individual variants that are physically and genetically linked to one another. Ideally, one would estimate recombination landscapes in multiple individuals in wild populations during adaptation. While that is far beyond the scope of this work, it does appear that the simplification of the genetic architecture of adaptation plays roles in maintaining adaptive haplotype variation, and increasing the rate of adaptation and the probability of parallelism across populations.

The structuring of variation onto distinct genetic backgrounds also has important consequences for linked variation. In Chapter III, I showed that chromosomes carrying freshwater alleles at divergently selected loci were more genetically diverse than those carrying marine alleles, but only in genomic regions adjacent to loci under divergent selection. Viewing the genome as a whole is key here, too, because the increase in diversity was observed in the aggregate: RAD loci that were otherwise unassociated with one another, and often on separate chromosomes, shared a common signal in large part because they were associated with loci under long-term selection. With increasing physical distance from a locus under divergent selection, the protective effects of selection broke down, homogenizing variation across chromosomal types.

Conceptualizing the genome as an integrated unit is not new (see Lewontin 1974), but over the intervening decades we have forgotten this reality through the focus of empirical studies of a small number of largely independent loci (McDonald and Kreitman 1991, Pritchard, Stephens et al. 2000). The field of empirical evolutionary genetics is reaching a moment when we can explicitly test these ideas. Over the past decade-and-a-half, population genomic studies have made great strides to understand patterns of variation in nature (Luikart, England et al. 2003, Begun, Holloway et al. 2007, Davey, Hohenlohe et al. 2011). These studies can easily end up, however, as genome-wide

studies of *genetic* variation: genomic regions are treated as independent units if they are further from one another than the average decay of linkage disequilibrium. Dense genomic sampling of simple genetic variants is used to ensure all such independent units are identified. While the treatment of the genome as a collection of units may be adequate in some instances as a first approximation, it prevents us from addressing other fundamental questions including, What is the realized variation in gametic types available to selection? How much variation in recombination rate exists in a population, and how quickly can recombination rate variation respond to selection? And is epistatic selection strong enough and common enough to structure genomic variation? Technological advances are no longer the limiting factor. Long-read and single-cell sequencing technologies are only years away from being broadly available for population and evolutionary genomic analyses. The limits are now the ability to design and execute appropriate experiments and to develop appropriate theory and models to test against.

GEOGRAPHY: ISOLATION AND CONNECTION

Geographic and ecological variation combine to structure and maintain genetic variation across a landscape (Wright 1932, Slatkin 1993, Andrew, Ostevik et al. 2012). This leads, in varying degrees, to a metapopulation — a web of semi-isolated yet interconnected populations. A number of my key results, including the preponderance of anciently derived adaptive variation and the asymmetric distribution of variation across marine and freshwater haplotypes, make little sense without invoking a substantial amount of population structure within and among habitat types. Populations structured by ecology have maintained variation much longer than would be expected otherwise. And within each habitat type, variable selection regimes (more stringent in the marine habitat) and geographic structuring (greater structure among freshwater habitats) have resulted in asymmetries in genetic diversity harbored on marine and freshwater chromosomes. While the stickleback may not represent the average species — though maybe it does — the influences of ecological and geographical heterogeneity on patterns of genomic variation will no doubt find analogs in other species. In general, studies seeking to identify the

sources of genetic variation, and understand its distribution across the genome, must account for contributions multiple populations.

The sources of genetic variation available to a population are not limited to traditionally recognized species boundaries (Huerta-Sánchez, Jin et al. 2014, Stankowski and Streisfeld 2015, Wallbank, Baxter et al. 2016). The evidence of recent gene flow between threespine and ninespine stickleback species (Chapter III) adds to a growing number of studies that, together, are expanding the known sources of genetic variation to include multiple species. Interspecific gene flow has been documented in insects (Fontaine, Pease et al. 2015), plants (Stankowski and Streisfeld 2015), and even our own lineage (Huerta-Sánchez, Jin et al. 2014). Together with incomplete lineage sorting (Tajima 1983, Cui, Schumer et al. 2013, Pease, Haak et al. 2016), gene flow between morphologically and genetically distinct species blurs the boundaries between microevolutionary and macroevolutionary processes. Species become analogous to populations that are almost — but not quite — entirely geographically isolated from one another. We cannot discount evolutionarily distant sources of variation, even when seeking to understand processes occurring on ecological timescales.

Time: Why Rapid Adaptation Takes Millions of Years

One of the most astounding facts to have come from studies of adaptation in the wild is how quickly it can proceed (Grant and Grant 2002, Lescak, Bassham et al. 2015). Rapid responses to selection are not limited to species with large population sizes and short generation times — in other words, the usual suspects under classical evolutionary theory (Fisher 1930, Lenormand, Bourguet et al. 1999, Gillespie 2004) — but have been observed across vertebrate lineages, including fish (Lescak, Bassham et al. 2015, Malinsky, Challis et al. 2015), reptiles (Brodie 1992, Grant and Grant 2002), and mammals (Domingues, Poh et al. 2012). The rate of adaptation has been variously attributed to an abundance of standing genetic variation for quantitative traits (Brodie 1992, Grant and Grant 2002, Miller, Metcalf et al. 2015) and/or a simple genetic basis of adaptation, with few alleles of large effect on phenotype and fitness (Barrett, Rogers et al.

2008) — combined, in either case, with strong selection (Kingsolver, Hoekstra et al. 2001, Grant and Grant 2002, Barrett, Rogers et al. 2008). My findings emphasize an additional consideration: that the genetic variation itself comes pre-shaped by a much longer, unobserved history of selection.

