THE EVOLUTIONARY GENOMICS OF ADAPTATION AND SPECIATION IN THE THREESPINE STICKLEBACK

by

Kieran Mikhail Samuk

H.B.Sc., The University of Toronto, 2008

M.Sc., The University of British Columbia, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Zoology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

July 2016

Abstract

Speciation and adaptation are key processes in biological evolution. Speciation creates genealogically discrete lineages, whereas adaptation causes organisms to become better matched to their environments. In this thesis, I conducted three studies that advanced our knowledge of speciation or adaptation. All three studies made use of a unique study system: threespine stickleback – small fish found in marine and fresh waters throughout the northern hemisphere. I first explored the potential of a newly-discovered "white" form of threespine stickleback for studying the early phases of speciation. Using a variety of population genomic methods, I showed that white stickleback are genetically distinct from other marine stickleback, and diverged recently in the face of substantial gene flow. These features make white stickleback an excellent system for studying the early phases of speciation. Next, I used white stickleback to examine the role of sexual and trophic divergence in the early phases of speciation. Using morphological and isotopic data, I found evidence for only weak trophic differentiation between white and common stickleback. Instead, genetic differences between the two forms are concentrated on genomic regions that harbour genes with male-biased expression. This suggests that, apart from difference in body size, strong trophic differentiation may not be necessary in the early phases of speciation. The final study explored the role of gene flow in shaping the genomic architecture of adaptation. Theory predicts that when adaptation occurs in the face of gene flow, genomic architectures in which adaptive loci are localized in regions of low recombination will be favored over others. I tested this prediction by quantifying the correlation between recombination rate and the density of adaptive loci in pairs of stickleback populations that varied in their degree of gene flow. In line with theory, we found that adaptive loci were more like to be found in regions of low recombination when divergent selection and gene flow co-occurred. Together, the studies presented in this thesis provide new tools and significant advances in our understanding of speciation and adaptation.

Preface

The work presented in Chapter 2 and 3 of this thesis was conceived and designed by myself, in close consultation with my advisor Dolph Schluter. Hannah Visty (co-author of Chapter 3) and Jacob Best (research technician) collected the morphological data presented in Chapter 3. Kate Ostevik provided field assistance for the collection of specimens, as well as assistance with preparing genomic sequencing libraries for all chapters. I performed the analyses and wrote the resulting manuscripts presented in Chapter 2 and 3, again in close consultation with Dolph Schluter.

Chapter 4 was a collaboration between myself and Greg Owens, Diana Rennison, Sara Miller, Kira Delmore and Dolph Schluter. As a group, we designed the study and prepared the genomic data presented in that chapter for analysis. I then carried out the statistical analysis and wrote the paper with input from the other authors.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

cM	Centimorgan
DNA	Deoxyribonucleic Acid
DS-GF	Divergent selection with gene flow
GATK	Genome Analysis Toolkit
GBS	Genotyping by sequencing
IM	Isolation with Migration
JSFS	Joint site frequency spectrum
KB	Kilobase
LD	Linkage disequilibrium
(M)ANOVA	(Multivariate) Analysis of Variance
MB	Megabase
NND	Nearest-neighbour distance
PCA	Principal Components Analysis
QTL	Quantitative trait locus
RAD	Restriction amplified digest
RI	Reproductive isolation
SNP	Single nucleotide polymorphism
VCF	Variant Call Format

Acknowledgements

This thesis could not have been possible without the help and support of many people. First and foremost, my advisor Dolph Schluter provided an enormous amount of support and sage advice on all aspects of the work presented here. Working in Dolph's lab has been a transformational experience, and I feel very lucky to have had the opportunity to do so. Thanks for making me think about The Big Picture™, Dolph!

Secondly, I am greatly indebted to my supervisory committee: Patricia Schulte, Michael Whitlock and Loren Rieseberg. The feedback provided during committee meetings and via comments on manuscripts was instrumental in the success of this thesis.

This thesis was partly inspired by the pioneering work of Dr. Max Blouw. Dr. Blouw was the first to study the white stickleback, and many of the ideas discussed in this thesis are based on his work. Dr. Blouw also generously shared a great deal of original data, notes, and advice on studying the white stickleback.

The members of the Schluter lab, particularly Diana Rennison, Sara Miller, Gina Conte, Seth Rudman, Monica Yau, Carling Gerlinsky, Thor Veen were all key to my success. Outside the Schluter lab, the following people also helped me immensely in various ways: Greg Owens, Brook Moyers, Kira Delmore, Greg Baute, Sam Yeaman, Armando Geraldes, Chris Grassa, Jasmine Ono, Andrew MacDonald, Marius Roesti, Matt Siegle, JS Moore, Jon Mee, Aleeza Gerstein, Jacob Best, Dave Toews, Laura Southcott, Nathaniel Sharpe, Julie Lee-Yaw, Kim Gilbert, Michael Scott. I couldn't have asked for a better group of friends and colleagues! Finally, many thanks to my wonderful partner Kate Ostevik for all her support over the years, and to my family: Dad, Tristan, Heidi, and my late mother and grandfather for their love and encouragement.

To Mom and Dad

Chapter 1: Introduction

The natural world is host to an astounding diversity of organisms. A careful examination of this diversity reveals two broad patterns. First, biological diversity is not a continuum – organisms tend to represent discrete, identifiable units (Coyne & Orr 2004). Secondly, organisms appear to be exquisitely matched to their environments (Williams 1966). These patterns are the outcome of two core evolutionary processes: speciation and adaptation. Speciation is the evolution of reproductive isolation. Reproductive isolation refers to the inability of two populations of organisms to produce viable offspring – and thus remain genealogically discrete (Coyne & Orr 2004). Adaptation is the processes by which organisms become better matched to their environment by way of natural selection (Williams 1966). Together, these two processes explain much of the organization of biological diversity. As such, understanding these processes forms the core of evolutionary research.

1.1 Aims of this thesis

In this thesis, I present three studies that each advance our knowledge of either speciation or adaptation. Here, I outline the broad motivation behind each chapter along with a general description of my approach and findings. Each chapter of the thesis makes use of the threespine stickleback as a study system, which is briefly introduced below.

1.2 Threespine stickleback

Threespine stickleback (herein "stickleback") are small ray-finned fishes found in both marine and fresh water environments throughout the northern hemisphere (McKinnon & Rundle 2002). For several reasons, stickleback have become a very popular system for studying evolution. First, stickleback have rapidly diversified into a variety of unique incipient species over the last ~12,000 years (Hendry *et al.* 2009). Many of these species have independently evolved multiple times, making them an attractive system for studying the role of natural selection in speciation. Secondly, stickleback are easy to rear and observe in laboratory conditions, and crossing divergent forms can be readily achieved via in vitro fertilization (Taylor *et al.* 2011). Thirdly, stickleback have well-developed genomic resources, including a high quality reference genome (Jones *et al.* 2012). This allows the exploration of genome-scale patterns, including the genomic distribution of divergence between sister species. Finally, and most importantly for the thesis, many incipient stickleback species likely evolved in the presence of substantial gene flow, while others did not (McKinnon & Rundle 2002; Hendry *et al.* 2009; Marques *et al.* 2016). This creates the natural variation needed to the study the role of gene flow in shaping both speciation and adaptation.

1.3 Outline of the thesis

1.3.1 Chapter 2: A new system for studying recent speciation

Speciation research focuses on studying the evolution of barriers to reproduction (Coyne & Orr 2004; Price 2008). Speciation researcher often ask question such as: What types of reproductive

barriers are involved in the formation of new species? In what order do these barriers evolve? The basic approach to answering these questions is to identify traits that cause reproductive isolation between two species, and then perform detailed studies of their evolution (Via 2009; Marie Curie Speciation Network *et al.* 2012). However, this approach has a fundamental problem: it tells us about the traits that cause reproductive isolation *currently*, not those that caused reproductive isolation *initially* (Coyne & Orr 2004). This is a problem because as reproductive isolation between two species evolves, new barriers eventually mask older ones (Coyne & Orr 2004). This results in an inability to resolve the order in which barriers evolved and thus their relative importance during evolution of reproductive isolation. The greater divergence time is between species, the greater this problem becomes.

How can we avoid this issue? One idea is to focus our research effort on very young species – those in which reproductive isolation evolved very recently, or remains incomplete (Hendry *et al.* 2009). Such species have the lowest potential for masked barriers, and are thus key to the study of speciation. An added benefit of studying young species is that most still actively hybridize with their sister species. This creates elevated signatures of differentiation in the genomes of the diverging species that we can use to identify regions involved in reproductive isolation (Oleksyk *et al.* 2009). However, in spite of these obvious advantages, speciation research has generally focused on older, more diverged species (Coyne & Orr 2004; Price 2008). This may be partly a function of the difficulty of recognizing very young species, which may lack clear phenotypic differences. Nevertheless, more young species study systems are required if we are to crack the nut that is speciation.

In Chapter 2, I examined whether a recently discovered "white" marine stickleback from Nova Scotia may be a suitable system for studying recent speciation. I used population genetic

methods to examine three aspects that would make this system ideal: genetically distinct diverging lineages (i.e. they are indeed speciating), recent divergence and the presence of gene flow.

Previous work has shown that white stickleback are differentiated from the common form in body size, male coloration and nest site preference (Blouw & Hagen 1990). However, it is unclear whether white stickleback represent an incipient species, a genetically determined male-specific mating strategy polymorphism, a conditional strategy or an ontogenetic-determined male strategy. Further, the timing of divergence, and the question whether white and common stickleback are connected by gene flow, remains unclear. We used a reduced representation genome-wide sequencing approach -- "genotyping by sequencing", GBS -- to examine each of these aspects (Elshire *et al.* 2011). We found that white stickleback form a genotypic cluster distinct from common stickleback. These clusters align well with known phenotypic differences between whites and commons. The white cluster also contained both males and females and was stable across sampling years, suggesting that the white stickleback is an incipient species. To infer the timing of divergence and presence of gene flow, we compared models of divergence with and without gene flow using two methods: TREEMIX and dadi (Gutenkunst *et al.* 2009; Pickrell & Pritchard 2012). Both methods suggested that a model of very recent divergence with gene flow was the best fit to the data.

In sum, our results suggest that the white stickleback is a recently-evolved incipient species that likely diverged in the presence of gene flow. Thus, white stickleback are an excellent system for studying recent speciation.

1.3.2 Chapter 3: The roles of sexual and ecological divergence in speciation

A first step in studying the evolution of reproductive isolation is to identify the phenotypic and genotypic differences between sister species. Comparative data suggest that species often differ in two major trait axes: ecological -- related to performance in a particular niche, or sexual -- related to the alternative roles of the sexes (Schluter 2001; Ritchie 2007). However, the relative importance of these two classes of traits for the evolution of reproductive isolation is still poorly understood. In Chapter 3, I asked if white and common stickleback differed mainly in ecological or sexual traits. To quantify ecological differences, I measured (i) a suite of morphological traits known to be connected with ecological adaptation in stickleback, and (ii) ratios of Carbon-13 and Nitrogen-15 stable isotopes in the tissues of wild-caught individuals. I found that white stickleback do not differ from common stickleback in any of the traditional morphological traits associated with trophic niche in stickleback, other than in body size. I also found no significant differences in stable isotope abundances, suggesting weak differences in diet and trophic position. White stickleback appear to differ from common stickleback only in overall size and body colour.

To examine the extent of sexual differences between white and common stickleback, we asked whether regions of the genome that were highly divergent between the two forms contained more genes with sex-biased expression than the genomic background. To do this, we made use of a previously published dataset on sex-biased gene expression in threespine stickleback (Leder *et al.* 2015). We found that genes in divergent regions of the genome had ~20% greater male-biased expression on average than the genomic background. Together, these results suggest that white stickleback diverged from common stickleback in sexual rather than mainly ecological trait axes.

This suggests that ecological differentiation may not be an essential component of early reproductive isolation.

1.3.3 Chapter 4: Gene flow and the genomics of adaptation

Adaptation is the process by which natural selection increases the frequency of alleles that confer higher fitness in a given environment. Classically, it was thought that these alleles could generally occur at loci anywhere in the genome; all that matters is that they somehow increase fitness (Coyne & Orr 2004). However, recent theoretical work has shown that when adaptation occurs in the face of gene flow, haplotypes in which adaptive alleles are tightly linked can be favored over others (Kirkpatrick 2006; Yeaman & Whitlock 2011). For simplicity, I will refer to the arrangement and linkage between loci underlying a trait (fitness in this case) as 'genomic architecture'.

Models of the effects of gene flow on the genomic architecture of adaptation generally consider two populations diverging via natural selection with gene flow (Kirkpatrick 2006; Yeaman & Whitlock 2011). Natural selection brings alternate sets of selected alleles together in each population, i.e. establishes linkage disequilibrium (LD) between co-selected alleles. Gene flow between the populations and subsequent recombination breaks down LD between co-selected alleles. The key result of these models is that the breakdown of adaptive LD can be slowed if co-selected alleles are in tight genetic linkage – for example within a chromosomal inversion or other region of low recombination (Yeaman & Whitlock 2011). Thus, there is a general prediction that gene flow should specifically favor the evolution of adaptive alleles with tightly-linked genomic architectures. However, there have been very few empirical tests of these models, partly due to the need to contrast multiple populations with and without gene flow (Renaut *et al.* 2013).

To test these theoretical predictions, we compiled a global genomic dataset of over 1300 individual threespine stickleback from 48 populations and statistically disentangled the effects of gene density, mutation, gene flow and divergent natural selection on the extent to which adaptive alleles are clustered in regions of low recombination. After controlling for gene density and mutation rate, we found that genomic signatures of local adaptation tend to concentrate in regions of low recombination even in the absence of gene flow between pairs. However, in support of theory, this tendency is far stronger when divergent natural selection and gene flow co-occur. Together, these results suggest that gene flow constrains adaptation not only by opposing changes in allele frequency, but also by limiting *where* divergence can occur in the genome. If common, this has far-reaching implications for our understanding of the genetics of adaptation.

1.3.4 Summary

Understanding the forces that generate and maintain biological diversity is the central goal of evolutionary biology. Here, I outlined three studies that address key issues in the study of speciation and adaptation. Using a variety of methods, I helped to develop a new system for the study of speciation, explored the role of ecology and sex in speciation, and tested predictions about the genomic architecture of adaptation. My hope is that the research presented in this thesis not only helps answer unsolved questions in evolutionary research, but provides ideas for new lines of inquiry.

Chapter 2: Genome wide genotyping reveals that the white stickleback is an incipient species

2.1 Introduction

Speciation (the evolution of reproductive isolation, RI) is central to the generation and maintenance of biological diversity. Accordingly, understanding the genetics of speciation is a major goal of evolutionary biology (Coyne & Orr 2004). While recent advances have provided insight into some aspects of speciation genetics (Noor & Feder 2006; Seehausen *et al.* 2014), there are still many unanswered questions about how isolation evolves in natural systems. What forms of selection (if any) cause speciation genes to evolve? What role do changes in genome architecture play in the process? Do the traits that cause RI tend to have particular genetic architectures?

Answering these questions requires appropriate model study systems. In recent years, it has become increasingly appreciated that the most fruitful systems for this task are those that are early in the speciation process (Coyne & Orr 2004). Young species are particularly useful for studying speciation because they avoid a key problem with the study of the evolution of reproductive isolation: the compounding of reproductive barriers (Price 2008; Via 2009). As new species proceed along the "speciation continuum", new reproductive barriers gradually evolve and can mask those that formed at the onset of speciation (Nosil & Feder 2011). These later forming barriers may help maintain RI (e.g. by causing post-zygotic RI), but they are not informative of the barriers that originally caused speciation to occur (Orr 2005). This masking effect is an issue for the identification of barriers at the phenotypic and genotypic scale. For example, at the phenotypic level, a late-

evolved lethal intrinsic incompatibility between two species prevents us from assessing ecological hybrid performance as a reproductive barrier because hybrids are never formed (Butlin *et al.* 2014). Such a barrier would also heavily attenuate gene flow and cause gene-wide divergence to increase, reducing the power of divergence-based methods for detecting loci involved in RI (Noor & Feder 2006; Egan *et al.* 2008). Thus, we can maximize our ability to find the genetic changes that initiate speciation by focusing our attention on the most recently diverged taxa.

Young species that actively exchange migrants with each other are also very useful for speciation studies (Feder *et al.* 2012). This is because gene flow tends to homogenize parts of the genome that are not involved in the maintenance of species differences, amplifying the genomic signature of divergence at RI loci (Rogers & Bernatchez 2006; Pavlidis *et al.* 2012). This requires RI between the species in question to be incomplete, and so augments the benefits of studying recently diverged taxa. Interestingly, in spite of these obvious advantages, there are still only a handful of developed systems for studying the very earliest stages of speciation, particularly in the presence of gene flow. Some examples of such systems include *Rhagoletus* apple/hawthorn flies, *Littorina* intertidal snails, dune sunflowers and *Timema* walking sticks, which have all begun to yield key insights into the speciation process (Feder *et al.* 2003; Nosil *et al.* 2005; Andrew *et al.* 2012). However, there is obviously a great need for additional developed systems, particularly those with developed genomic resources.

Threespine stickleback (*Gasterosteus aculeatus*) are thought to harbor many such systems. However, most of the described "species pairs" actually have moderate to high levels of genome wide differentiation (Hohenlohe *et al.* 2010; Roesti *et al.* 2012; Jones *et al.* 2012), and the handful of recently diverged pairs are all of the stream-lake type (Roesti *et al.* 2012; Marques *et al.* 2016). Thus, we have an incomplete sample of the speciation continuum, and are limited in our ability to truly

probe the crucial changes that occur at the onset of speciation. To amend this, we need a stickleback species that is both very young, and still exchanges migrants with its sister group. Could such an experimental system exist?

2.1.1 The white stickleback

The so-called "white" threespine stickleback from Nova Scotia, Canada, may be one such species (Blouw & Hagen 1990). White stickleback appear to be a distinct form of marine stickleback (*Gasterosteus aculeateus*, hereafter "common stickleback"), and both types are broadly sympatric in marine environments in Nova Scotia (Appendix Figure A.1). Male white stickleback build nests near shore (sometimes in the intertidal), and are often found using filamentous algae rather than sand/gravel as nesting substrate (Jamieson *et al.* 1992b; Macdonald *et al.* 1995). When on the breeding grounds, male white stickleback exhibit unusual pearlescent-white dorsal breeding colors, instead of the more common olive/blue colors (Blouw & Hagen 1990). Intriguingly, male white stickleback also appear to lack the classic paternal care behaviors characteristic of male common stickleback: instead of caring for eggs after fertilizing them, white males carry them away from their nest (often out of their territory entirely), disperse the eggs into the surrounding algae, and return to soliciting matings from females (Jamieson *et al.* 1992a; Blouw & Blouw 1996). Male white stickleback are also on average ~20% shorter in body length than common male stickleback, resulting in a bimodal distribution of male body sizes at sites where both are found (Blouw & Hagen 1990).

In spite of these striking phenotypic differences, whether the white stickleback represents an incipient species remains unclear. Two lines of evidence suggest that reproductive isolation between white and common stickleback is likely very weak, if it exists at all. First, white and common

stickleback are fully interfertile in advanced generation laboratory crosses (Blouw 1996). Second, an allozyme study found no evidence of genetic differentiation between the two types (Haglund *et al.* 1990). On the other hand, some evidence suggests that white stickleback may be a nascent species. For one, a distinct class of small-bodied females are always found at sites with small-bodied male white stickleback (Blouw & Hagen 1990; Jamieson *et al.* 1992b). Mate-choice experiments and field observations also suggest that small females and small white males mate assortatively, consistent with experimental evidence of the role of body size in mate choice in stickleback (Jamieson *et al.* 1992a; b).

Together, the work by Blouw and colleagues suggests that white stickleback may be a promising system for studying recent speciation. However, assessing the stage of speciation (recent or old), as well as the presence of on-going gene flow requires a detailed genetic study of the evolutionary relationship between the white and common stickleback.

2.1.2 Approach and hypotheses

Here, I use the genomic resources recently developed for threespine stickleback to explore the evolutionary relationship between white and common stickleback. I attempt to answer the following two broad questions: (1) are white stickleback an incipient species? and (2) do white stickleback actively exchange migrants with common stickleback? I address these questions with analyses of a rich genome-wide polymorphism dataset derived from collections of both species I made in Nova Scotia.

2.1.2.1 Incipient species or alternative strategy?

While white stickleback may be an incipient species, Blouw and colleagues suggested a number of alternative explanations (Blouw & Hagen 1990). These are based on the various forms of male mating strategy polymorphisms – the situation in which males in a population exhibit alternative, discrete mating strategies such as "sneaker" or "guarder" (Gross 1996, Taborsky et al. 2008). The first such possibility is that male white common male phenotypes represent a genetically-based male mating strategy polymorphism, such as that in ruffs (Gross 1996, Brockmann 2001, Kupper et al. 2015). Secondly, white stickleback could be an ontogenetically determined strategy. For example, given their small body size, white stickleback may represent young males (perhaps first year breeders). Finally, it is possible that white sticklebacks are a condition dependent strategy – for example, low body condition males may become whites (Gross 1996).

If white stickleback are indeed an incipient species, there are a number of key predictions that can be tested. First, white and common males should from discrete genetic classes (e.g. clusters in a PCA of genetic variation). Second, females collected at sites with males should fall into the same two genetic classes as males (because they are part of the same lineage). Third, these classes should be stable over geographic locations and through time. Finally, the genetic differences between white and common types that underlie the discrete classes should be found on multiple chromosomes, rather than restricted to a single region -- as has been found in all fully described genetically determined male strategy polymorphisms (Gross 1996, Taborsky 2008, Kupper et al. 2015). If these predictions are met, then we reject the alternative explanations based on alternative male strategies.

2.1.2.2 Divergence with gene flow

Along with examining the possibility that white sticklebacks are an incipient species, I also test the hypothesis that white stickleback diverged and are diverging in the face of gene flow, rather than in the absence of gene flow. To do this, I fit isolation with migration (IM) models and TREEMIX migration-edge models to the polymorphism dataset. If there has been appreciable gene flow between white and common stickleback (now or in the past), the best fit model of both types should include significantly non-zero migration terms/edges.

2.2 Methods

2.2.1 Sample collection

In early May-July of 2012 and 2014, I collected white and common threespine stickleback at 12 sites in Nova Scotia, Canada (Figure 2.1 C). I determined sites using the list of sites in Blouw et al. (1992) as a guide. I focused on sites where both types were most likely to co-occur according to Blouw's environmental analysis: brackish water with abundant filamentous algae. This sampling scheme ultimately resulted in examining every accessible freshwater estuary I could access by car or short hike along the southern coast of Nova Scotia, from Yarmouth and through the Straight of Canso to Antigonish. In 2014, I also sampled estuarine sites I could access in the Bras d'Or Lake (and inland sea on Cape Breton Island) with a radius of approximately 100km centred on the town of Whycocomagh (Figure 2.1 D).

At all sites, I caught fish by setting unbaited Gee brand ¼ inch mesh stainless steel minnow traps in shallow regions where I observed males courting females. These were set according to the general methods described in Schluter & McPhail (1992). Upon retrieving the traps, I evenly sampled white and common males (identified by breeding color) and kept all females (identified by gravidity), until I had approximately 16 of each type of male and 32 unclassified females from each site. If I could not sex an individual by color or gravidity, or if a male had faded breeding colors, I did not collect it. All fish were euthanized using 0.5g/L tricaine methanesulfonate (MS-222) in sea water.

I placed all the individuals from each site into a single 1 L Nalgene container containing 95% ethanol, and moved each fish to an individual 50 mL Falcon tube containing 95% ethanol as soon as possible (usually ~6 hours later). Upon returning from the field (1-6 weeks after collections), I removed the pectoral and tail fins of each individual and placed them in 1.5mL microcentrifuge tubes filled with 95% ethanol.

2.2.2 Preparation of genotyping by sequencing libraries

I extracted DNA from the clipped fins of each individual using the protocol described in Peichel *et al.* (2001). Briefly, the tails were digested with proteinase-K and I used a standard phenol-chloroform extraction to isolate DNA. I eluted the resultant DNA in 1X TE and assessed DNA concentrations using a QuBit flourometer (Qiagen Corp, Germany). After DNA quality control, I retained DNA from 365 individuals.

I then prepared three genotyping-by-sequencing (GBS) libraries using an adapted version of the original protocol (Elshire *et al.* 2011). The first library contained DNA from 96 males from 2012,

randomized in 96-plate well position. Based on the number of sites obtained from the 2012 data, we increased the number of individuals to 148 for the second and third libraries. These two libraries contained DNA from 296 males and females from 2014, randomized among library, plate and in 96-plate well position. I aimed for an insert size of 300-400 basepairs, and used a gel-extraction method to size-select fragments from the prepared libraries. I confirmed the final fragment size distribution using a microeletrophoretic Bioanalyzer assay (Agilent Technologies, California). The completed libraries were then sequenced in individual lanes of an Illumina Hi-Seq 2000 at the University of British Columbia Biodiversity Next Gen Sequencing facility.

2.2.3 Variant Identification

I identified variants using a custom pipeline based on the GATK best practices guidelines (Supplement 1) (McKenna *et al.* 2010; DePristo *et al.* 2011). After demultiplexing the data using a Perl script, I used Trimmomatic version 0.32 (Bolger *et al.* 2014) to trim and filter sequences for quality. I then aligned the filtered reads to the revised stickleback reference genome (Glazer *et al.* 2015) using BWA version 0.7.10 "mem" algorithm (Li & Durbin 2010). I then realigned these reads using the GATK version 3.3.0 RealignTargetCreator, and IndelRealigner. Finally, I identified variants using the HaplotypeCaller, and genotyped the entire dataset using GenotypeGVCFs. To facilitate analyses that required an outgroup (e.g. TREEMIX) I also identified variants from whole genome data from six marine individuals from Denmark (Ferchaud *et al.* 2014). I processed these using the same pipeline, but with a separate run of GenotypeGVCFs.

I combined the final VCFs from the Nova Scotia and Denmark samples using the "merge" function in bcftools (Li 2011). I encoded the resultant single nucleotide polymorphism data in VCF

or tabular format, depending on the analysis. To simplify analyses, I only included sites with biallelic single nucleotide polymorphisms (SNPs). For all analyses, I also required sites to have genotype calls in at least 80% of all individuals.

I filtered this final dataset using slightly different criteria for each analysis (described below). These filters were meant to reduce bias and/or facilitate specific statistical analyses (e.g. by reducing interdependence in the data). Unless otherwise stated, I excluded SNPs with a minor allele frequency (MAF) of < 0.05. For some analyses, I "pruned" the dataset to reduce linkage disequilibrium (LD) between sites. This was done using the "--indep-pairwise 50 5 0.2" function in PLINK (Purcell *et al.* 2007). This function calculates pairwise linkage disequilbrium (r^2) between all SNPs in a window of 50 SNPs, which is moved along the genome at 5 SNP increments. If any SNPs in the window exceed the LD threshold (0.2), a single SNP is randomly chosen to be representative of SNPs in that window and the others are dropped. This reduces the statistical interdependence between SNPs caused by physical linkage, which is undesirable for most types of phylogenetic and demographic inference (Gutenkunst *et al.* 2009; Pickrell & Pritchard 2012). The final dataset ranged from ~55 000 – 19 000 SNPs in 354 individuals, depending on the filtration applied and the populations included.

Threespine stickleback have chromosomal sex determination, with males as the heterogametic sex (coded as XY, as in humans). The male sex chromosome also shares a small pseudoautosomal region with the X chromosome. However, the male sex chromosome is currently not included in the stickleback reference genome. This complicates allele-frequency based inference, because representation biases (e.g. more females sampled from a population) can have large effects on frequency estimates. Thus, for the analyses presented here, we removed all markers associated with the X chromosome.

2.2.4 Genotypic clustering

To test the hypotheses that white stickleback are genetically distinct from commons I used a variety of methods for detecting structure, i.e. increased genetic similarity among discrete groups, in the polymorphism dataset.

2.2.4.1 PCA

To assess whether white stickleback represent a distinct genotypic cluster, I first ordinated the pruned SNP data using principal components analysis (PCA). I did this using the R package *snpRelate* (Zheng *et al.* 2012). Principal components analysis attempts to find orthogonal vectors ("principal components") that explain the maximal variance in the raw data (in this case, the number of non-reference alleles at each biallelic SNP). These principal components represent composite indices of variation at many loci, i.e. a single component could represent allelic variation at hundreds or thousands of SNPs, each with a different scaled contribution (its "loading"). By projecting the original data onto these components, we can get a summarized version of the maximal sources of variation in the dataset.

To assess the presence and statistical significance of clusters in the SNP data, I first extracted the PC1 and PC2 scores for each individual, and identified clusters using the K-means algorithm (Jain & Dubes 1988). To assess the statistical significance of these clusters, I performed an analysis of cluster stability using the function *clusterboot* in the R package *fpc* (Hennig 2013). This method use assesses the significance of clusters in data using three methods: bootstrapping, addition of random outliers, and addition of random noise to the data, each followed by a re-fit of the K-means

algorithm. The program performs each perturbation/refit many times, and calculates the proportion of runs in which clusters are recovered with the same membership as those in the original data. I used the default settings, except for specifying the K-means algorithm and increasing number of iterations to 1000.

2.2.4.2 Genomic distribution of differentiation

The assess the genomic distribution of the alleles underlying the formation of genotypic clusters, I extracted the PC loadings of each locus. These values are approximately normally distributed with a mean of zero. Extreme positive or negative loading values are indicative of loci with particularly large effects on the PC position of individuals (and hence their cluster membership). If a single large region or single locus is responsible for cluster membership, these extreme values should be highly localized in the genome – for example clustered in an inverted region (e.g. as shown in Kupper et al. 2015). I classified extreme loading values by first taking the absolute value of each set of loadings (sign is irrelevant in this case), and then identified loci that exceeded the 95th percentile of this distribution of absolute loadings. I then visualized these extreme loadings, along with the rest of the loadings, by plotting them across the genome.

2.2.4.3 fastSTRUCTURE

To compliment the PCA results, I analyzed the pruned SNP data using *fastSTRUCTURE*, the modern implementation of the original *STRUCTURE* method (Raj *et al.* 2014). This method is conceptually similar to PCA in that it attempts to summarize data by reducing its dimensionality.

However, instead of finding maximal variance, *STRUCTURE* assumes that there is some fixed number of ancestral populations (the "K" value in the model) from which each individual draws some proportion of its overall genotype. Thus, fitting the model with different K-values can help reveal different discrete source "populations" in the underlying data. While there has been some argument over exactly how to assess the significance of these results, i.e. determine the "best K", *fastSTRUCTURE* simply provides a range of K values that are considered equally successful in describing the variation in the data (Raj *et al.* 2014).

2.2.5 Temporal stability of genotypic clusters

To assess the temporal stability of genotypic clusters in the dataset, I calculated the percentage of shared SNPs between each individual in the dataset using the *snpgdsIBS* function in *snpRelate* (Zheng *et al.* 2012). I then asked if male white stickleback collected in different years were more similar to each other than to male common stickleback from the same year they were collected. I assessed the significances of this test using a permutation test. For each iteration of the permutation, I resampled the pairwise differences between each individual, and computed a difference in means (e.g. mean distance between clusters within a year minus mean distance within clusters between years). After 10000 permutations, I computed the number of observations in the null distribution that exceeded the empirical value plus one, and divided this by 10001 to obtain a one-tailed p-value. I then multiplied this number by two to obtained a two-tailed p-value.

2.2.5.1 F_{ST}

To assess overall genetic distance for the groups identified using PCA and structure, I calculated genome-wide F_{ST} between each previously identified cluster. This was done to allow comparison of the divergence between white and common sticklebacks to other stickleback species pairs. Mean F_{ST} is a common metric for summarizing the proportion of genetic variation within vs. between two populations (Beaumont 2005). Using an unpruned dataset and the populations identified in the PCA / fastSTRUCTURE analyses, I estimated pairwise, genome-wide average F_{ST} using Weir and Cockerham's estimator (Weir & Cockerham 1984) provided by the function "--weir-pop" in *vcftools* (Danecek *et al.* 2011).

2.2.6 Assessing phylogenetic relationships and gene flow

In following sections, I examined phylogenetic relationships and assessed evidence of gene flow between the populations identified using the cluster-identification methods above. I did this using *TREEMIX* and *dadi* (Pickrell & Pritchard 2012, Gutenkust et al. 2009).

2.2.6.1 TREEMIX

The *TREEMIX* software (Pickrell & Pritchard 2012) provides several simple approaches to jointly assessing phylogenetic relationships and the likelihood of gene flow between populations. The method works as follows. Using SNP data, the program estimates a covariance matrix that summarizes genetic similarity between populations. It then fits a standard phylogenetic tree that best

explains the genetic similarity between populations. Using this tree as a base, *TREEMIX* can then add a set number of discrete admixture branches (i.e. unconstrained branches between points on the tree) to explain residual covariance between populations. If a tree with admixture branches provides a better fit to the data compared a tree without admixture (assessed via likelihood ratio test), we can infer that a gene flow event may likely occurred between populations in the tree. While this method is somewhat coarse, and assumes discrete admixture events, its ability to recover broad trends in genetic data is robust to heterogeneity in demographic history and continuous gene flow (Pickrell & Pritchard 2012).

I estimated *TREEMIX* trees using a pruned dataset, only including individuals collected in 2014. To help place the putative white-common divergence in phylogenetic context, I first fit trees with both Denmark and British Columbia as outgroups. To assess confidence in the topology of this tree, I performed 1000 bootstrap replicates of the tree fit using the included bootstrapping mode in *TREEMIX*. This method resamples the input dataset with replacement and fits a tree each resampled dataset. I then combined the bootstrap trees into a consensus tree and marked each node with its bootstrap confidence using the functions *consensus* and *prop.clades* in the R package *ape* (Paradis *et al.* 2004).

Next, to test for the signal of admixture, I estimated a base tree using Denmark as an outgroup, followed by sequential addition of migration edges (starting with 0). For each new migration edge, I compared the goodness of fit of the new tree in to the previous tree using a likelihood ratio test (Pickrell & Pritchard 2012). I classified the "best" tree as the last tree to offer a significant increase in likelihood.