The maintenance of variation, at least in the threespine stickleback, is due to the persistence of variable selective pressures through time, even as the populations experiencing those pressures have partially or entirely turned over (Figure 5.1). The marine stickleback population has been a constant presence since its initial evolution (Bell 1994, Bell and Foster 1994). In contrast, freshwater populations have come and gone over the course of evolutionary time. The selective pressures favoring freshwater-

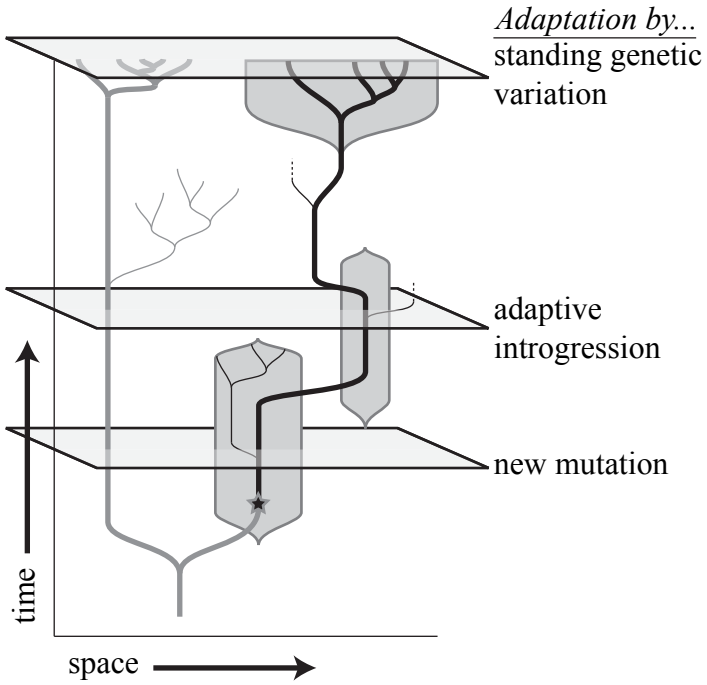

Figure 5.1. Hypothetical history of a locus under divergent selection. Gray and white backgrounds represent different selective environments. Darker lineages are selected for in the gray environments and against in the white environment. Bold lineages are ancestral to those sampled at the present (top). Star denotes the original adaptive mutation, leading to adaptation via new mutation in an ancestral population (the first slice in time). Gene flow among geographically isolated populations later leads to adaptation via adaptive introgression. Finally, a lineage survives in the alternative environment to be present in the founding pool of variation in a new population, leading to adaptation from standing genetic variation.

adaptive variation, however, have likely persisted in some form throughout the history of the threespine stickleback lineage. Initial adaptive mutations that arose early on in the species' history, by traversing geography via gene flow among populations, have been maintained until the present, even though the population in which they arose have long since gone extinct. Over the millennia, additional adaptive mutations have accumulated on persistent genetic backgrounds, perhaps aided in their survival by a favorable, and even adaptable, recombination landscape. The result, which we observe today, is drastic phenotypic change happening over decades driven by strong selection on genomic variation that has evolved piecemeal throughout the entire history of the species.

So does adaptive evolution happen rapidly in the stickleback? No doubt strong selection can change allele frequencies over few generations in a population of interest. From the standpoint of standard population genetic theory, then, the answer is yes. But that selection is acting on a diverse and highly structured genomic architecture, which has evolved adaptively over timescales that dwarf recent events. From the standpoint of the molecular evolution of the genome, then, the answer is clearly no. That we arrive at different conclusions by studying different parts of the process highlights the need for empirical and theoretical integration across spatial and temporal scales to fully understand adaptation in the wild.

REFERENCES CITED

Albert, A. Y. K., S. Sawaya, T. H. Vines, A. K. Knecht, C. T. Miller, B. R. Summers, S. Balabhadra, D. M. Kingsley and D. Schluter. "The genetics of adaptive shape shift in stickleback: Pleiotropy and effect size." *Evolution* 62 (2008): 76-85.

Aldenhoven, J. T., M. A. Miller, P. S. Corneli and M. D. Shapiro. "Phylogeography of ninespine sticklebacks (*Pungitius pungitius*) in North America: glacial refugia and the origins of adaptive traits." *Molecular Ecology* 19 (2010): 4061-4076.

Altshuler, D., L. D. Brooks, A. Chakravarti, F. S. Collins, M. J. Daly, P. Donnelly, R. A. Gibbs, J. W. Belmont, A. Boudreau, S. M. Leal, P. Hardenbol, S. Pasternak, D. A. Wheeler, T. D. Willis, F. L. Yu, H. M. Yang, C. Q. Zeng, Y. Gao, H. R. Hu, W. T. Hu, C. H. Li, W. Lin, S. Q. Liu, H. Pan, X. L. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. R. Zhang, H. B. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, Y. Shen, Z. J. Yao, W. Huang, X. Chu, Y. G. He, L. Jin, Y. F. Liu, Y. Y. Shen, W. W. Sun, H. F. Wang, Y. Wang, Y. Wang, Y. Wang, X. Y. Xiong, L. Xu, M. M. Y. Waye, S. K. W. Tsui, H. Xue, J. T. F. Wong, I. L. M. Galver, J. B. Fan, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, P. Y. Kwok, D. M. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, P. C. Sham, Y. Q. Song, P. K. H. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, B. Fry, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. Varilly, L. D. Stein, L. Krishnan, A. V. Smith, G. A. Thorisson, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. H. Guan, H. M. Munro, Z. H. S. Qin, D. J. Thomas, G. McVean, L. Bottolo, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, H. C. Zhang, C. Q. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Z. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, D. J. Richter, L. Ziaugra, B. W. Birren, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L. Archeveque, G. Bellemare, K. Saeki, H.

G. Wang, D. C. An, H. B. Fu, Q. Li, Z. Wang, R. W. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, C. R. Bird, M. S. Guyer, P. J. Nailer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, J. Witonsky, L. F. Zacharia, K. Kennedy, R. Jamieson, J. Stewart and I. H. Consortium. "A haplotype map of the human genome." *Nature* 437 (2005): 1299-1320.

Andrew, R. L., K. L. Ostevik, D. P. Ebert and L. H. Rieseberg. "Adaptation with gene flow across the landscape in a dune sunflower." *Molecular Ecology* 21 (2012): 2078-2091.

Andrew, R. L. and L. H. Rieseberg. "Divergence is focused on few genomic regions early in speciation: Incipient speciation of sunflower ecotypes." *Evolution* 67 (2013): 2468-2482.

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart and P. A. Hohenlohe. "Harnessing the power of RADseq for ecological and evolutionary genomics." *Nature Reviews Genetics* 17 (2016): 81-92.

Arendt, J. and D. Reznick. "Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation?" *Trends in Ecology and Evolution* 23 (2008): 26-32.