2.2.6.2 dadi

A more formal test for gene flow between two populations can be achieved by fitting an explicit evolutionary-demographic model, such as an "isolation with migration" (IM) model to the data. I did this using the software package *dadi* (Gutenkunst *et al.* 2009). Unlike traditional coalescent-based approaches, *dadi* fits models to a summary of genetic polymorphism in two populations, the so-called joint site frequency spectrum (JSFS). Along with several other features, this has the benefit of massively reducing the computational complexity of fitting demographic models to polymorphism data.

Using *dadi*, I fit full IM models (dadi.Demographics2D.IM) to the polymorphism data from two separate pairs of sympatric white and common stickleback (Salmon River and Canal Lake, Figure 2.1B & C). Because the IM model contains many free parameters and overfitting is a concern, I also fit a standard neutral model (no divergence or gene flow, i.e. a single panmictic population) to both data sets. This neutral model has a single free parameter (the effective population size), and comparing it to the full IM model can serve as a baseline for the improvement in fit offered by the full IM model. This is especially useful for in the case of very high migration and/or recent divergence, which may be better modelled as panmixia.

To avoid computationally intensive bootstrapping of the *dadi* model fits, I fit all *dadi* models to the pruned dataset (i.e. there was little to no linkage disequilibrium between markers). This allows standard likelihood ratio tests to be used for comparing demographic models.

Importantly, *dadi* also assumes that the sites underlying the JSFS are evolving neutrally. To identify sites under selection, I performed an F_{ST} outlier scan using vcftools. The complete details of this scan are outlined in Chapter 3. Briefly, I classified loci as "outliers" on the basis of extreme

values of F_{ST} (>99th percentile). I then divided the genome into 75 000 bp windows and performed a permutation test to determine windows that were significantly enriched for extreme outliers. I then removed all sites in the data set that occurred in these windows.

dadi uses a stochastic optimization algorithm to determine model parameters, and thus any given set of iterations may fail to find the best fit (Gutenkunst *et al.* 2009). Based on the recommendations in Gutenkust et al. 2009, I ran initial 20 replicates of the IM model for each comparison. I used arbitrary starting points and highly permissive bounds for the model parameters in these initial runs (see code supplement). I then determined the model parameters with the highest log likelihood from these runs, and performed an additional 20 replicates using this set of parameters as the starting point.

2.3 Results

2.3.1 Genotypic clustering

PCA projection of the genotypic data revealed three distinct genotypic clusters in the data (Figure 2.1A). Two of these clusters (white and blue dots in Figure 2.1A) closely coincided with the classification of white and common individuals based on body size classes (Figure 2.1B). The cluster containing the smallest individuals (putatively white stickleback) contained nearly every male I had classified as "white" in the field (light breeding colours, nesting in algae) in both 2012 and 2014. These two types (white/common) co-occurred at many of the sites where I sampled (Figure 2.1 C &D), suggesting they do not represent neutral geographic structure. The third cluster (green dots, Figure 2.1A) appeared to comprise stickleback exclusively from the Bras d'Or region (Figure 2.1D).

This cluster also contained three males that I scored as "white" in the field, although there is no evidence of small bodied males at any Bras d'Or sites (Figure 2.1B). Application of the bootstrap and perturbation method of Hennig (2007) showed all three clusters to be highly significant (Table 2.1).

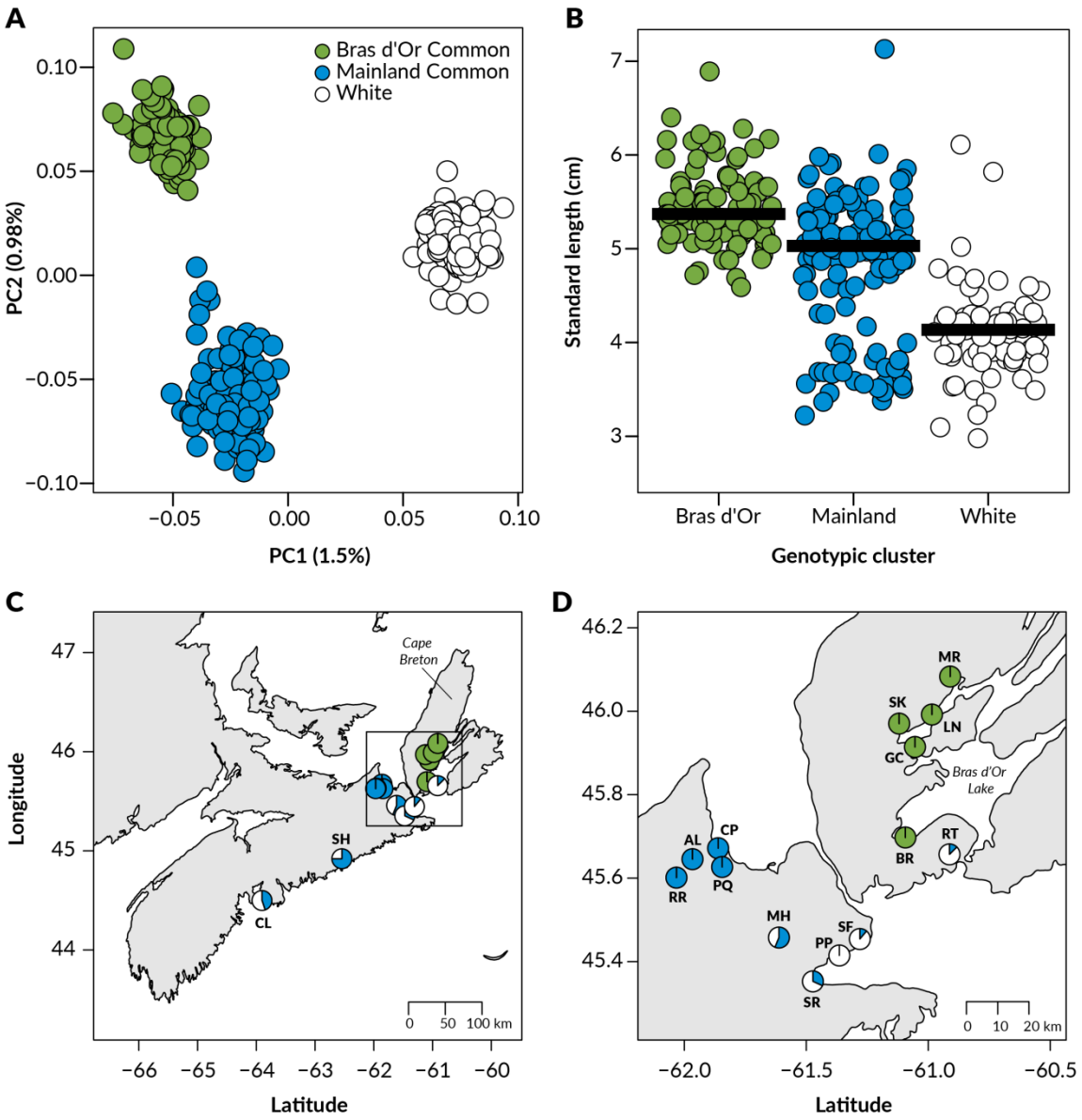

Figure 2.1 | A Genotype clustered in stickleback polymorphism data on two principal components. Cluster colors represent K-means cluster groupings (K=3). **B** Standard lengths of individual stickleback from each genotypic cluster. Black bars represent cluster means. **C,D** Map of the geographic distribution of genotypic clusters. Pie chart sections represent the proportion individuals at each site belonging to each genotypic cluster in A. Site labels: CL = Canal Lake, SH = Sheet Harbour, RR = Rights River, AL = Antigonish Landing, CP = Captain's Pond, PQ = Pomquet, MH = Milford Haven River, SF = St. Francis Harbour, PP = Porper Pond, SR = Salmon River, RT = River Tillard, BR = Black River, GC = Gillies Cover, SK = Skye River, LN = Little Narrows, MR = Middle River.

Table 2.1 | Results of Henning (2007) cluster stability tests on the three stickleback genotypic clusters shown in Figure 2.1. Each value represents the proportion of cases out of 1000 replicates in which the original cluster membership was recovered after perturbation. High proportions indicate more cluster stability.

Perturbation	White	Common	Bras d'Or
Bootstrap	0.920	0.906	0.885
Noise	0.943	0.928	0.910
Jitter	0.936	0.927	0.912

2.3.1.1 Genomic distribution of differentiation

PC1 was the key axis separating white and common stickleback (Figure 2.1). I thus focused on the distribution of PC1 axis loadings across the genome. The loci contributing the most to the formation of the genotypic clusters (loci in the 95th percentile of axis loadings) were distributed across every chromosome (Figure 2.2). This suggests that divergence between white and common stickleback has occurred genome wide, rather than in a single region.

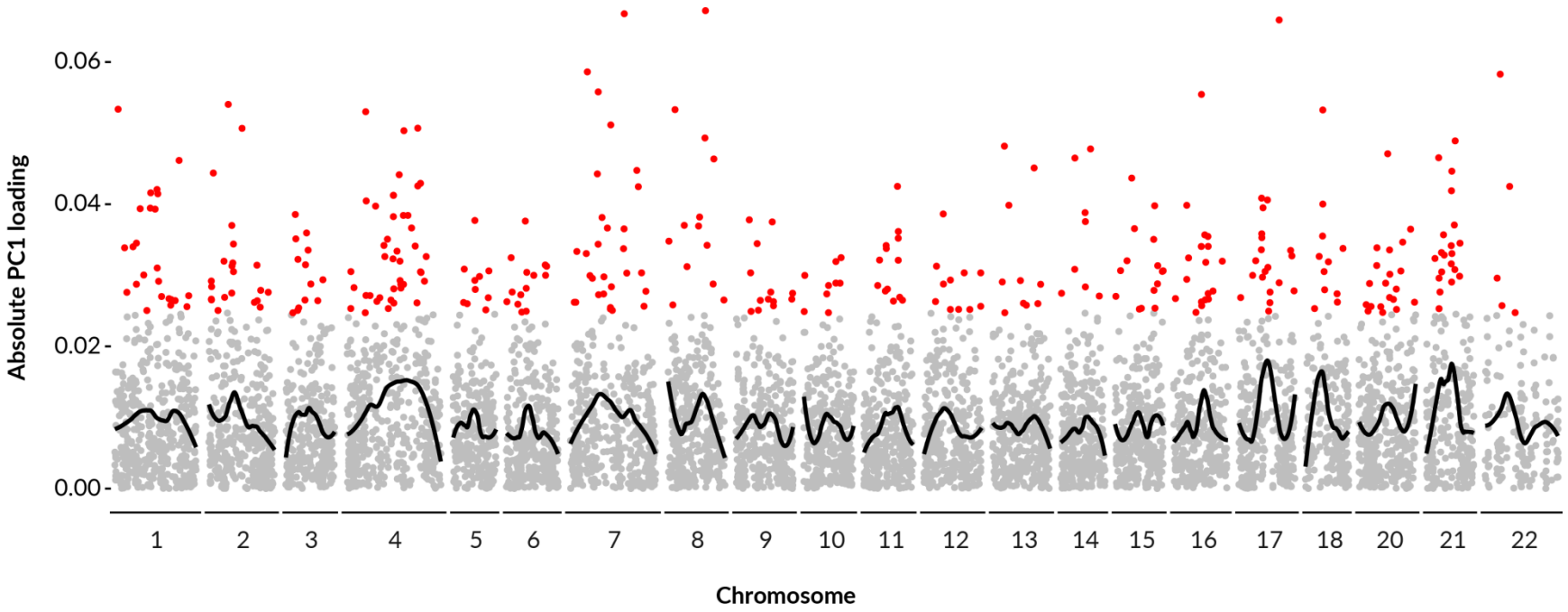

Figure 2.2| The genomic distribution of principal component (PC) axis loadings for the PC on which white and common sticklebacks are separated (PC1). Each point indicates the absolute value of the axis loading for a single locus. Red points indicate that the PC loading for that locus exceeds the 95th percentile of all axis loadings. Black lines show a LOESS smooth of axis loadings for each chromosome. The sex chromosomes were omitted (see text for details). Chromosome "22" is concatenated, unplaced scaffolds.

2.3.1.2 fastSTRUCTURE

fastSTRUCTURE *chooseK* identified a range of 1-3 clusters in the genotypic data (Figure 2.3). Examining the assigned groups in the plot revealed that they are identical to the clusters identified by PCA (Figure 2.3, note regional groupings). This includes the existence of discrete "white", "mainland" and "Bras d'Or" groups. Mainland common stickleback from areas near the Bras d'Or lake (e.g. Guysborough) appear to have small amounts of mixed ancestry from both Bras d'Or lake common stickleback and white stickleback (Figure 2.3, K=3).

2.3.2 Temporal stability of genotypic clusters

When we examined years separately, individuals in the 2012 and 2014 white clusters were more closely related to individuals in white clusters sampled in different years than they were to individuals from common clusters sampled in the same year (Figure 2.4, Permutation test: within year between cluster vs. between year within cluster, $p = 0.0018$, 10000 permutations).

2.3.3 Overall genomic divergence

In spite of forming discrete genotypic clusters, white and common stickleback were only weakly genetically differentiated. Because three distinct groups emerged in the cluster analysis, I computed genome-wide Weir and Cockerham F_{ST} between each pair of clusters in the 2014 and 2012 data set. The F_{ST} estimates for 2014 were as follows: White vs. Mainland: 0.015 (n = 177), White vs. Bras d'Or: 0.02 (n = 166), Common vs. Bras d'Or: 0.013 (n = 197). The overall F_{ST} for

White vs. Mainland commons in 2012 was 0.019 (n = 90) (I did not sample in Bras d'Or during 2012).

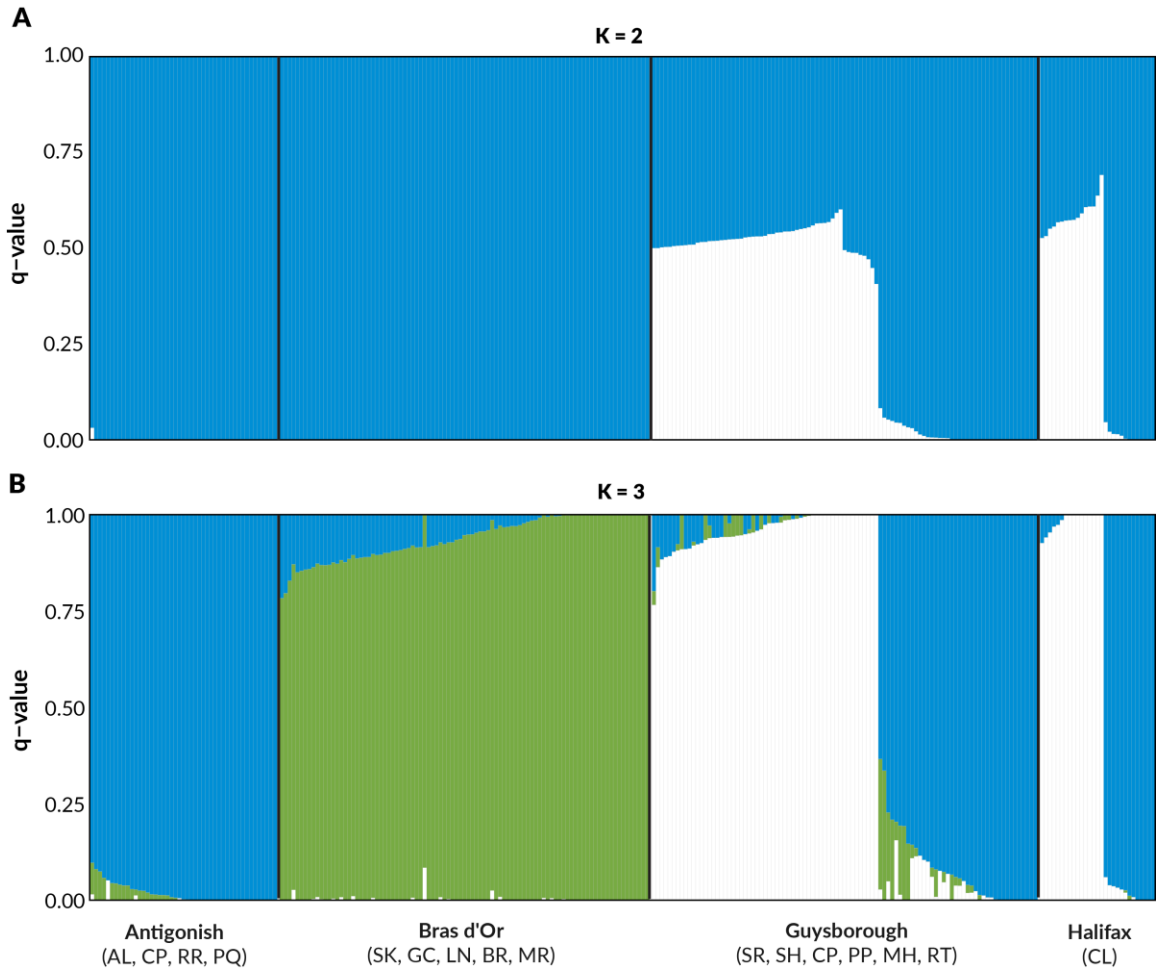

Figure 2.3 | Stacked bar plot of fastSTRUCTURE ancestry proportions for Nova Scotian stickleback sampled in 2014. Bars are grouped by general geographic area. The proportion of each color in each individual bar represents the estimated proportion of that individual's genome belonging to a population. Results from K = 2 and K = 3 are shown in A and B respectively. The "chooseK" script provided with fastSTRUCTURE reports a best fit between 1 and 3 groups. Complete location labels are provided in Appendix Figure A.1.

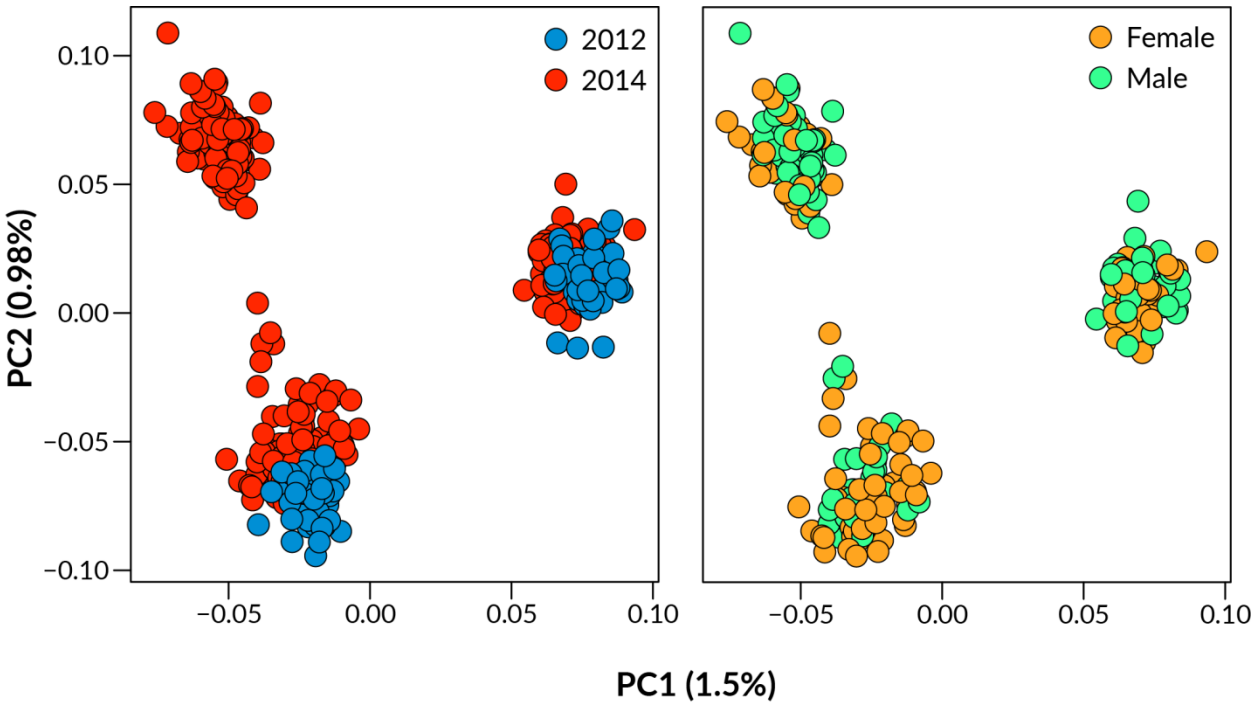

Figure 2.4 | Projection of Nova Scotia stickleback genetic polymorphism data on two principal components, colored based on collection year (left panel) or sex (right panel). 2012 males (blue points in left panel) were removed in the right panel for clarity. Note that only males were collected in 2012. Percentages in axes labels indicate the percent of total genetic variation explained by each principal component.

2.3.4 Assessing phylogenetic relationships and gene flow

2.3.4.1 TREEMIX

Like PCA and fastSTRUCTURE, the consensus TREEMIX tree recovered three groups of Nova Scotian stickleback (Figure 2.1). The white, common and Bras d'Or clades all have strong

support for their individual monophyly and strong support for whites and mainland commons as sister groups (Figure 2.5A, 0.85-1.00 bootstrap support),.

The best fit for number of migration edges for the Nova Scotia/Denmark tree was three (Figure 2.5B, likelihood ratio test: $\chi^2_1 = 15.52$, $p = 0.000081$). The strongest migration edge connected a western white population (Canal Lake) to another white population to the east (Sheet Harbour) (Figure 2.5B, red edge). The remaining two edges connected the white clade to Guysborough common populations, mirroring the signal of admixture from the fastSTRUCTURE analysis (Figure 2.5B, yellow edges).

2.3.4.2 dadi

Both the standard neutral model (i.e. panmixia) and isolation with migration model produced reasonable fits to the JSFS for the Canal Lake and Salmon River population pairs (Figure 2.6). However, likelihood ratio tests revealed that the isolation with migration model provided a significantly better fit than the neutral model for both cases (Canal Lake: $\chi^2_6 = 3569.52$, $p << 1 \times 10^{-6}$, Salmon River: $\chi^2_6 = 492.52$ $p << 1 \times 10^{-6}$). The isolation with migration fits for Canal Lake and Salmon River both estimated moderate divergence times (~1 million years), moderate population size contractions relative to the ancestral population (~25% reductions in effective population size), and moderate migration rates (~2 migrants per generation) (Table 2.2).

There was a large degree in variability of the parameter estimates over subsequent runs (Appendix Figure A.3). This may mean that the parameters that best fit the JSFS for Salmon River and Canal Lake exist in a flat part of the likelihood surface. This behavior is well known for demographic models in cases where migration is very high and/or divergence is very recent

(Gutenkunst *et al.* 2009). That said, migration rates were generally estimated to be high and divergence times estimated to be low (Appendix Figure A.3). While this suggests that recent divergence and gene flow are indeed likely, the estimates for divergence time and migration rate should be interpreted cautiously.

Table 2.2 | Demographic model parameters estimated by *dadi* for the standard neutral model and isolation with migration model. Values shown were back-transformed from *dadi* units using the formulas in the dadi manual (Gutenkunst *et al.* 2009). Original parameter estimates are shown in parentheses. Parameters are: n_{ref}, effective population size of the ancestral population; s, relative size of population 1 after split (size of population 2 1-s); n_1 and n_2 relative size of population 1 and 2 at present; t, divergence time (generations) between population 1 & 2; m_{12}, migration rate (individuals per generation) from population 1 to population 2; m_{21}, migration rate from population 2 to population 1.

	Isolation with migration parameter estimates						
Population	n_{ref}	s	n_1	n_2	t	m_{12}	m_{21}
Salmon River 30 individuals 26 003 loci	823 972 (θ = 10048.34)	143 700 (0.17)	121 031 (0.20)	175 773 (0.31)	1 976 126 (1.12)	3 (39.60)	2 (18.91)
Canal Lake 30 individuals 25 844 loci	862 741 (θ = 10048.34)	237 253 (0.27)	126 726 (0.15)	184 043 (0.21)	997 353 (0.58)	3 (39.60)	2 (21.10)

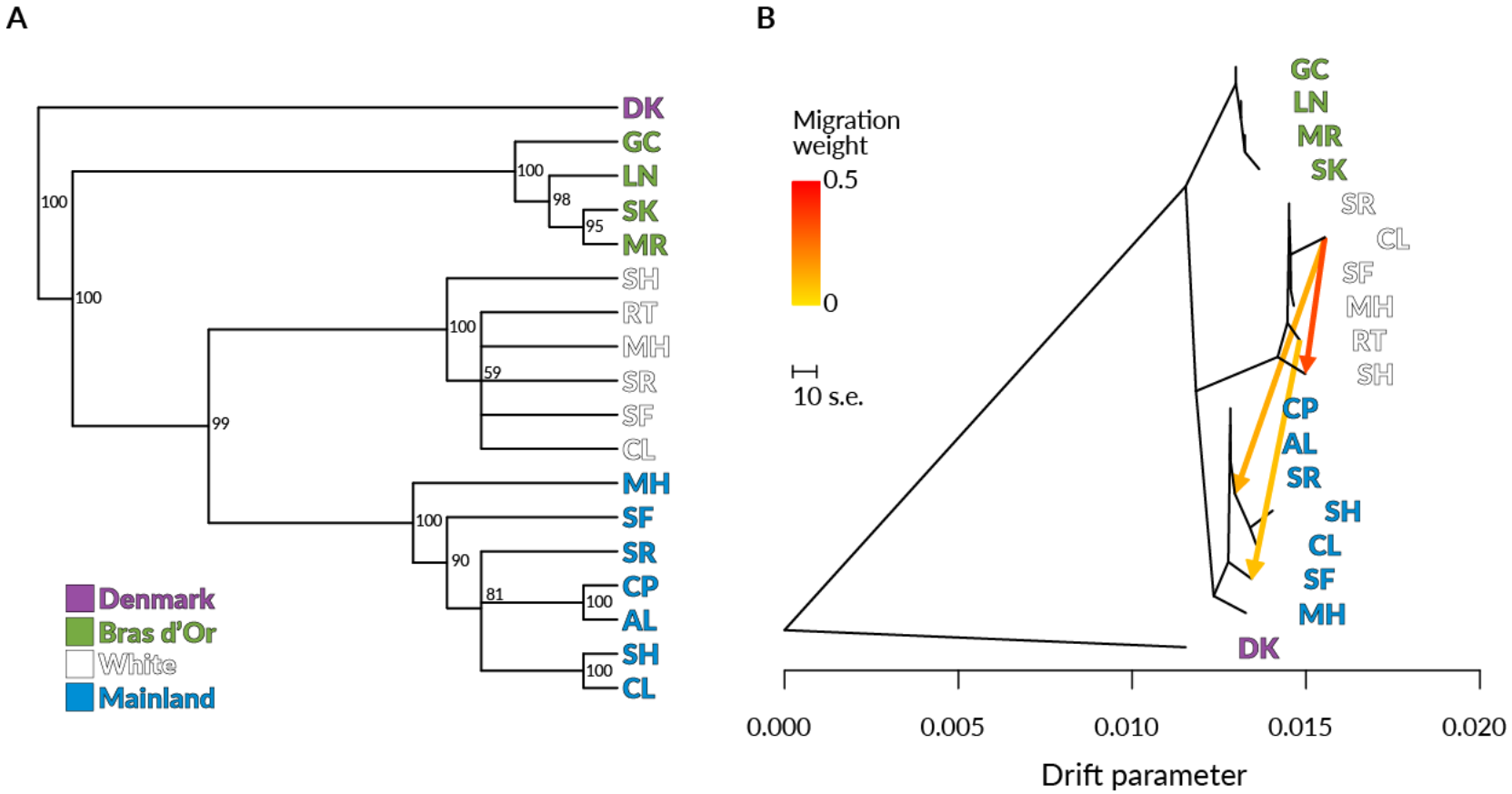

Figure 2.5| TREEMIX trees for Nova Scotian stickleback sampled in 2014, with a Denmark marine population as the outgroup. Trees were fit using putatively neutral markers only (see text for details). **A** A consensus tree derived from 1000 bootstrap replicates, assuming no migration. Node labels represent the percentage of trees in which each node exists. Larger numbers represent more confidence in the node. Branch lengths are arbitrary. Tip label colours correspond to genotypic clusters. **B** The maximum likelihood TREEMIX (m = 3 migration edges). Migration edges are colored according to their weight (more red = higher relative migration). The drift parameter corresponds to the estimated amount of genetic drift that has occurred between populations.

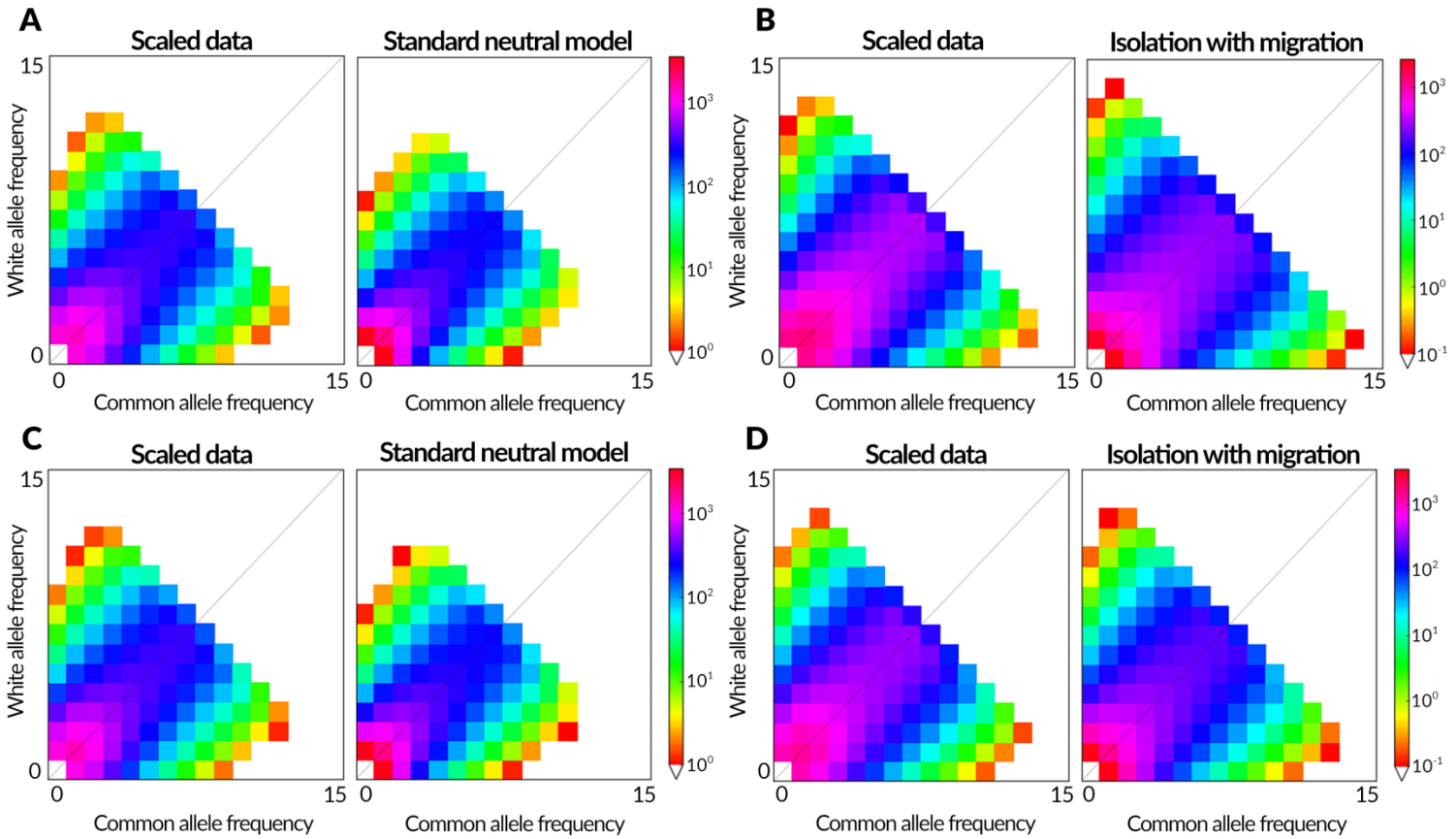

Figure 2.6| Scaled empirical joint site frequency spectra (left panel in each pair) and model fits (right panel in each pair) from two demographic models fit using *dadi*. All spectra are displayed as 'folded' and thus allele frequencies (all axes) represent scaled minor allele frequencies. The density of observations in each 2D bin is represented by hue (scales vary slightly among plots). **A, B** Standard neutral model and isolation with migration models fit to white and common populations from Canal Lake, Nova Scotia. **C, D** Standard neutral model and isolation with migration models fit to white and common populations from Salmon River, Nova Scotia. Parameter estimates for isolation with migration models are listed in Table 2.

2.4 Discussion

Species in the early stages of divergence, particular those that still exchange genes with their sister taxa, are excellent systems for studying speciation. However, identifying such systems is inherently difficult. Here, I used modern population genomics methods and attempted to clarify whether a novel "white" threespine stickleback found in Nova Scotia is likely in in the early stages of divergence. Using a variety of methods, I found compelling evidence that the white stickleback is genetically distinct from, but very closely related to, the common stickleback with which it is sympatric. This close relationship appears to the result of the joint effects of gene flow and very recent divergence between the two types. Below, I discuss the evidence for these findings, as well as the possibility that white sticklebacks represent some form of alternative male reproductive strategy .

2.4.1 White sticklebacks: alternative mating strategy?

There are three forms of male alternative reproductive strategy that the white stickleback could represent: genetic, ontogenetic, and condition-dependent. I address the evidence for each of these below.

The white stickleback is unlikely to represent a genetically-determined alternative male strategy. Alternative male strategies are typically predicted to have a simple genetic basis, in order to preclude the possibility of intermediates (Gross 1996, Kopp and Heimson 2006, Taborsky 2008). These predictions are consistent with known examples (citations) – in all cases a single locus, or a large non-recombining supergene (Gross 1996, Kupper et al. 2015). The genome-wide differentiation between white and common stickleback is not consistent with this model. Instead,

these groups show genetic changes distributed over many loci and an overall reduction in gene flow (i.e. genealogical independence) – such as we might expect for in the case of partial reproductive isolation.

My results also indicate that the white stickleback is not a conditional strategy. A pure conditional strategy involving no genetic polymorphism would involve a genetic basis that is shared among all individuals (Gross 1996). The fact that white and common stickleback are genetically differentiated across their range, including at sites where they are sympatric, argues strongly against this possibility.