Arnold, B., R. B. Corbett-Detig, D. Hartl and K. Bomblies. "RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling." *Molecular Ecology* 22 (2013): 3179-3190.

Aronesty, E. "ea-utils: Command-line tools for processing biological sequencing data." (2011)

Avise, J. C. *Phylogeography: The History and Formation of Species*. Cambridge, MA: Harvard University Press. 2000.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko and E. A. Johnson. "Rapid SNP discovery and genetic mapping using sequenced RAD markers." *PLoS ONE* 3 (2008): 1-7.

Barrett, R. D. H., S. M. Rogers and D. Schluter. "Natural selection on a major armor gene in threespine stickleback." *Science* 322 (2008): 255-257.

Barrett, R. D. H. and D. Schluter. "Adaptation from standing genetic variation." *Trends in Ecology and Evolution* 23 (2008): 38-44.

Barton, N. H. and P. D. Keightley. "Understanding quantitative genetic variation." *Nature Reviews Genetics* 3 (2002): 11-21.

Beck, E. A., A. C. Thompson, J. Sharbrough, E. Brud and A. Llopart. "Gene flow between *Drosophila yakuba* and *Drosophila santomea* in subunit V of cytochrome c oxidase: A potential case of cytonuclear co-introgression." *Evolution* 69 (2015): 1973-1986.

Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y. P. Poh, M. W. Hahn, P. M. Nista, C. D. Jones, A. D. Kern, C. N. Dewey, L. Pachter, E. Myers and C. H. Langley. "Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*." *PLoS Biol* 5 (2007): e310.

Bell, M. A. “Paleobiology and evolution of the threespine stickleback.” In *The Evolutionary Biology of the Threespine Stickleback*, ed. M. A. Bell and S. A. Foster, 438-471. New York: Oxford University Press, 1994.

Bell, M. A. and S. A. Foster. *The Evolutionary Biology of the Threespine Stickleback*. New York: Oxford University Press, 1994.

Bell, M. A. and S. A. Foster. “Evolutionary inference: the value of viewing evolution through stickleback-tinted glasses.” In *The Evolutionary Biology of the Threespine Stickleback*, ed. M. A. Bell and S. A. Foster, 472-486. New York: Oxford University Press, 1994.

Bell, M. A. and S. A. Foster. “Introduction to the evolutionary biology of the threespine stickleback.” In *The Evolutionary Biology of the Threespine Stickleback*, ed. M. A. Bell and S. A. Foster, 1-27. New York: Oxford University Press, 1994.

Bell, M. A., J. D. Stewart and P. J. Park. "The World's Oldest Fossil Threespine Stickleback Fish." *Copeia* 2009 (2009): 256-265.

Bernatchez, L. "On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes." *Journal of Fish Biology* 89 (2016): 2519-2556.

Blount, Z. D., J. E. Barrick, C. J. Davidson and R. E. Lenski. "Genomic analysis of a key innovation in an experimental *Escherichia coli* population." *Nature* 489 (2012): 513-518.

Boughman, J. W. "Divergent sexual selection enhances reproductive isolation in sticklebacks." *Nature* 411 (2001): 944-948.

Boughman, J. W. "Genetics and the nature of selection on reproductive isolation in sticklebacks." *Integrative and Comparative Biology* 49 (2009): E18-E18.

Brawand, D., C. E. Wagner, Y. I. Li, M. Malinsky, I. Keller, S. H. Fan, O. Simakov, A. Y. Ng, Z. W. Lim, E. Bezault, J. Turner-Maier, J. Johnson, R. Alcazar, H. J. Noh, P.

Russell, B. Aken, J. Alfoldi, C. Amemiya, N. Azzouzi, J. F. Baroiller, F. Barloy-Hubler, A. Berlin, R. Bloomquist, K. L. Carleton, M. A. Conte, H. D'Cotta, O. Eshel, L. Gaffney, F. Galibert, H. F. Gante, S. Gnerre, L. Greuter, R. Guyon, N. S. Haddad, W. Haerty, R. M. Harris, H. A. Hofmann, T. Hourlier, G. Hulata, D. B. Jaffe, M. Lara, A. P. Lee, I. MacCallum, S. Mwaiko, M. Nikaido, H. Nishihara, C. Ozouf-Costaz, D. J. Penman, D. Przybylski, M. Rakotomanga, S. C. P. Renn, F. J. Ribeiro, M. Ron, W. Salzburger, L. Sanchez-Pulido, M. E. Santos, S. Searle, T. Sharpe, R. Swofford, F. J. Tan, L. Williams, S. Young, S. Y. Yin, N. Okada, T. D. Kocher, E. A. Miska, E. S. Lander, B. Venkatesh, R. D. Fernald, A. Meyer, C. P. Ponting, J. T. Streelman, K. Lindblad-Toh, O. Seehausen and F. Di Palma. "The genomic substrate for adaptive radiation in African cichlid fish." *Nature* 513 (2014): 375-381.

Brodie, E. D. "Correlational selection for color pattern and antipredator behavior in the garter snake *Thamnophis ordinoides*." *Evolution* 46 (1992): 1284-1298.

Burke, M. K. and A. D. Long. "What paths do advantageous alleles take during short-term evolutionary change?" *Molecular Ecology* 21 (2012): 4913-4916.

Burri, R., A. Nater, T. Kawakami, C. F. Mugal, P. I. Olason, L. Smeds, A. Suh, L. Dutoit, S. Bures, L. Z. Garamszegi, S. Hogner, J. Moreno, A. Qvarnstrom, M. Ruzic, S. A. Saether, G. P. Saetre, J. Torok and H. Ellegren. "Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers." *Genome Research* 25 (2015): 1656-1665.

Catchen, J., S. Bassham, T. Wilson, M. Currey, C. O'Brien, Q. Yeates and W. A. Cresko. "The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing." *Molecular Ecology* 22 (2013): 2864-2883.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores and W. A. Cresko. "Stacks: An analysis tool set for population genomics." *Molecular Ecology* 22 (2013): 3124-3140.

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

Charlesworth, B. "Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation." *Nature Reviews Genetics* 10 (2009): 195-205.