White stickleback are also likely not an ontogenetically-determined strategy. If this were the case, the genetic basis of the strategy would again be shared between white and common sticklebacks. However, one possibility is that the genetic differences we see are a cohort effect, and the common and white sticklebacks are different age classes (perhaps one and two year old breeders). This is unlikely to be the case. First, although we do not have direct estimates of the age of the fish, marine stickleback generally have a one year life cycle. Secondly, the genetic differences we see between white and common stickleback genetic clusters are stable through time, ruling out a cohort effect as the cause. For example, under a pure cohort model there is no reason why 2012 and 2014 white stickleback should be closely related. The ontogenetic and conditional strategies are also both not consistent with Blouw's (1996) qualitative reports (Blouw 1996) that white and common sticklebacks breed true under laboratory conditions (we would eventually expect some generation of the "wrong" type if the strategy is purely conditional or ontogenetic). Thus, my results again better fit Blouw and Hagen's (1990) model of two partially reproductively isolated lineages rather than an alternative male strategy.

2.4.2 White sticklebacks are likely an incipient species

While the high-resolution methods we used indicate white stickleback are genotypically distinct from common stickleback, they are still very closely related to the sympatric common stickleback. The overall F_{ST} between white and common stickleback in Nova Scotia is ~0.02. For a phenotypically identifiable species, this is unusually low: compare to ~0.4 for benthic-limnetic species pairs and ~0.2 for lake-stream ecotype pairs in British Columbia (Taylor & McPhail 2000; Roesti *et al.* 2012).

This high genetic similarity appears to be the result of very recent divergence coupled with on-going gene flow. The short-branched topology of the TREEMIX tree suggests very recent divergence between white and common stickleback, particularly in context of the Atlantic/Pacific split (Appendix Figure A.2). In spite of the branch lengths themselves likely being shortened to some degree by gene flow, the addition of three white-to-common migration edges further improves the fit the TREEMIX tree. In sum, TREEMIX clearly favors a model in which white and common stickleback diverged recently and continue to exchange genes.

The parameters of the demographic models fit using dadi support this idea. The best-fit parameter estimates for the isolation with migration models all indicated high migration and low divergence times. This appeared to the case in both pairs of populations I examined. Further, the standard neutral model (i.e. panmixia) alone provided a reasonable fit to the data, suggesting that white-common gene flow is likely substantial.

In light of the recent divergence and large amount of inferred gene flow, it is surprising that there do not appear to be any clear hybrid individuals. Assuming the genetic differences between white and common stickleback are numerous and found throughout the genome, hybrid individuals

ought to have manifested as intermediates in the PCA projection, or as having large amounts (~50% for an F1) mixed ancestry in the fastSTRUCTURE plots. The only discernable evidence of this is a small amount of mixed ancestry in the Guysbourough common populations, which may also be attributable to unsorted ancestral polymorphism. There are several possible explanations for a lack of hybrids. First, we did not collect ambiguous-looking males, which may have been more likely to be hybrid individuals. This does not, however, account for a lack of hybrid females in my sample. Secondly, it is possible that our sampling method was somehow biased against finding hybrids. For example, perhaps hybrids have transgressive preferences for nest sites or timing of mating, precluding sampling. Finally, it is possible that reproductive isolation between white and common stickleback has become nearly complete extremely recently. Indeed, Blouw's laboratory and field trials suggested that there is near-perfect assortative mating between white and common stickleback. If pre-mating isolation is indeed as strong as these experiments suggest, it is not surprising that we did not detect any hybrids.

2.4.3 The Bras d'Or common clade

While not the focus of this study, we discovered a genetically distinct group of common sticklebacks in the Bras d'Or region. These populations may represent a geographically isolated group of common stickleback. While the Bras d'Or "lake" (actually an inland sea) and mainland coastal waters may seem currently navigable by stickleback, the southern channel out of the Bras d'Or lake was only opened as a commercial canal ~100 years ago. Previously, the only oceanic passage into the lake was through two openings over 150 km away (the Great Bras d'Or and St. Andrew's channels). Further, the Bras d'Or lake has historically fluctuated between being glaciated, a

freshwater body disconnected from the ocean (i.e. a true lake), and its current state as an inland sea (Shaw *et al.* 2006). Thus, there has been ample opportunity for geographic isolation between mainland and Bras d'Or stickleback, both via distance and physical barriers.

2.4.4 Unexplained patterns

While the genetic data paints a clear picture of the white stickleback as an incipient species, there are some unexplained patterns. First, while the 'white' genotypic cluster clearly contains much smaller individuals than the two other clusters, there still seems to be bimodality in the body size distribution in the 'mainland' common marine stickleback cluster (Figure 2.1 B, blue dots). It is not clear what these 'small common' individuals represent. One possibility is that it represents an age class (first year instead of second year breeders). If so, we are unable to explain why we do not see them in the other clusters. Another possibility is that perhaps the alleles that cause small body size in white stickleback are present at an appreciable frequency in the common population, creating a genetically based size polymorphism.

The second strange element is the complete lack of individuals from the 'white' genotypic cluster in the Bras d'Or lake, even at sites where I saw males with light coloration (Appendix Figure A.1). This is made stranger by the fact that the handful of males I directly scored as "white" at these populations failed to cluster with the rest of the white stickleback from the mainland. Like the body size issue above, this may be the result of the alleles that cause white nuptial colors segregating in the Bras d'Or population. Perhaps persistent gene flow between Bras d'Or and whites (directly or via mainland commons) and/or selection of some form maintains a color polymorphism independent

of the other loci involved in full blown 'whiteness'. Further work dissecting the genetics of the white/common difference will hopefully elucidate these issues.

2.4.5 Future work

Our findings suggest that the white stickleback is an excellent candidate system for studying the genetics of speciation. This opens a number of exciting possibilities for future work. The most obvious is a closer examination of the genomic distribution of loci contributing to divergence between white and common populations. This could be complemented by a QTL mapping project examining the genetic basis of white nuptial color, loss of parental care, etc.

In addition, the white stickleback may serve has an excellent test-bed for theories about the interacting roles of ecological and sexual selection in speciation. White stickleback appear to have diverged in mostly sexually-related traits rather than ecological traits – is this indeed the case? If so, perhaps sexual selection was an important driver of the evolution of reproductive isolation in this system. While theory suggests that speciation via sexual selection alone is difficult in the face of gene flow (Servedio & Kopp 2012; Servedio & Bürger 2014), a recent model suggests that spatial variation in resources (e.g. nest sites) can dramatically increase the probably of speciation via sexual selection (M'Gonigle *et al.* 2012). This model is particularly interesting in light of the fact that the white stickleback syndrome includes a shift in nest site preference. Overall, white stickleback will likely be a fruitful system for empirical evaluation of models of sexual-selection speciation models.

2.4.6 Conclusions

Here, I used high-throughput genomic methods to explore whether white stickleback represent a nascent species. While much remains to be learned about these strange animals, the evidence I presented here suggests that they are likely a young species. White stickleback likely do not represent a genetically determined male mating strategy polymorphism, nor young or low-condition individuals. Instead, they form a unique genotypic class, distinct from common stickleback. In spite of this, their genome-wide differentiation remains very low, and gene flow has likely been very strong throughout their divergence. Together, my findings suggest that white stickleback are not only a nascent species, but an excellent system for studying speciation.

Chapter 3: White stickleback exhibit sexual and genomic divergence in the absence of ecological differentiation

3.1 Introduction

The evolution of reproductive isolation – speciation – is responsible for much of the diversity on earth. Accordingly, understanding this process remains a central goal of evolutionary biology (Marie Curie Speciation Network *et al.* 2012). Recent years have seen a genomics-fueled renaissance in speciation research, and many new approaches to solving Darwin's "mystery of mysteries" (Noor & Feder 2006; Butlin 2010). Much of this new research has so far focused on species exhibiting strong reproductive isolation and substantial genomic divergence (Seehausen *et al.* 2014). This is problematic, as the traits that matter for the initial evolution of reproductive isolation become masked as divergence proceeds (Coyne & Orr 2004). There is thus a serious dearth of knowledge about the types of traits involved in the initial build-up of reproductive isolation.

3.1.1 Trophic and sexually dimorphic traits

Reproductive isolation often evolves as a by-product of divergent natural selection (Dobzhanky 1951). In many cases, this divergence is mediated by adaptation to different habitats and diets (Schluter 2009). Comparative studies of animals have shown that this adaptation is overwhelmingly focused on coping with the biotic aspects of the environment: predators, competitors, food and mates (Benton 2009). Traits linked to these aspects can be broadly grouped

into two classes: those involved in trophic interactions such as feeding or anti-predatory traits, and sexually dimorphic traits linked to reproduction, such as sex-specific mating cues and behaviors (Schluter 2001; Coyne & Orr 2004). For simplicity, I will refer these as "trophic" and "sexual" traits.

The extent to which sister taxa are diverged along trophic vs. sexual lines varies considerably among groups of animals. In the threespine stickleback species complex, all described species pairs differ in trophic characters such as gill rakers, head morphology and body shape (McKinnon & Rundle 2002). Further, extensive research has shown that reproductive isolation in stickleback is directly linked to trophic differentiation (Hendry et al. 2002; Rundle & Schluter 2004; Taylor *et al.* 2011). In contrast, among the swordtail cricket species of Hawaii, trophic differentiation is weak (Mendelson & Shaw 2005). Instead, reproductive isolation among these insects appears to be largely maintained by differences in sexually dimorphic traits – male mating signals and matching female preferences (Grace & Shaw 2011) .

There are a variety of reasons why *both* trophic and sexual traits are likely important during the early phases of speciation. Divergence in many sexually dimorphic characters is likely to generate strong pre-mating isolation as by-product, and thus could provide a key initial boost to reproductive isolation (Ritchie 2007). This could be particularly important for speciation without geographic isolation, where speciation is far more likely to occur if initial reproductive barriers are strong (Servedio & Kopp 2012). Sexual divergence can also occur in the absence of ecological differences (e.g. via Fisherian sexual selection), suggesting that there may be fewer constraints on the evolution of sexual characters than trophic characters (Safran *et al.* 2013). However, divergence in trophic characters can also rapidly generate reproductive isolation -- extrinsic post-zygotic isolation arising from reduced hybrid ecological performance being a prime example (e.g. Arnegard *et al.* 2014).

Overall, much more work is needed to begin to resolve the relative importance of trophic vs. sexual divergence in the initial stages of speciation (Maan & Seehausen 2011b, Safran *et al.* 2013). Some authors argue that trophic divergence is an essential ingredient in the early stages of the evolution of reproductive isolation via natural selection (Muschick et al. 2012), while others have speculated that divergence in sexual characters can be sufficient (perhaps even in the absence of ecological differentiation) (Safran 2013). Testing these alternatives will ultimately require studies of trophic and sexual divergence in a variety of very young species.

3.1.2 Studying the roles of trophic and sexual traits

There are a number of approaches to studying trophic differences between species. A simple approach is to examine morphological traits known to be associated with feeding and/or anti-predatory traits. For example, fish species that make use of different niches often differ in body size, body shape, armor, and the number and length of gill rakers (part of the feeding apparatus) (McKinnon & Rundle 2002). This approach has proven useful across a variety of systems.

Another powerful approach to assessing trophic differences is stable isotopic analysis (Hobson & Wassenaar 1999; Reimchen *et al.* 2008). In fish, the abundances of Carbon-13 and Nitrogen-15 have been shown to be sensitive markers for the amount of dietary carbon derived from littoral sources and trophic position respectively (Reimchen *et al.* 2008; Matthews *et al.* 2010; Stasko et al. 206). The use of stable isotopes also enjoys the benefit of not requiring previous knowledge about the morphological traits involved in trophic differentiation. As such, stable isotopes remain a widely-used method of measuring trophic differences (Fry 2007).

Measuring differences between nascent species in sexual traits is more challenging. In most species, many sexual traits behaviors and pheromonal traits are cryptic or difficult to measure. One approach, and the one used here, is to measure divergence in genes with sex-biased expression. This approach works as follows. First, The genomic regions underlying species differences are identified using divergence-based methods such as F_{ST} (Beaumont & Balding 2004). Second, the degree of sex-biased expression at genes within divergent regions is examined. This can be done by directly measuring expression, or using data from studies examining sex-biased expression in the species of interest (Viitaniemi & Leder 2011). This approach has been widely applied to the study of sexual selection, as genes with sex biased expression are more likely to be the targets of sexual selection (Zhang *et al.* 2007; Ellegren & Parsch 2007; Perry *et al.* 2014). Such a genomic approach also has the benefit of being able to reveal differences between the sexes in unmeasured traits (Ellegren & Parsch 2007).

This approach is not without limitations. One key assumption of our application of this approach is that divergence in sex-biased genes serves as a proxy for divergence in purely sexual characters. However, trophic differences can themselves be sexually dimorphic (Reimchen & Nosil 2008). Secondly, this approach does not allow us to detect the magnitude of phenotypic difference in sexual characters between species (although genetic divergence is perhaps a more informative of a traits role in reproductive isolation) . Sex-biased divergence data must thus be interpreted in the context of other morphological or isotopic data in order to explore this possibility.

3.1.3 Aims and study system

Here, we make use of a recently discovered young species of stickleback – the white stickleback -- to test the relative importance of differences between the species in trophic traits and sexually dimorphic traits linked to reproduction. White stickleback are a unique form of marine threespine stickleback (*Gasterosteus aculeatus*) found only in marine coastal Nova Scotia, Canada (Blouw & Hagen 1990). Compared to "common" marine stickleback, white stickleback have smaller body sizes, unique pearlescent-white male nuptial colors, and reversed brain size sexual dimorphism (Blouw & Hagen 1990, Samuk et al. 2014). White stickleback are sympatric with common marine stickleback throughout Nova Scotia (Jamieson *et al.* 1992b). Previous work by our group has shown that white stickleback are a recently evolved, genetically distinct group that likely has diverged from the common form in the face of gene flow (this thesis, Chapter 1). Work by Blouw and colleagues has shown that during the breeding season, white stickleback specialize on habitats rich in filamentous algae, which male white stickleback use as substrate and material for their nests (Blouw & Hagen 1990). Unusually, these males appear to lack the parental care behaviors for which stickleback are widely known (Jamieson *et al.* 1992a; Macdonald *et al.* 1995; Blouw 1996).

White stickleback are an ideal system in which to study the relative roles of trophic and sexual divergence during speciation. For one, because they evolved so recently, the barriers that currently separate the species are likely to have played a key role in the initial (and perhaps on-going) evolution of reproduction isolation (Coyne & Orr 2004). Secondly, a large amount is known about the phenotypic traits important for trophic interactions in threespine stickleback, allowing for a realistic assessment of trophic divergence (Schluter & McPhail 1992; Gow *et al.* 2007; Arnegard *et al.* 2014). Finally, threespine stickleback have a high quality reference genome, allowing us to pinpoint

the genomic regions involved in reproductive isolation (Jones *et al.* 2012; Glazer *et al.* 2015) and examine divergence in sex-biased genes.

3.1.4 Questions

Here, we test whether the white stickleback has diverged along trophic lines, akin to other threespine stickleback species, or along an alternative route focused on sexual traits. To do this, we leverage three types of data: morphological, isotopic, and genomic. We ask the following questions:

(1) Do white and common stickleback differ in the key morphological traits associated with trophic differences in other stickleback species?
(2) Do white and common stickleback differ in isotopic ratios of carbon and nitrogen, which are associated with diet and habitat differences in other stickleback ?
(3) Are highly diverged genomic regions between white and common stickleback enriched for genes displaying sex-biased expression in stickleback?

3.2 Methods

3.2.1 Collection of specimens and species identification

We used the same specimens here as described in Chapter 1. In total, we included 296 individuals sampled in 2014, and 96 individuals sampled in 2012. We collected specimens of both types from 15 locations in Nova Scotia (see Chapter 1). Common stickleback were present at all

sites, whereas white stickleback were present at a total of seven sites (see Appendix 1). We used genetic methods to classify individuals as common or white stickleback as described in Chapter 1. We excluded 4 males from Little Narrows that displayed a mix of white and common traits.

3.2.2 Morphological traits

We measured a suite of morphological characters known to correlate with trophic divergence in other stickleback and fish species ("trophic traits"). The first of these was body size and shape, which are well known to correlate with trophic differentiation in fish, particularly along the benthic - limnetic axis of trophic specialization (Schluter & McPhail 1992, Sharpe 2008). Next, we examined body armor – spines and lateral plates – which are known to evolve in response to the presence of different types of predators (Reimchen & Nosil 2004; Marchinko 2009). We also measured the number of short and long gill rakers, which are cartilaginous projections on the gill arches whose number is positively correlated with the amount of zooplankton (instead of macroinvertebrates) in the diet of stickleback (Bolnick & Lau 2008; Conte *et al.* 2015). General visual brightness of the body (see below) serves important anti-predatory functions in fish (e.g. crypsis, Greenwood et al. 2012), and we thus included this in our panel of traits.

Finally, to represent sexual traits we measured egg size (mass) in females and mass of the testes in males, as energetic investment in these organs are known to correlate with life history differences that arise as a result of ecological adaptation (Soulsbury 2010). Increased male-male competition can also lead to the evolution of increased testes mass (Pitcher *et al.* 2005).