Charlesworth, B., D. Charlesworth and N. H. Barton. "The effects of genetic and geographic structure on neutral variation." *Annual Review of Ecology Evolution and Systematics* 34 (2003): 99-125.

Charlesworth, B., M. Nordborg and D. Charlesworth. "The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations." *Genetics Research* 70 (1997): 155-174.

Colosimo, P. F., K. E. Hosemann, S. Balabhadra, G. Villarreal, M. Dickson, J. Grimwood, J. Schmutz, R. M. Myers, D. Schluter and D. M. Kingsley. "Widespread parallel evolution in sticklebacks by repeated fixation of *Ectodysplasin* alleles." *Science* 307 (2005): 1928-1933.

Comeron, J. M., R. Ratnappan and S. Bailin. "The Many Landscapes of Recombination in *Drosophila melanogaster*." *PLoS Genetics* 8 (2012): e1002905.

Coop, G., X. Q. Wen, C. Ober, J. K. Pritchard and M. Przeworski. "High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans." *Science* 319 (2008): 1395-1398.

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

Coyne, J. A. and H. A. Orr. *Speciation*. Sunderland, MA: Sinauer Associates, 2004.

Cresko, W. A., A. Amores, C. Wilson, J. Murphy, M. Currey, P. Phillips, M. A. Bell, C. B. Kimmel and J. H. Postlethwait. "Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations." *Proceedings of the National Academy of Sciences of the United States of America* 101 (2004): 6050-6055.

Crow, J. F. “Mid-century controversies in population genetics.” *Annual Review of Genetics* 42 (2008): 1-16.

Cui, R., M. Schumer, K. Kruesi, R. Walter, P. Andolfatto and G. G. Rosenthal. "Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes." *Evolution* 67 (2013): 2166-2179.

Darwin, C. *The Origin of Species By Means of Natural Selection; Or, The Preservation of Favored Races in the Struggle For Life*. New York, NY: The Modern library, 1859.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen and M. L. Blaxter. "Genome-wide genetic marker discovery and genotyping using next-generation sequencing." *Nature Reviews Genetics* 12 (2011): 499-510.

Dobzhansky, T. and A. H. Sturtevant. “Inversions in the chromosomes of *Drosophila pseudoobscura.*” *Genetics* 23 (1938): 28-64.

Domingues, V. S., Y. P. Poh, B. K. Peterson, P. S. Pennings, J. D. Jensen and H. E. Hoekstra. "Evidence of adaptation from ancestral variation in young populations of beach mice." *Evolution* 66 (2012): 3209-3223.

Drummond, A. J. and A. Rambaut. "BEAST: Bayesian evolutionary analysis by sampling trees." *BMC evolutionary biology* 7 (2007): 214.

Drummond, A. J., M. A. Suchard, D. Xie and A. Rambaut. "Bayesian phylogenetics with BEAUti and the BEAST 1.7." *Molecular Biology and Evolution* 29 (2012): 1969-1973.

Dumont, B. L., K. W. Broman and B. A. Payseur. "Variation in genomic recombination rates among heterogeneous stock mice." *Genetics* 182 (2009): 1345-1349.

Durrett, R., L. Buttel and R. Harrison. "Spatial models for hybrid zones." *Heredity* 84 (2000): 9-19.

Endler, J. A. *Geographic Variation, Speciation, and Clines*. Princeton, NJ: Princeton University Press, 1977.

Endler, J. A. *Natural Selection in the Wild*. Princeton, NJ: Princeton University Press, 1986.

Ewing, E. P. "Genetic variation in a heterogeneous environment .7. Temporal and spatial heterogeneity in infinite populations." *American Naturalist* 114 (1979): 197-212.

Feder, J. L., S. P. Egan and P. Nosil. "The genomics of speciation-with-gene-flow." *Trends in Genetics* 28 (2012): 342-350.

Feder, J. L., R. Gejji, S. Yeaman and P. Nosil. "Establishment of new mutations under divergence and genome hitchhiking." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (2012): 461-474.

Feder, J. L. and P. Nosil. "The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation." *Evolution* 64 (2010): 1729-1747.

Felsenstein, J. "Theoretical population genetics of variable selection and migration." *Annual Review of Genetics* 10 (1976): 253-280.

Feulner, P. G. D., F. J. J. Chain, M. Panchal, C. Eizaguirre, M. Kalbe, T. L. Lenz, M. Mundry, I. E. Samonte, M. Stoll, M. Milinski, T. B. H. Reusch and E. Bornberg-

Bauer. "Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks." *Molecular Ecology* 22 (2013): 635-649.

Fisher, R. A.. "The correlation between relatives on the supposition of Mendelian inheritance." *Philosophical Transactions of the Royal Society of Edinburgh* 52 (1918): 399-433.

Fisher, R. A. *The Genetical Theory of Natural Selection*. Oxford: Clarendon, 1930.

Flaxman, S. M., A. C. Wacholder, J. L. Feder and P. Nosil. "Theoretical models of the influence of genomic architecture on the dynamics of speciation." *Molecular Ecology* 23 (2014): 4074-4088.

Fontaine, M. C., J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey, I. V. Sharakhov, X. F. Jiang, A. B. Hall, F. Catteruccia, E. Kakani, S. N. Mitchell, Y. C. Wu, H. A. Smith, R. R. Love, M. K. Lawniczak, M. A. Slotman, S. J. Emrich, M. W. Hahn and N. J. Besansky. "Extensive introgression in a malaria vector species complex revealed by phylogenomics." *Science* 347 (2015): 1258524.

Francis, R. C., J. V. Baumgartner, A. C. Havens and M. A. Bell. "Historical and ecological sources of variation among lake populations of threespine sticklebacks, *Gasterosteus aculeatus*, near Cook Inlet, Alaska." *Canadian Journal of Zoology* 64 (1986): 2257-2265.

Gillespie, J. H. (1973). "Polymorphism in random environments." *Theoretical Population Biology* 4 (1973): 193-195.

Gillespie, J. H. *Population genetics: a concise guide*. Baltimore, MD: Johns Hopkins University Press, 2004.

Glazer, A. M., E. E. Killingbeck, T. Mitros, D. S. Rokhsar and C. T. Miller. "Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing." *G3: Genes Genomes Genetics* 5 (2015): 1463-1472.