3.2.3 Quantification of morphological differences

Body shape

To measure body shape, we determined the coordinates of morphometric landmarks on digital photographs of individual fish using *imageJ*. We used the landmarks described in Sharpe et al. (2008) (shown in Figure 3.1). We then imported the coordinates of these landmarks into *R* where we analyzed them using the *geomorph* 2.0 package (Adams *et al.* 2014). We performed generalized Procrustes analysis to align and scale the landmarks, followed by principal components analysis to identify the major axes of variation. We examined the first six principal components, which accounted for the majority (77%) of the variation in body shape. As found in other studies (Albert *et al.* 2007), the first principal component (PC1) represented differences in the degree of bending in specimens due to preservation. We thus restricted our analysis to PCs 2-5.

Skeletal traits

To measure skeletal traits, we first stained the fish using Alizarin red following the protocol in Arnegard *et al.* (2014). We then took digital photographs of the stained specimens, and counted the number of lateral armor plates and then measured the length of spines using *imageJ* (Rasband 2012). To count gill rakers, we dissected out the first gill arch and examined it under a dissecting microscope. We then counted the number of short and long gill rakers, again following the methods of Arnegard *et al.* (2014).

Body brightness

We quantified brightness of the body by extracting a 1 cm^2 section of the flank of each fish from a digital photograph taken under constant light conditions using *imageJ*. We then obtained the mean RGB values (0-255 for each channel) for these segments using *imageJ*, and calculated an overall luminance score: R+G+B/(255*3).

Testes and egg mass

We quantified testes mass by dissecting out the testes from each male, drying them in a desiccator and weighing them using a XS3DU microbalance (Mettler-Toldeo, Ohio). When both testes were developed, we weighed both and took the average of the resulting measurement; otherwise, we measured the single developed testis. We quantified egg size by extracting up to ten individual eggs from each female, and weighing them in the same manner as the testes. We divided the final weight by the number of eggs measured.

Analysis of morphological differences

We tested for differences in mean trait values between species using a one-way ANOVA (for continuous traits) or via a Chi-Squared test applied to an Analysis of Deviance (for meristic traits). When comparing morphological traits among species, it is conventional to control for body size via ANCOVA. Thus, we also performed separate ANCOVAs for each trait, examining differences after controlling for standard length by including it as a covariate. Finally, we also performed all analysis

(ANOVA and ANCOVA) with sex as a covariate. All tests were performed using *R* 3.2.2 (Team 2015).

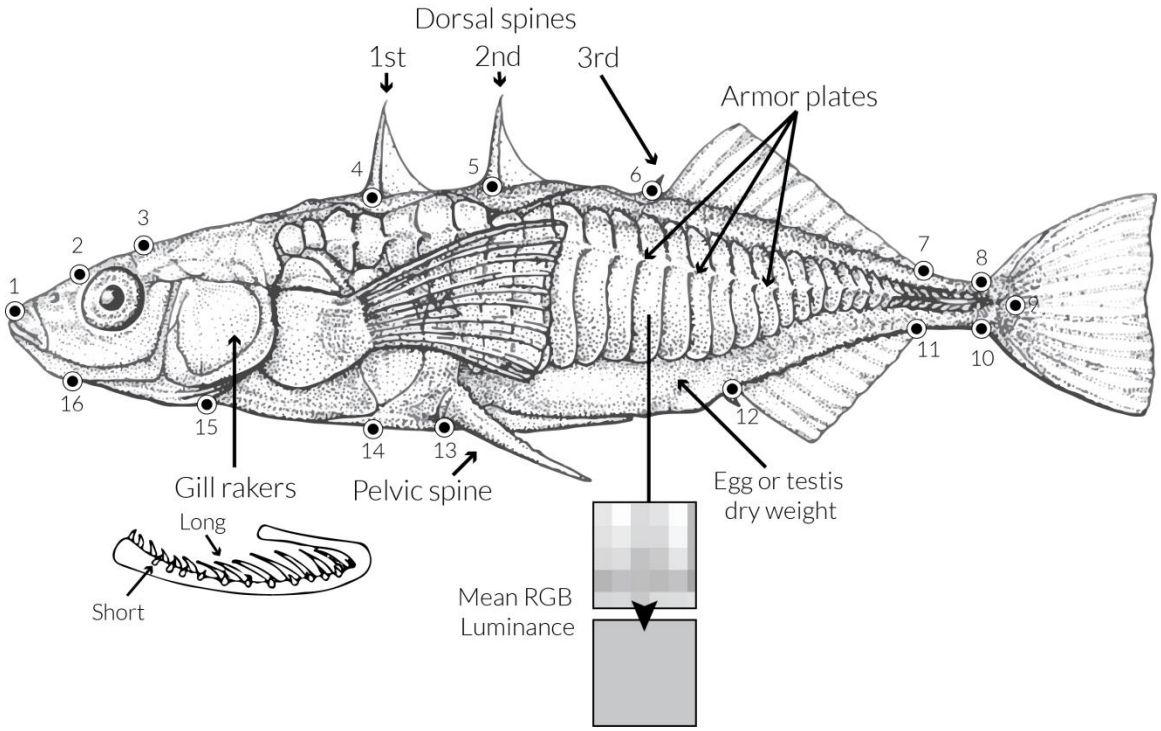

Figure 3.1| Morphological traits of the threespine stickleback measured to assess trophic differentiation. Numbered points indicate geometric landmarks used for morphometric analyses, described in Sharpe et al. (2008). Points are as follows: (1) anterior tip of the upper jaw; (2) dorsal outlier at anterior edge of eye orbit; (3) dorsal outlier at posterior edge of eye orbit; (4-6) anterior insertions of 1st-3rd dorsal spines; (7) intersection of dorsal fin ray on dorsal midline; (8) origin of caudal fin membrane on dorsal midline; (9) caudal border of hypural plate at midline; (10) origin of caual fin membrane on ventral midline; (11) insertion of anal fin membrane on ventral midline; (12) anterior insertion of the first anal fin ray; (13) anterior insertion of pelvic spine; (14) intersection of ventral outlier with first dorsal spine; (15) intersection of operculum with ectocoracoid; (16) posterior edge of angular bone. Mean RGB luminance was determined by calculating the average luminance (grey value) at each pixel in a 1cm^2 region of the flank centred on the cloaca. Standard length was calculated as the distance between points 1 and 9, and body depth was calculated as the distance between point 4 and 14. Stickleback artwork is reproduced with permission from Bell and Aguirre (2013).

3.2.4 Isotopic ratios

We quantified isotopic ratios following the method described in Reimchen *et al.* (2008). We began by dissecting out the dorsal muscle from each individual, drying it in a desiccator, and weighing out exactly 1.00 mg of dried tissue from each individual using a microbalance. We placed each unit of weighed tissue into an individual nickel capsule. The capsules were placed in a 96-well plate, and shipped to the UC Davis Stable Isotope Facility for Carbon-13/Carbon 12 and Nitrogen-15/Nitrogen-14 ratio quantification. We did not obtain environmental samples for calibration of the isotopic ratios, and instead focused on relative isotopic differences between the whites and commons. We compared differences in Carbon and Nitrogen ratios separately using a Kruskal-Wallace test and together using a MANOVA. For both analyses, we included sampling location as a covariate to control for geographic variance in isotopic abundances.

3.2.5 Assessing divergence in sex-biased genes

Identifying highly diverged regions of the genome

We identified highly differentiated regions of the genome between white and common stickleback using an F_{ST} outlier approach. To do this, we compiled the SNPs derived from GBS genotyping in Chapter 1. We calculated Weir and Cockerham's unbiased estimator of F_{ST} (Weir & Cockerham 1984) for each SNP using the function *wc.pop* in *vcftools* (Danecek *et al.* 2011). Our previous work on this system discovered that *common* stickleback in Nova Scotia group into two geographic clades – one located on the Nova Scotia mainland and another in the Bras d'Or region to

the northeast (Chapter 1). Thus, we separately computed SNP-wise divergence between whites versus mainland commons and whites versus Bras d'Or commons. We also computed F_{ST} values between mainland commons versus Bras d'Or commons. Differences between these two latter groups appear to be merely geographical (see Chapter 1), and they can thus serve as a control for regions of the genome that tend to exhibit elevated differences (e.g. regions of low recombination), but are unrelated to species differences (Marques *et al.* 2016). For simplicity, we only used SNPs from the 2014 data set.

To identify specific regions of the genome with unusually elevated F_{ST}, we first classified SNPs as outliers if they fell in the top 99th percentile of the total F_{ST} distribution of each comparison. We then separated the genome into non-overlapping 75 kilobase-pair windows, and assigned all SNPs into the resulting bins. To find windows with unusually high numbers of outliers, we performed a permutation test. In each iteration of the permutation, each SNP (outlier and non-outlier) was randomly re-assigned to a window, followed by a computation of the average number of outliers in each window. We performed 10000 permutations, resulting in a null distribution for the number of outliers in a window expected by chance. We then compared the observed number of outliers in each window to this distribution. If a window exceeded the 95th percentile of the permuted distribution, we classified it as an "outlier window".

The female sex chromosome was analyzed separately from the autosomes. Stickleback have chromosomal sex determination, with males as the heterogametic sex (Peichel *et al.* 2004). The female sex chromosome (chromosome 19) is included in the reference genome, but the male sex chromosome (the "Y" morph of chromosome 19) is not. Thus, we only examined divergence on female copies of chromosome 19 by first filtering out all the male individuals in our dataset. We then computed F_{ST} (as above) for all SNPs on chromosome 19.

To determine the windows involved in species differences between white and common stickleback, we adopted the following classification scheme. If a window was an outlier in the white vs. mainland common or white vs. Bras d'Or common comparisons (interspecific comparisons) and not an outlier in the Bras d'Or vs. mainland common comparison (an intraspecific geographic comparison), we considered it to be a "species difference" window. We discarded all outlier windows that did not fit this criterion.

Divergence and sex-biased expression

To examine sex biased expression, we used the gene annotations provided by Glazer *et al.* (2015) to create a list of ENSEMBL gene identifiers found in each 75kb genomic window (both outlier and non-outlier). For each ENSEBML gene identifier, we found the associated level of sex-biased expression reported in the comprehensive dataset of Leder *et al.* (2015). These data are in the form of standardized regression coefficients, with negative values indicating female bias and positive values representing male bias. For simplicity, we classified each gene as either male or female biased (negative or positive coefficient), and took the absolute value of the coefficient. This resulted in a bias score of 0–1.75 for each gene, along with a "male" or "female" categorization. We compared the difference in male and female bias between outlier and non-outlier windows using a Kruskal-Wallis test via the *kruskal.test* function in *R*.

3.3 Results

3.3.1 Trophic differentiation

Morphology

Based on our panel of traits, the primary morphological differences between white and common stickleback are paleness of the body and body size *per se*. In spite of a large sample size, we did not detect any body shape differences between white and common stickleback (Figure 3.2, Table 3.1). This was true for every principal component we examined (Table 3.1).

In contrast, before accounting for body size, we found that whites and commons differed in a number of other morphological traits (Table 3.1, "no covariate"). However, many of these traits (egg number, body depth, etc.) are-known to scale with body size in stickleback, and white and common stickleback differ in body size (Chapter 1). After controlling for body size via ANCOVA, the only trait that remained significantly different was a slight reduction in length of pelvic spine in the white stickleback (Table 3.1, "standard length as covariate"). That said, there is probably no reason to expect body lightness (RGB Luminance) to scale with body size, so applying a correction to that particular trait is questionable, although we include it for completeness. Finally, controlling for sex did not change the statistical significance of any of our findings (Appendix Table B.2).

Isotopic ratios

After controlling for geographic variation, we found no difference between white and common stickleback in the individual ratio of Carbon12/13 or Nitrogen isotopes14/15 (Figure 3.3, Nitrogen-15: Kruskal-Wallis test $\chi^2_1 = 1.5363$, $p = 0.2152$; Carbon-13: Kruskal-Wallis $\chi^2_1 = 0.60412$, $p = 0.437$). This result also held when Carbon and Nitrogen values were analyzed together via MANOVA (Pillai's Trace = 0.0443, $F_{2,110} = 2.55$, $p = 0.082$). Compared to the other sites, individuals of both species from Canal Lake and Little Narrows had unusual Carbon-12:13 and Nitrogen-14:15 signatures, perhaps indicating unique food sources for fish at these sites (Appendix Figure B.1).

3.3.2 Divergence in sex-biased genes

Highly divergent genomic regions

Our genome scan for F_{ST} outlined windows resulted in a total of 72 "species difference" outlier windows (Appendix Table B.1). These windows were distributed across nearly every chromosome (Figure 3.4), indicating that the genetic architecture of white-common divergence is polygenic. Most of the outlier windows contained tight clusters of SNPs that greatly exceeded the average F_{ST} of nearby SNPs (Figure 3.4, red points vs. black lines).

Genomic windows with the most F_{ST} outlier windows were on chromosomes 7, 9, 11, and 15. The most divergent of these was the gene-rich window on chromosome 11, with an average F_{ST} of over 0.7 for both white vs. common comparisons (Appendix Table B.1).

The chromosomes with the greatest number of outlier windows were chromosomes 4 and 7, which are known to harbor many regions and QTLs involved in adaptation to various environments in other stickleback (freshwater, benthic/limnetic, lake/stream, etc.). These chromosomes are also large in size, and have extensive regions of low recombination (Roesti *et al.* 2013). There was no obvious increase in average white-common F_{ST} or enrichment of white-common outlier windows on the female sex chromosome (Figure 3.4, chromosome XIX).

Divergence and sex-biased expression

The outlier windows contained 106 of 29,207 total annotated genes (Table S1). Compared to genes in non-outlier windows, genes in outlier windows were significantly more male biased according to measurements of Leder et al. (2015) (Figure 5; Kruskal-Wallis test: $\chi^2_1 = 8.68$ p = 0.0032;). In contrast, there was no significant enrichment for female biased genes in outlier regions (Figure 3.5, Kruskal-Wallis test: $\chi^2_1 = 0.17$, p = 0.67). The most male-biased gene in any outlier window was *IGFBP4* (Table S1, insulin like growth factor binding protein 4), a member of a key pathway involved in body size regulation in animals (Hyun 2013).

Table 3.1| Statistical summaries of morphological trait distributions in white and common stickleback. The mean and standard deviation (sd) are reported for each type along with sample size (n). Cohen's D is the difference in means (white minus common) in units of the pooled standard deviation. Tests of statistical significance either took the form of an ANOVA (F test), or a chi-squared test following an analysis of deviance (D statistic). The "no covariate" and "standard length as covariate" columns refer to tests of significance without and with body size correction. All p-values were corrected for multiple comparisons via the false discovery rate method.

			Common		White			No covariate		Standard length as covariate	
Trait	**Function**	**Unit**	**Mean (sd)**	**n**	**Mean (sd)**	**n**	**Cohen's D**	**Test statistic**	**p-value (FDR)**	**Test statistic**	**p-value (FDR)**
Body depth	Trophic	cm	1.23 (0.12)	161	0.95 (0.12)	73	-2.29	F1,232 = 263.96	**<0.00001**	F1,231 = 0.52	0.67903
Egg number	Sexual	eggs	86.28 (25.43)	36	50.52 (17.46)	23	-1.58	D1,57 = 260.3879	**<0.00001**	D1,54 = 0.6548	0.65222
Standard length	Trophic, sexual	cm	4.86 (0.84)	230	3.76 (0.58)	132	-1.45	F1,360 = 175.89	**<0.00001**	-	-
Pelvic spine	Trophic	cm	0.83 (0.13)	157	0.68 (0.12)	72	-1.14	F1,227 = 63.92	**<0.00001**	F1,226 = 8.75	**0.01515**
Testis weight	Sexual	mg	0.78 (0.24)	53	0.55 (0.13)	31	-1.12	F1,82 = 24.72	**0.00002**	F1,81 = 1.61	0.40852
2nd dorsal spine	Trophic	cm	0.56 (0.14)	226	0.49 (0.09)	131	-0.55	F1,355 = 25.35	**0.00001**	F1,353 = 1.25	0.46798
1st dorsal spine	Trophic	cm	0.51 (0.16)	226	0.44 (0.12)	131	-0.52	F1,355 = 22.03	**0.00002**	F1,353 = 0.4	0.69933
Body lightness	Sexual?	Intensity	474.81 (38.91)	166	491.7 (49.99)	73	0.4	F1,237 = 7.98	**0.02090**	F1,229 = 1.54	0.40990
Long gill rakers	Trophic	Rakers	20.26 (1.5)	70	20.83 (1.42)	59	0.39	D1,127 = 0.5125	0.67903	D1,125 = 0.1489	0.78885
Armor plate count	Trophic	Plates	31.41 (1.01)	70	31.02 (1.12)	59	-0.37	D1,127 = 0.1619	0.78885	D1,125 = 0.0168	0.91421
Egg diameter	Sexual	mm	1.23 (0.28)	36	1.32 (0.17)	23	0.35	F1,57 = 1.76	0.40852	F1,54 = 6.07	0.06415
Short gill rakers	Trophic	Rakers	15.14 (1.09)	70	15.44 (0.93)	59	0.29	D1,127 = 0.1858	0.78885	D1,125 = 0.0982	0.79928
Egg weight	Sexual	mg	2.83 (1.39)	36	2.53 (0.57)	23	-0.26	F1,57 = 0.93	0.55980	F1,54 = 1.67	0.40852
3rd dorsal spine	Trophic	cm	0.16 (0.06)	160	0.15 (0.05)	72	-0.23	F1,230 = 2.65	0.27745	F1,229 = 0.45	0.68454
Body shape PC2	Trophic	-	0 (0.01)	122	0 (0.01)	70	0.22	F1,190 = 2.06	0.36800	F1,189 = 0.75	0.62229
Body shape PC3	?	-	0 (0.02)	122	0 (0.01)	70	0.19	F1,190 = 1.64	0.40852	F1,189 = 0.18	0.78885
Body shape PC6	?	-	0 (0.01)	122	0 (0.01)	70	-0.12	F1,190 = 0.59	0.67105	F1,189 = 0.11	0.79814
Body shape PC5	?	-	0 (0.02)	122	0 (0.01)	70	0.07	F1,190 = 0.2	0.78885	F1,189 = 1.76	0.40852
Body shape PC4	?	-	0 (0.01)	122	0 (0.01)	70	0	F1,190 = 0	0.97495	F1,189 = 2.42	0.30630

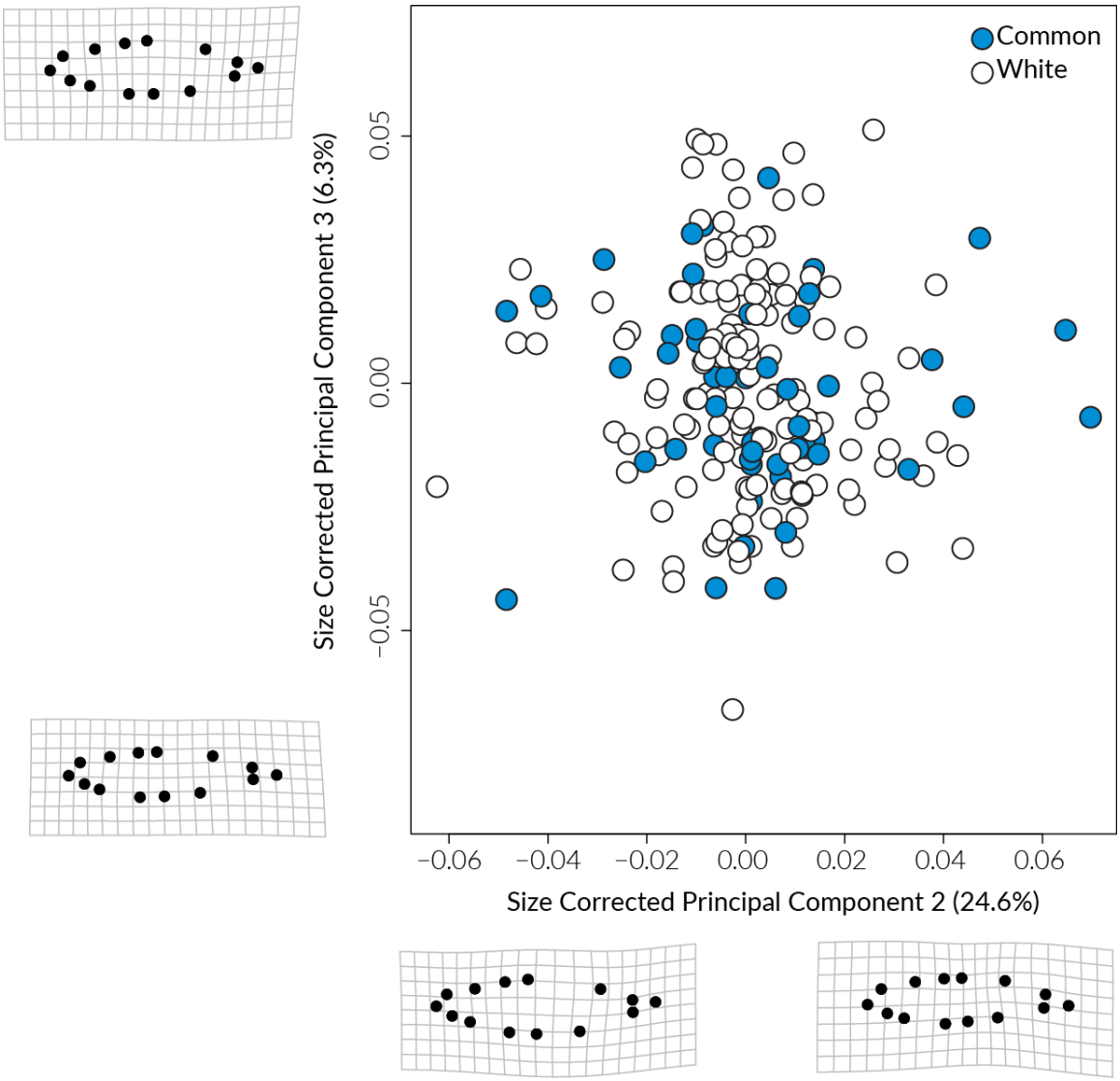

Figure 3.2| Principal components analysis of morphometric landmark positions in white and common stickleback. Percentages represent the amount of total variation explained by each principal component (PC1 represents "bending" of specimens and was omitted). Deformation grids below each axis depict the projected body shape of individuals at the minimum and maximum values of each principal component, exaggerated by a factor of 2 for visualization. Principal components data were size corrected by regressing each on the generalized Procrustes analysis scaling factor, and extracting the resulting residual.

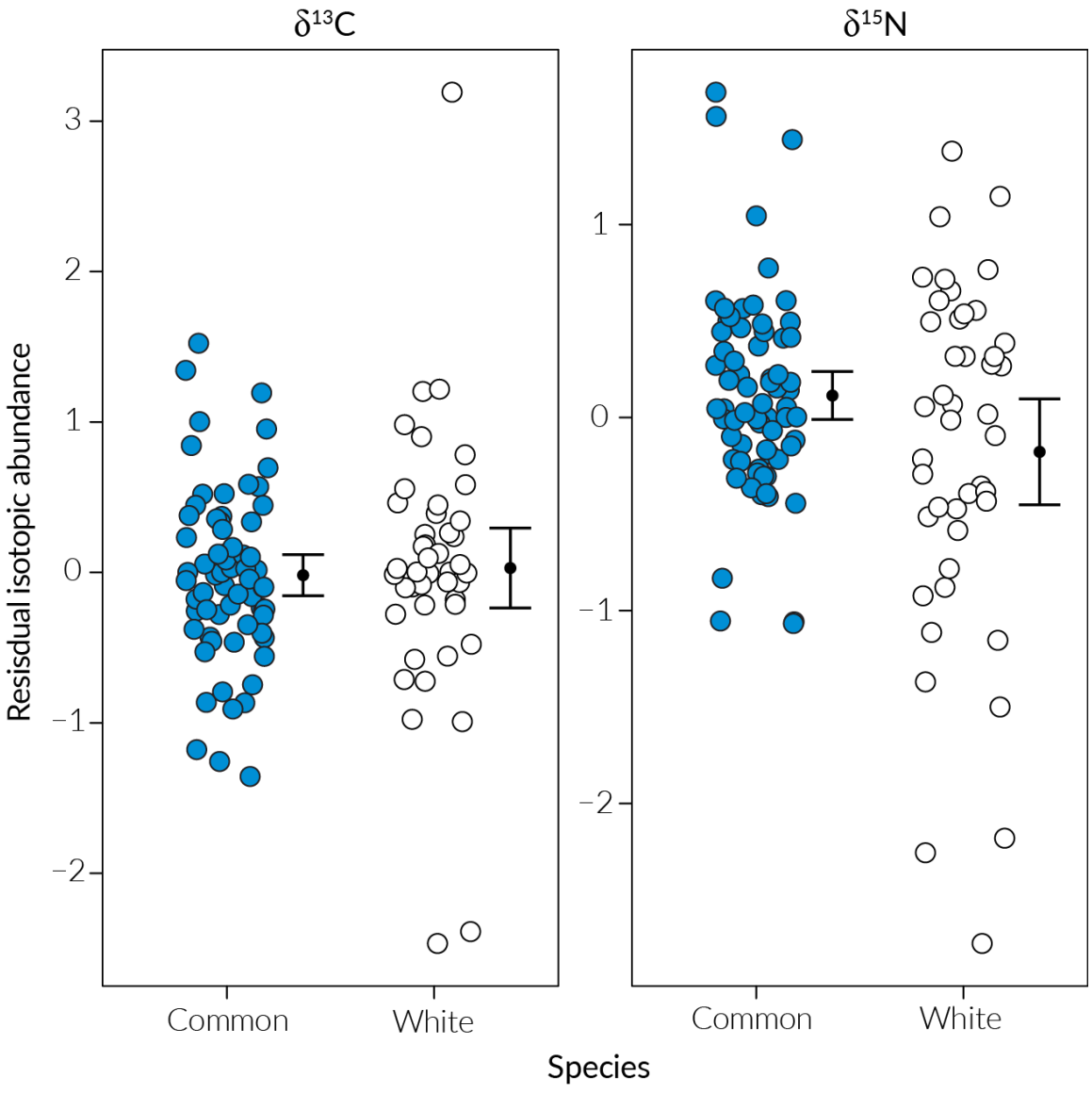

Figure 3.3| ^{13}C and ^{15}N stable isotope ratios in white and common stickleback. Black points and error bars represent means and 95% confidence intervals respectively. Isotopic abundances were corrected for geographic variation by subtracting the mean abundance of each population (irrespective of species) from each individual abundance measure. Uncorrected values are provided in Appendix Figure B.1.

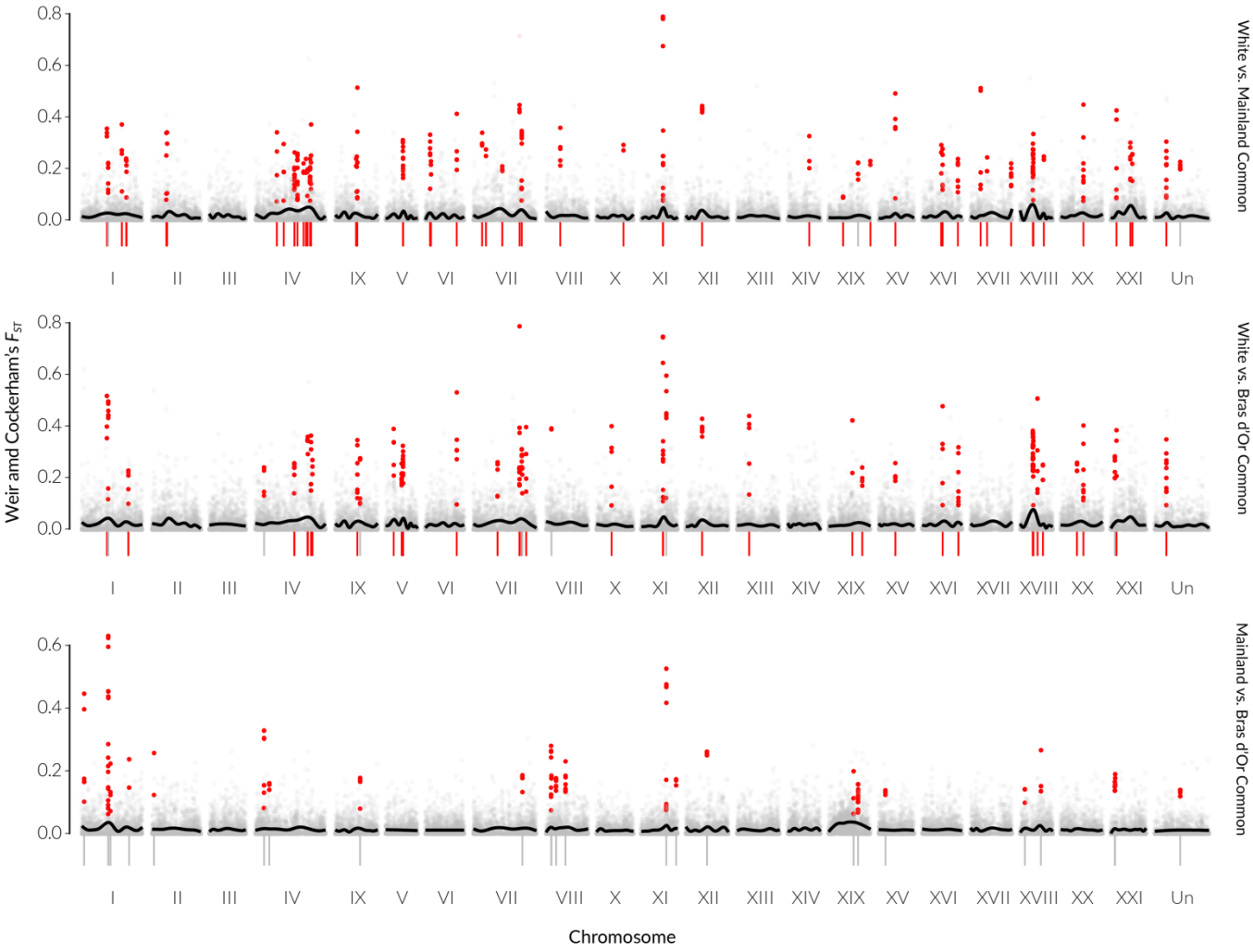

Figure 3.4| The genomic distribution of SNP F_{ST} values between three pairs of populations of stickleback. Roman numerals below plots indicate individual chromosomes ("Un" represents unassembled scaffolds). Grey points represent non-outlier SNPs, and red points represent outlier (>95th percentile of F_{ST} distribution) SNPs in significant outlier windows. Rectangles below the x-axis indicate the presence of a 75kb window containing more outlier SNPs than expected by chance. Grey rectangles represent outlier windows that occurred in common-common and common-white comparisons, whereas red rectangles represent those only occurring white-common comparisons.

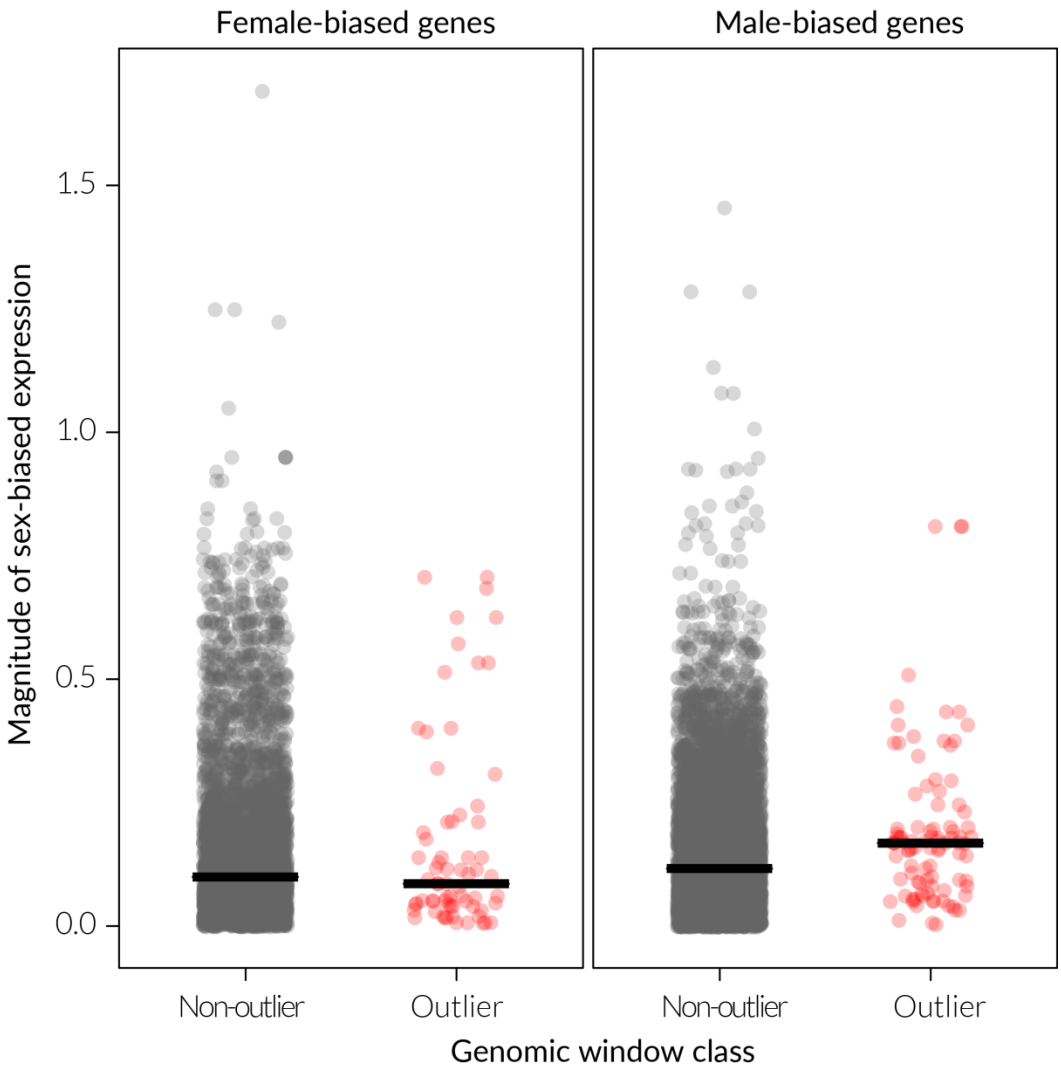

Figure 3.5| The magnitude of male and female sex biased expression of genes found in non-outlier (grey) and outlier (red) windows. Sex biased expression values correspond to normalized expression coefficients reported in Leder et al. (2015). Black lines represent median values.