Grant, P. R. and B. R. Grant. "Unpredictable evolution in a 30-year study of Darwin's finches." *Science* 296 (2002): 707-711.

Guerrero, R. F., F. Rousset and M. Kirkpatrick. "Coalescent patterns for chromosomal inversions in divergent populations." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (2012): 430-438.

Haasl, R. J. and B. A. Payseur. "Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication." *Molecular Ecology* 25 (2016): 5-23.

Hahn, M. W. "Toward a selection theory of molecular evolution." *Evolution* 62 (2008): 255-265.

Hahn, M. W. and L. Nakhleh. "Irrational exuberance for resolved species trees." *Evolution* 70 (2008): 7-17.

Harr, B. "Genomic islands of differentiation between house mouse subspecies." *Genome Research* 16 (2006): 730-737.

Hedrick, P. W. "Genetic polymorphism in heterogeneous environments: The age of genomics." *Annual Review of Ecology Evolution and Systematics* 37 (2006): 67-93.

Hendry, A. P. and E. B. Taylor. "How much of the variation in adaptive divergence can be explained by gene flow? An evaluation using lake-stream stickleback pairs." *Evolution* 58 (2004): 2319-2331.

Hendry, A. P., E. B. Taylor and J. D. McPhail. "Adaptive divergence and the balance between selection and gene flow: lake and stream stickleback in the Misty system." *Evolution* 56 (2002): 1199-1216.

Hermisson, J. and P. S. Pennings. "Soft sweeps: Molecular population genetics of adaptation from standing genetic variation." *Genetics* 169 (2005): 2335-2352.

Hill, W. G. and A. Robertson. "Effect of linkage on limits to artificial selection." *Genetics Research* 8 (1966): 269-294.

Hohenlohe, P. A., S. Bassham, M. Currey and W. A. Cresko. "Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (2012): 395-408.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson and W. A. Cresko. "Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags." *PLoS Genetics* 6 (2010): e1000862.

Hubby, J. L. and R. C. Lewontin. "A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*." *Genetics* 54 (1966): 577-594.

Hudson, R. R., M. Slatkin and W. P. Maddison. "Estimation of levels of gene flow from DNA sequence data." *Genetics* 132 (1992): 583-589.

Huerta-Sánchez, E., X. Jin, Asan, Z. Bianba, B. M. Peter, N. Vinckenbosch, Y. Liang, X. Yi, M. He, M. Somel, P. Ni, B. Wang, X. Ou, Huasang, J. Luosang, Z. X. P. Cuo, K. Li, G. Gao, Y. Yin, W. Wang, X. Zhang, X. Xu, H. Yang, Y. Li, J. Wang, J. Wang and

R. Nielsen. "Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA." *Nature* 512 (2014): 194-197.

Ingram, T., R. Svanbäck, N. J. B. Kraft, P. Kratina, L. Southcott, D. Schluter and R. Svanb. "Intraguild predation drives evolutionary nich shift in threespine stickleback." *Evolution* 66 (2012): 1819-1832.

Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. T. Lu, K. M. Roskin, C. F. Chen, M. A. Thomas, D. Haussler and H. J. Jacob. "Comparative recombination rates in the rat, mouse, and human genomes." *Genome Research* 14 (2004): 528-538.

Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody, S. White, E. Birney, S. Searle, J. Schmutz, J. Grimwood, M. C. Dickson, R. M. Myers, C. T. Miller, B. R. Summers, A. K. Knecht, S. D. Brady, H. Zhang, A. A. Pollen, T. Howes, C. Amemiya, P. Broad Institute Genome Sequencing, T. Whole Genome Assembly, J. Baldwin, T. Bloom, D. B. Jaffe, R. Nicol, J. Wilkinson, E. S. Lander, F. Di Palma, K. Lindblad-Toh and D. M. Kingsley. "The genomic basis of adaptive evolution in threespine sticklebacks." *Nature* 484 (2012): 55-61.

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

Kane, N. C., M. G. King, M. S. Barker, A. Raduski, S. Karrenberg, Y. Yatabe, S. J. Knapp and L. H. Rieseberg. "Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species." *Evolution* 63 (2009): 2061-2075.

Karasov, T., P. W. Messer and D. A. Petrov. "Evidence that adaptation in *Drosophila* is not limited by mutation at single sites." *PLoS Genetics* 6 (2010): e1000924.

Kimura, M. *The Neutral Theory of Molecular Evolution*. New York: Cambridge University Press, 1983.

Kingman, J. F. C. "The coalescent." *Stochastic Processes and their Applications* 13 (1982): 235-248.

Kingman, J. F. C. "On the genealogy of large populations." *Journal of Applied Probability* 19 (1982): 27-43.

Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gibert and P. Beerli. "The strength of phenotypic selection in natural populations." *American Naturalist* 157 (2001): 245-261.

Kirkpatrick, M. "How and why chromosome inversions evolve." *PLoS Biology* 8 (2010): e1000501.

Kirkpatrick, M. and N. Barton. "Chromosome inversions, local adaptation and speciation." *Genetics* 173 (2006): 419-434.

Kong, A., G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G. B. Walters, A. Jonasdottir, A. Gylfason, K. T. Kristinsson, S. A. Gudjonsson, M. L. Frigge, A. Helgason, U. Thorsteinsdottir and K. Stefansson. "Fine-scale recombination rate differences between sexes, populations and individuals." *Nature* 467 (2010): 1099-1103.

Kreitman, M. "Nucleotide polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster*." *Nature* 304 (1983): 412-417.

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

Lackey, A. C. R. and J. W. Boughman. "Evolution of reproductive isolation in stickleback fish." *Evolution* 71 (2017): 357-372.

Langley, C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider, J. E. Pool, S. A. Langley, C. Suarez, R. B. Corbett-Detig, B. Kolaczkowski, S. Fang, P. M. Nista, A. K. Holloway, A. D. Kern, C. N. Dewey, Y. S. Song, M. W. Hahn and D. J. Begun. "Genomic variation in natural populations of *Drosophila melanogaster*." *Genetics* 192 (2012): 533-598.

Leaver, S. D. and T. E. Reimchen. "Abrupt changes in defence and trophic morphology of the giant threespine stickleback (*Gasterosteus sp.*) following colonization of a vacant habitat." *Biological Journal of the Linnean Society* 107 (2012): 494-509.

Lee, C.-R., B. Wang, J. P. Mojica, T. Mandáková, K. V. S. K. Prasad, J. L. Goicoechea, N. Perera, U. Hellsten, H. N. Hundley, J. Johnson, J. Grimwood, K. Barry, S. Fairclough, J. W. Jenkins, Y. Yu, D. Kudrna, J. Zhang, J. Talag, W. Golser, K. Ghattas, M. E. Schranz, R. Wing, M. A. Lysak, J. Schmutz, D. S. Rokhsar and T. Mitchell-Olds. "Young inversion with multiple linked QTLs under selection in a hybrid zone." *Nature Ecology & Evolution* 1 (2017): 0119.

Lee, Y. W., L. Fishman, J. K. Kelly and J. H. Willis. "A segregating inversion generates fitness variation in yellow monkeyflower (*Mimulus guttatus*)." *Genetics* 202 (2016): 1473-1484.

Lekberg, Y., B. Roskilly, M. F. Hendrick, C. A. Zabinski, C. M. Barr and L. Fishman. "Phenotypic and genetic differentiation among yellow monkeyflower populations from thermal and non-thermal soils in Yellowstone National Park." *Oecologia* 170 (2012): 111-122.

Lenormand, T. "Gene flow and the limits to natural selection." *Trends in Ecology & Evolution* 17 (2002): 183-189.

Lenormand, T., D. Bourguet, T. Guillemaud and M. Raymond. "Tracking the evolution of insecticide resistance in the mosquito *Culex pipiens*." *Nature* 400 (1999): 861-864.

Lescak, E. A., S. L. Bassham, J. Catchen, O. Gelmond, M. L. Sherbick, F. A. von Hippel and W. A. Cresko. "Evolution of stickleback in 50 years on earthquake-uplifted islands." *Proceedings of the National Academy of Sciences of the United States of America* 112 (2015): E7204–E7212.

Leslie, S., B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, K. Hutnik, E. C. Royrvik, B. Cunliffe, D. J. Lawson, D. Falush, C. Freeman, M. Pirinen, S. Myers, M. Robinson, P. Donnelly, W. Bodmer, Wellcome Trust Case Control Consortium and the International Multiple Sclerosis Genetics Consortium. "The fine-scale genetic structure of the British population." *Nature* 519 (2015): 309-314.

Levin, D. A. "Metapopulations - an arena for local speciation." *Journal of Evolutionary Biology* 8 (1995): 635-644.

Lewontin, R. C. *The Genetic Basis of Evolutionary Change*. New York: Columbia University Press, 1974.

Lewontin, R. C. and J. L. Hubby. "A molecular approach to study of genic heterozygosity in natural populations .II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*." *Genetics* 54 (1966): 595-609.

Linnen, C. R., E. P. Kingsley, J. D. Jensen and H. E. Hoekstra. "On the origin and spread of an adaptive allele in deer mice." *Science* (2009): 1095-1098.

Lowry, D. B., R. C. Rockwood and J. H. Willis. "Ecological reproductive isolation of coast and inland races of *Mimulus guttatus*." *Evolution* 62 (2008): 2196-2214.

Lowry, D. B. and J. H. Willis. "A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation." *PLoS Biology* 8 (2010): e1000500.

Luikart, G., P. R. England, D. Tallmon, S. Jordan and P. Taberlet. "The power and promise of population genomics: from genotyping to genome typing." *Nature Reviews Genetics* 4 (2003): 981-994.

Malinsky, M., R. J. Challis, A. M. Tyers, S. Schiffels, Y. Terai, B. P. Ngatunga, E. A. Miska, R. Durbin, M. J. Genner and G. F. Turner. "Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake." *Science* 350 (2015): 1493-1498.

Martin, S. H., M. Most, W. J. Palmer, C. Salazar, W. O. McMillan, F. M. Jiggins and C. D. Jiggins. "Natural selection and genetic diversity in the butterfly *Heliconius melpomene*." *Genetics* 203 (2016): 525-541.

Maynard Smith, J. and J. Haigh. "The hitch-hiking effect of a favourable gene." *Genetical Research* 23 (1974): 23-35.

McDonald, J. H. and M. Kreitman. "Adaptive protein evolution at the *Adh* locus in *Drosophila*." *Nature* 351 (1991): 652-654.

McKinnon, J. S. and H. D. Rundle. "Speciation in nature: the threespine stickleback model systems." *Trends in Ecology & Evolution* 17 (2002): 480-488.

Miller, C. T., A. M. Glazer, B. R. Summers, B. K. Blackman, A. R. Norman, M. D. Shapiro, B. L. Cole, C. L. Peichel, D. Schluter and D. M. Kingsley. "Modular skeletal evolution in sticklebacks is controlled by additive and clustered quantitative trait loci." *Genetics* 197 (2014): 405-420.

Miller, S. E., D. Metcalf and D. Schluter. "Intraguild predation leads to genetically based character shifts in the threespine stickleback." *Evolution* 69 (2015): 3194-3203.

Morgan, T. H., A. H. Sturtevant and C. B. Bridges. "The evidence for the linear order of the genes." *Proceedings of the National Academy of Sciences of the United States of America* 6 (1920): 162-164.

Nachman, M. W. "Variation in recombination rate across the genome: evidence and implications." *Current Opinion in Genetics & Development* 12 (2002): 657-663.

Nadeau, N. J., A. Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra, S. W. Baxter, M. A. Quail, M. Joron, R. H. Ffrench-Constant, M. L. Blaxter, J. Mallet and C. D. Jiggins. "Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing." *Philosophical Transactions of the Royal Society B-Biological Sciences* 367 (2012): 343-353.

Nei, M. *Molecular Evolutionary Genetics*. New York: Columbia University Press, 1987.

Nei, M. and W. H. Li. “Mathematical model for studying genetic variation in terms of restriction endonucleases." *Proceedings of the National Academy of Sciences of the United States of America* 76 (1979): 5269-5273.

Nosil, P. *Ecological Speciation*. Oxford: Oxford University Press, 2012.

Nosil, P. and J. L. Feder. "Genomic divergence during speciation: causes and consequences." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (2012): 332-342.

Nosil, P. and J. L. Feder. "Widespread yet heterogeneous genomic divergence." *Molecular Ecology* 21 (2012): 2829-2832.

Nosil, P., D. J. Funk and D. Ortiz-Barrientos. "Divergent selection and heterogeneous genomic divergence." *Molecular Ecology* 18 (2009): 375-402.

Orr, H. A. "The genetic theory of adaptation: a brief history." *Nature Reviews Genetics* 6: (2005): 119-127.

Ortiz-Barrientos, D., J. Engelstadter and L. H. Rieseberg. "Recombination rate evolution and the origin of species." *Trends in Ecology & Evolution* 31 (2016): 226–236.

Palumbi, S. R. "Evolution - Humans as the world's greatest evolutionary force." *Science* 293 (2001): 1786-1790.

Pannell, J. R. and B. Charlesworth. "Effects of metapopulation processes on measures of genetic diversity." *Philosophical Transactions of the Royal Society B-Biological Sciences* 355 (2000): 1851-1864.

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

Paradis, E., J. Claude and K. Strimmer. "APE: Analyses of phylogenetics and evolution in R language." *Bioinformatics* 20 (2004): 289-290.

Pease, J. B., D. C. Haak, M. W. Hahn and L. C. Moyle. "Phylogenomics reveals three sources of adaptive variation during a rapid radiation." *PLoS Biology* 14 (2016): e1002379.

Pfeifer, B., U. Wittelsburger, S. E. Ramos-Onsins and M. J. Lercher. "PopGenome: An efficient swiss army knife for population genomic analyses in R." *Molecular Biology and Evolution* 31 (2014): 1929-1936.

Phillips, P. C. "Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems." *Nature Reviews Genetics* 9 (2008): 855-867.

Poelstra, J. W., N. Vijay, C. M. Bossu, H. Lantz, B. Ryll, I. Muller, V. Baglione, P. Unneberg, M. Wikelski, M. G. Grabherr and J. B. W. Wolf. "The genomic landscape underlying phenotypic integrity in the face of gene flow in crows." *Science* 344 (2014): 1410-1414.

Prasad, K. V. S. K., B.-H. Song, C. Olson-Manning, J. T. Anderson, C.-R. Lee, M. E. Schranz, a. J. Windsor, M. J. Clauss, a. J. Manzaneda, I. Naqvi, M. Reichelt, J. Gershenzon, S. G. Rupasinghe, M. a. Schuler and T. Mitchell-Olds. "A gain-of-function polymorphism controlling complex traits and fitness in nature." *Science* 337 (2012): 1081-1084.

Pritchard, J. K., M. Stephens and P. Donnelly. "Inference of population structure using multilocus genotype data." *Genetics* 155 (2000): 945-959.

Prüfer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. F. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso and S. Pääbo. "The complete genome sequence of a Neanderthal from the Altai Mountains." *Nature* (2014): 43-49.

Ralph, P. L. and G. Coop. "Convergent evolution during local adaptation to patchy landscapes." *PLoS Genetics* 11 (2015): e1005630.

Rastas, P., F. C. F. Calboli, B. C. Guo, T. Shikano and J. Merila. "Construction of ultradense linkage maps with Lep-MAP2: Stickleback F_2 recombinant crosses as an example." *Genome Biology and Evolution* 8 (2016): 78-93.

Reger, R. D. and D. S. Pinney. “Late Wisconsin glaciation of the Cook Inlet region with emphasis on Kenai lowland and implications for early peopling.” In *Adventures through time: Readings in the Anthropology of Cook Inlet, Alaska*, ed. N. Y. Davis and W. E. Davis, 13-35. Anchorage, AK: Cook Inlet Historical Society, 1996.

Roda, F., H. L. Liu, M. J. Wilkinson, G. M. Walter, M. E. James, D. M. Bernal, M. C. Melo, A. Lowe, L. H. Rieseberg, P. Prentis and D. Ortiz-Barrientos. "Convergence and divergence during the adaptation to similar environments by an Australian groundsel." *Evolution* 67 (2013): 2515-2529.

Roesti, M., S. Gavrilets, A. P. Hendry, W. Salzburger and D. Berner. "The genomic signature of parallel adaptation from shared genetic variation." *Molecular Ecology* 23 (2014): 3944-3956.

Roesti, M., A. P. Hendry, W. Salzburger and D. Berner. "Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs." *Molecular Ecology* 21 (2012): 2852-2862.

Roesti, M., B. Kueng, D. Moser and D. Berner. "The genomics of ecological vicariance in threespine stickleback fish." *Nature communications* 6 (2015): 8767.

Roesti, M., D. Moser and D. Berner. "Recombination in the threespine stickleback genome - patterns and consequences." *Molecular Ecology* 22 (2013): 3014-3027.

Schluter, D. and G. L. Conte. "Genetics and ecological speciation." *Proceedings of the National Academy of Sciences of the United States of America* 106 (2009): 9955-9962.

Schluter, D., K. B. Marchinko, R. D. H. Barrett and S. M. Rogers. "Natural selection and the genetics of adaptation in threespine stickleback." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 365 (2010): 2479-2486.

Schultz, J. and H. Redfield. "Interchromosomal effects on crossing over in Drosophila." *Cold Spring Harbor Symposia on Quantitative Biology* 16 (1951): 175-197.

Shapiro, M. D., M. E. Marks, C. L. Peichel, B. K. Blackman, K. S. Nereng, B. Jónsson, D. Schluter and D. M. Kingsley. "Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks." *Nature* 428 (2004): 717-723.

Skoglund, P., S. Mallick, M. C. Bortolini, N. Chennagiri, T. Hünemeier, M. L. Petzl-Erler, F. M. Salzano, N. Patterson and D. Reich. "Genetic evidence for two founding populations of the Americas." *Nature* 525 (2015): 104-108.

Slatkin, M.. "Isolation by distance in equilibrium and non-equilibrium populations." *Evolution* 47 (1993): 264-279.

Smukowski, C. S. and M. A. F. Noor. "Recombination rate variation in closely related species." *Heredity* 107 (2011): 496-508.

Soria-Carrasco, V., Z. Gompert, A. A. Comeault, T. E. Farkas, T. L. Parchman, J. S. Johnston, C. A. Buerkle, J. L. Feder, J. Bast, T. Schwander, S. P. Egan, B. J. Crespi and P. Nosil. "Stick insect genomes reveal natural selection's role in parallel speciation." *Science* 344 (2014): 738-742.

Stankowski, S., J. M. Sobel and M. A. Streisfeld. "The geography of divergence-with-gene-flow facilitates multi-trait adaptation and the evolution of pollinator isolation in *Mimulus aurantiacus*." *Evolution* 69 (2015): 3054-3068.

Stankowski, S. and M. A. Streisfeld. "Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers." *Proceedings of the Royal Society B: Biological Sciences* 282 (2015): 20151666.

Stephens, M., N. J. Smith and P. Donnelly. "A new statistical method for haplotype reconstruction from population data." *American Journal of Human Genetics* 68 (2001): 978-989.

Tajima, F. "Evolutionary relationship of DNA sequences in finite populations." *Genetics* 105 (1983): 437-460.

R Core Team. R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing (2016).

Tenaillon, O., J. E. Barrick, N. Ribeck, D. E. Deatherage, J. L. Blanchard, A. Dasgupta, G. C. Wu, S. Wielgoss, S. Cruveiller, C. Medigue, D. Schneider and R. E. Lenski. "Tempo and mode of genome evolution in a 50,000-generation experiment." *Nature* 536 (2016): 165-170.

Toews, D. P. L., S. A. Taylor, R. Vallender, A. Brelsford, B. G. Butcher, P. W. Messer and I. J. Lovette. "Plumage genes and little else distinguish the genomes of hybridizing warblers." *Current Biology* 26 (2016): 2313-2318.

Turner, T. L., M. W. Hahn and S. V. Nuzhdin. "Genomic islands of speciation in *Anopheles gambiae*." *PLoS Biology* 3 (2005): 1572-1578.

Twyford, A. D. and J. Friedman. "Adaptive divergence in the monkey flower *Mimulus guttatus* is maintained by a chromosomal inversion." *Evolution* 69 (2015): 1476-1486.

Van Belleghem, S. M., P. Rastas, A. Papanicolaou, S. H. Martin, C. F. Arias, M. A. Supple, J. J. Hanly, J. Mallet, J. J. Lewis, H. M. Hines, M. Ruiz, C. Salazar, M. Linares, G. R. P. Moreira, C. D. Jiggins, B. A. Counterman, W. O. McMillan and R. Papa. "Complex modular architecture around a simple toolkit of wing pattern genes." *Nature Ecology & Evolution* 1 (2017): 0052.

Via, S.. "Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (2012): 451-460.

Vitti, J. J., S. R. Grossman and P. C. Sabeti. "Detecting natural selection in genomic data." *Annual Review of Genetics* 47 (2013): 97-120.

Wallbank, R. W. R., S. W. Baxter, C. Pardo-Diaz, J. J. Hanly, S. H. Martin, J. Mallet, K. K. Dasmahapatra, C. Salazar, M. Joron, N. Nadeau, W. O. McMillan and C. D.

Jiggins. "Evolutionary novelty in a butterfly wing pattern through enhancer shuffling." *PLoS Biology* 14 (2016): e1002353.

Wilson, B. A., D. A. Petrov and P. W. Messer. "Soft selective sweeps in complex demographic scenarios." *Genetics* 198 (2014): 669-684.

Wolf, J. B. W. and H. Ellegren. "Making sense of genomic islands of differentiation in light of speciation." *Nature Reviews Genetics* 18 (2016): 87-100.

Wright, S. "The roles of mutation, inbreeding, crossbreeding and selection in evolution." *Proceedings of the Sixth International Congress on Genetics* 1 (1932): 356-366.

Wu, T. D. and S. Nacu. "Fast and SNP-tolerant detection of complex variants and splicing in short reads." *Bioinformatics* 26 (2010): 873-881.

Yatabe, Y., N. C. Kane, C. Scotti-Saintagne and L. H. Rieseberg. "Rampant gene exchange across a strong reproductive barrier between the annual sunflowers, *Helianthus annuus* and *H. petiolaris*." *Genetics* 175 (2007): 1883-1893.

Yeaman, S.. "Local adaptation by alleles of small effect." *American Naturalist* 186 (2015): S74-S89.

Yeaman, S., K. A. Hodgins, K. E. Lotterhos, H. Suren, S. Nadeau, J. C. Degner, K. A. Nurkowski, P. Smets, T. L. Wang, L. K. Gray, K. J. Liepe, A. Hamann, J. A. Holliday, M. C. Whitlock, L. H. Rieseberg and S. N. Aitken. "Convergent local adaptation to climate in distantly related conifers." *Science* 353 (2016): 1431-1433.

Yeaman, S. and M. C. Whitlock. "The genetic architecture of adaptation under migration-selection balance." *Evolution* 65 (2011): 1897-1911.

Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, H. C. Zheng, T. Liu, W. M. He, K. Li, R. B. Luo, X. F. Nie, H. L. Wu, M. R. Zhao, H. Z. Cao, J. Zou, Y. Shan, S. Z. Li, Q. Yang, Asan, P. X. Ni, G. Tian, J. M. Xu, X. A. Liu, T. Jiang, R. H. Wu, G. Y. Zhou, M. F. Tang, J. J. Qin, T. Wang, S. J. Feng, G. H. Li, Huasang, J. B. Luosang, W. Wang, F. Chen, Y. D. Wang, X. G. Zheng, Z. Li, Z. M. Bianba, G. Yang, X. P. Wang, S. H. Tang, G. Y. Gao, Y. Chen, Z. Luo, L. Gusang, Z. Cao, Q. H. Zhang, W. H. Ouyang, X. L. Ren, H. Q. Liang, H. S. Zheng, Y. B. Huang, J. X. Li, L. Bolund, K. Kristiansen, Y. R. Li, Y. Zhang, X. Q. Zhang, R. Q. Li, S. G. Li, H. M. Yang, R. Nielsen, J. Wang and J. A. Wang. "Sequencing of 50 human exomes reveals adaptation to high altitude." *Science* 329 (2010): 75-78.