3.4 Discussion

Studying recently-diverged species is the key to understanding the initial barriers that cause reproductive isolation. In this study, we examined trophic and sexual divergence between common marine stickleback and recently-evolved white stickleback. We defined "trophic" traits as those involved in trophic interactions (feeding and predation) and "sexual" traits as those that differ between the sexes and are involved in mating or reproduction.

The key trophic trait separating white and common sticklebacks was body size. However, apart from a slight difference in pelvic spine length, white stickleback and common stickleback are not differentiated in any of the other classic morphological traits associated with trophic differentiation in stickleback. White and common stickleback also did not differ in Nitrogen or Carbon isotopic ratios. The morphological and isotopic data thus in indicate weak evidence for trophic differentiation between white and common stickleback.

Genomic divergence between white and common stickleback was limited to a small number of regions. However, within these regions F_{ST} was unusually high, often exceeding 60 times the genomic average. The annotated genes within these highly diverged regions were also significantly more likely to include genes with male biased expression, suggesting that divergence in sexual traits, as defined here, plays a key role in divergence between white and common stickleback.

3.4.1 Trophic or sexual divergence?

Our results suggest that the divergence between white and common stickleback was likely not driven by the evolution of strong trophic differences. This is somewhat surprising, as every

other described threespine stickleback morph show some trophic differentiation (McKinnon & Rundle 2002). However, the context of divergence between white and common stickleback is very different from previous studied stickleback systems. For one, whites and commons appear to have diverged while remaining in the marine environment, whereas all other stickleback ecotypes evolved by initially colonizing post-glacial freshwater bodies (McKinnon & Rundle 2002). This means there was limited scope for an obvious and sudden ecological opportunity, a colonization bottleneck, and for periods of allopatry . Attenuation of any of these characteristics is known to decrease the likelihood of ecological speciation with gene flow (Flaxman *et al.* 2013; Feder *et al.* 2013), and thus may have precluded the evolution of traditional trophic differences between white and common stickleback.

That said, white and common stickleback do differ significantly in overall body size – a trait linked to trophic differences in other stickleback (McKinnion & Rundle 2002). Could this trait represent the key trophic trait separating white and common sticklebacks? If the body size differences between white and common stickleback resulted in strong trophic differences, we would also expect isotopic differences between the types (Arnegard et al. 2014). We did not see such a difference, although an analysis of stomach contents (outside of the breeding season) might help uncover differences in prey size (vs. prey trophic level). Finally, in some stickleback populations, females also use body size *per se* as a mating cue, leaving open the possibility that divergence in body size may represent both a trophic and sexual difference.

That said, if stickleback speciation without trophic differentiation is possible, why has it not been described elsewhere? One possibility is that it has indeed occurred, but has yet to have been discovered – perhaps a symptom of stickleback research focusing largely on trophic differentiation and the common assumption that the marine form represents a homogenous "living ancestral"

population (McKinnon 2002; Jones *et al.* 2012). Another idea is that speciation via sexual divergence is rare in stickleback, but conditions in Nova Scotia are particularly favorable for it. For example, perhaps the particularly species composition of filamentous algae beds in Nova Scotia represents a unique "nesting niche", which white stickleback have evolved to exploit. Blouw and colleagues suggested that the filamentous algae allow male white stickleback to to eschew parental care by taking advantage of the oxygenation and crypsis from predators offered by the filamentous algae.It is possible that this change in parental care strategy was the key trait leading to divergence (the nuptial color change perhaps being secondary) (Blouw & Blouw 1996). This idea is consistent with the fact that in spite of no physical barriers to dispersal, white stickleback appear to be restricted to Nova Scotia (Blouw & Hagen 1990).

Together, our phenotypic and isotopic results suggest that divergence in trophic niche is not necessarily the first step in stickleback speciation. Instead, it appears that divergence in sexual characters can provide equally strong, early barriers to gene flow. The view that divergence in sexual traits can be a main driver of speciation at one point prevalent in the literature, although its popularity has waned in recent years (Maan & Seehausen 2011a). This is perhaps due to the growing view that ecological differentiation is an essential ingredient for speciation, and particularly for speciation with gene flow (Feder *et al.* 2012). Indeed, in some models of speciation – such as speciation via reinforcement – divergence in sexual characters only proceeds after ecological or intrinsic isolation have evolved (Liou & Price 1994). The white stickleback appears to provide an example counter to this view.

3.4.2 Barriers to reproduction in the white stickleback

We have shown that the traits separating white and common stickleback are likely mostly sexual. However, the specific traits that maintain reproductive isolation between these populations remain obscure. The obvious candidates are still the sexual characters outlined by Blouw and colleagues: body size, male nuptial color, nest site preference and the loss of parental care (Blouw & Hagen 1990; Jamieson *et al.* 1992b). Various pairs of stickleback species have been found to mate assortatively based on body size (so called 'self-referential phenotype matching') (Nagel & Schluter 1998; Conte & Schluter 2013). Indeed, some authors consider body size to be an 'automatic magic-trait' for stickleback, as ecologically-based divergence in this trait can cause both extrinsic post-zygostic isolation and premating isolating in one fell swoop (Servedio & Kopp 2012; Conte & Schluter 2013). Differences in male coloration and nest site preference are also known to be linked to assortative mating in stickleback, although this requires concomitant changes in female preference (Boughman *et al.* 2005; Foster *et al.* 2008).

One possibility is that these sexual traits form a co-adapted male "syndrome", centered on the reallocation of male energy away from parental care and toward mate acquisition. The key to this change in strategy would be the shift to nesting in filamentous algae, which might have relaxed selection on the maintenance of parental care (Blouw & Blouw 1996). A loss of parental care then might favor a smaller body size (and faster maturation) -- male stickleback usually expend large amounts of stored energy during the parental phase, necessitating a large body size (Jamieson *et al.* 1992a; Smith 1999). Because males normally halt all courtship during the parental phase, this would then massively increase the number of males actively courting females in a given area, intensifying

male-male competition (Jamieson *et al.* 1992b). This increased competition might then intensify sexual selection on male secondary characters, such as nuptial color (Ritchie 2007).

3.4.3 Genomic islands?

The patterns of genomic divergence between white and common stickleback are consistent with the "genomic islands" pattern reported in many other studies. These highly diverged "islands" of high F_{ST} surrounded by a "sea" of undifferentiated loci are thought to be typical (and some argue indicative) of speciation with gene flow. Many authors argue that highly diverged loci are likely those that underlie reproductive isolation and/or local adaptation. Indeed, in this study, we assumed this was generally true and used this to our advantage – i.e. as a proxy for divergence in sexual characters. However, there is much debate around the appropriateness of so called "relative" measures of genetic differentiation, such as F_{ST}, for identifying loci involved in reproductive isolation. For example, because the denominator of F_{ST} is usually some function of overall genetic diversity (e.g. total expected heterozygosity), F_{ST} can become elevated in regions of unusually low heterozygosity, even in the absence of divergent selection between population. The most well-known example of this is a reduction in heterozygosity at sites linked to those undergoing strong purifying selection – "background selection" (Charlesworth *et al.* 1993). Could the outlier regions we identified here be the result of background selection?

Two lines of evidence suggest that this is not the case. First, if background selection has caused a given region to become an F_{ST} outlier, we would expect the region to be an outlier in both the white vs. common and mainland common vs. Bras d'Or common comparisons. And indeed, we do see outlier windows that exist in both these comparisons. However, we explicitly sought to

remove these "global" outliers specifically to account for sources of genomic variation unrelated to reproductive isolation. Secondly, in general, background selection is not predicted to generate divergence specifically in autosomal genes with male-biased expression (Charlesworth 2012).

Another approach to examining divergence is to use so called "absolute" measures of genomic divergence such as D_a or d_{xy} (Noor & Bennett 2009; Cruickshank & Hahn 2014). These measures are useful in some circumstances, but can be highly problematic in others. For example, d_{xy} is highly insensitive to allele frequency differences, particularly when a given marker is polymorphic in both populations (Cruickshank & Hahn 2014) . Instead, d_{xy} relies largely on the presence of new mutations to generate a signal (Cruickshank & Hahn 2014). This makes the application of d_{xy} to very young species of questionable utility, particularly when reduced representation data (RAD or GBS) are involved.

3.4.4 Future work

Our results suggest a number of fruitful avenues for future investigation. First, the white stickleback seems like a prime system in which to test the role of sexual and natural selection in speciation. This is a major unresolved issue in speciation research. For example, sexual selection theory predicts that female preference should be genetically correlation with male nuptial traits (Ritchie 2007) – this prediction could be readily tested in white stickleback. In additional, one could test whether speciation in the white stickleback was driven by environment dependent sexual selection – a form of ecological speciation (Schluter 2001). This could be followed up with experiments using aimed at assessing the phenotypes (behavioral and morphology) and fitness of hybrids – particularly in the presence and absence of filamentous algae.

Another line of inquiry could focus on the role of standing variation versus new mutation in speciation. The repeated ecological divergence in other stickleback systems is thought to have been caused (in part) by repeated selection on standing variation shared via gene flow (Schluter & Conte 2009). Perhaps the geographic isolation of the white stickleback from other stickleback systems has forced them to rely on new mutation instead of gene flow as a source of phenotypic novelty. This would provide a rare case in which we could compare the phenotypic and genomic consequence of speciation from standing variation versus new mutation.

3.4.5 Conclusion

In spite of over a hundred years of research, the mechanisms behind the origin of new species remain poorly understood. We still have a murky picture of which barriers are key to the initial evolution of reproductive isolation, and which evolve later. Studying recently diverged species is a promising approach to solving this problem. Here, we explored morphological, isotopic and genotypic differentiation between the recently diverged white and common stickleback from Nova Scotia. We found that white and common stickleback are only weakly diverged in morphology and isotopic abundances, though they differ in body size. Genomic differences between the two types are extremely strong in some regions of the genome (but globally weak), highly heterogeneous and biased toward genes with male specific expression. Together, these results suggest that sexual rather than trophic traits underlies divergence and reproductive isolation between white and common stickleback. This stands in contrast with the view that trophic differentiation is a necessary first step in the speciation process, particularly in fish. While more work is needed to explore these ideas, the

white stickleback will no doubt provide a fruitful system for future inquiry into the role of sexual, trophic and genomic differentiation during the early phases of speciation.

Chapter 4: Clustering of adaptive alleles is favored by gene flow in a globally distributed species

4.1 Introduction

Understanding the genetic basis of adaptation is a fundamental goal of evolutionary biology. Yet, we still know little about the myriad interacting factors that determine the number, genomic location and effect size of loci underlying adaptive traits. Recent work suggests that interactions between two core evolutionary forces, natural selection and gene flow, may profoundly shape where adaptation occurs in the genome (Noor *et al.* 2001a; Kirkpatrick & Barton 2006; Yeaman & Whitlock 2011; Nachman & Payseur 2012). When divergent selection and gene flow co-occur (hereafter 'DS-GF'), hybridization between migrant and local individuals breaks down linkage disequilibrium (LD) between sets of locally adapted alleles, impeding adaptation (Noor *et al.* 2001a; Kirkpatrick & Barton 2006; Yeaman & Whitlock 2011; Nachman & Payseur 2012; Sousa & Hey 2013). This decay of LD can be slowed if locally adapted alleles are tightly genetically linked, i.e. very close together on the same chromosome, or occurring within a region of low recombination (Rieseberg 2001; Noor *et al.* 2001a; Navarro & Barton 2003; Yeaman & Whitlock 2011). Accordingly, theory predicts that DS-GF will drive a tendency for locally adapted alleles to be tightly linked in regions of low recombination (Yeaman & Whitlock 2011).

However, this prediction has never been formally tested, as it requires replicated comparisons of the genomic distribution of adaptive alleles between populations with and without gene flow, and populations with and without divergent selection. It is also necessary to disentangle the effects of

selection and gene flow from other processes that can generate clustering of adaptive alleles. For example, linked selection – hitchhiking and background selection – is widely known to cause clustering of diverged loci, an effect that is amplified in regions of low recombination even in the absence of gene flow (Charlesworth *et al.* 1993; Cutter & Payseur 2013). In addition, variation in mutation rate across of the genome (perhaps mediated by variation in cross-over rates) can cause clustering of co-segregating alleles (Noor *et al.* 2001a; Kirkpatrick & Barton 2006; Yeaman & Whitlock 2011; Nachman & Payseur 2012; Rattray *et al.* 2015).

4.2 Outline of methods

To approach this problem, we assembled a large population genomic dataset of threespine stickleback (*Gasterosteus aculeatus*) from across the northern hemisphere (Appendix Figure C.1). Threespine stickleback are a holarctic species of fish that have evolved into a variety of unique forms over the last 10,000 years (Bell & Foster 1994; Kirkpatrick & Barton 2006; Nachman & Payseur 2012; Sousa & Hey 2013). Leveraging the wealth of data on this species, we obtained DNA sequences from databases and generated new genomic data for several populations (Rieseberg 2001; Noor *et al.* 2001a; Navarro & Barton 2003; Yeaman & Whitlock 2011). The final dataset includes genotype information from 1350 individuals from 48 unique populations, each belonging to one of seven described ecotypes: marine (including anadromous), lake, stream, benthic, limnetic, white, and Sea of Japan (Appendix Figure C.1, Appendix Table C.1). The genomic data were a mixture of Restriction Amplified Digest (RAD), Genotyping-By-Sequencing (GBS), and whole genome sequencing (Yeaman & Whitlock 2011). We used a single bioinformatic pipeline to standardize the identification of single nucleotide polymorphisms (SNPs) across all study populations (Charlesworth

et al. 1993; Cutter & Payseur 2013). We then classified all pair-wise comparisons between our 48 populations (n = 1128 comparisons) into four classes (herein "regimes") based on whether the pairs have evolved divergent or similar ecotypes ("divergent" / "parallel") and whether there has been opportunity for gene flow between populations, based on geographic distance ("gene flow" / "allopatry"). For the latter, we classified populations as having the opportunity for gene flow (now or in the recent past) if they were within 500 km of one another by great circle distance. This threshold was based on the size of gaps in geographic sampling that were inherent in the data – a cutoff of 500km divides populations into clusters in which all populations are all within ~20-50km of another population (Appendix Figure C.1). Although this factorial approach is simplified, it allowed us to broadly disentangle how the genomic distribution of divergent alleles is affected by selection and gene flow. We also explored the use of continuous measures of geographic isolation, which produced similar results (see Detailed Materials and Methods).

We identified adaptively differentiated regions of the genome by locating SNPs and 75 kilobase pair (kbp) windows exhibiting unusually high levels of genetic divergence (>=95th percentile values for F_{ST} and D_{XY}) in each pair-wise comparison. Rather than identifying specific SNPs or windows as targets of selection, this approach provides a sample of loci enriched for those subject to divergent selection (Narum & Hess 2011). For convenience, we refer to these hereafter as 'outlier SNPs' and 'outlier windows'. For each window, we also estimated mutation rates using a phylogenetic approach, and obtained estimates of gene density for each window from the ENSEMBL database.

We then carried out two separate analyses. First, we estimated the concentration of outlier windows in regions of low recombination. For each pairwise comparison we fit outlier status of windows (0 = includes no outlier SNPs, 1 = at least one outlier SNP) to their estimated rates of

recombination, while controlling for mutation rate and gene density, using logistic regression. The slopes of these regressions were then compared among the four gene flow/selection regimes. Second, we quantified clustering of outlier SNPs over the whole genome. We calculated (a) the nearest neighbor distance in centimorgans (cM) between outlier SNPs relative to nearest neighbor distance between all SNPs; and (b) the coefficient of variation of genetic distances (in cM) between outlier SNPs. Importantly, these clustering metrics control for variation in SNP density among genomic regions, and thus are not biased by differences in sequencing coverage (see section 4.4, Detailed Materials and Methods).

4.3 Results and discussion

As expected, F_{ST} outlier windows occurred most often in regions of low recombination, even between allopatric populations and between populations inhabiting similar environments (Fig 1), when mutation rate and gene density are controlled statistically. Therefore, the occurrence of outliers in regions of low recombination is a general feature of divergence in this species. However, as predicted, this tendency was most extreme in DS-GF comparisons (Figures 4.1A-C, Appendix Figure C.2; permutation test: $p = 0.0002$). The use of a continuous measure of geographic distance led to the same result (Appendix Figure C.4). Further, DS-GF comparisons did not exhibit unusually low levels of average intra-population heterozygosity (H_S) in regions of low recombination, suggesting that the tendency for outliers to occur in regions of low recombination is not an artefact of background selection (Appendix Figure C.2; permutation test: $p = 0.755$). d_{XY} outliers also showed a non-significant tendency to be occur most often in regions of low recombination (Appendix Figure C.2; permutation test: $p = 0.475$, see also Appendix Figure C.5).

This lack of significance may be because $d_{XY\,Y}$ is less sensitive than F_{ST} at detecting (a) differences in very recently diverged populations such as these, and (b) adaptation from standing variation, which is thought to be common in stickleback (Cruickshank & Hahn 2014). In aggregate, these results suggest that between populations experiencing divergent selection with gene flow, adaptation occurs disproportionally often in regions of low recombination.

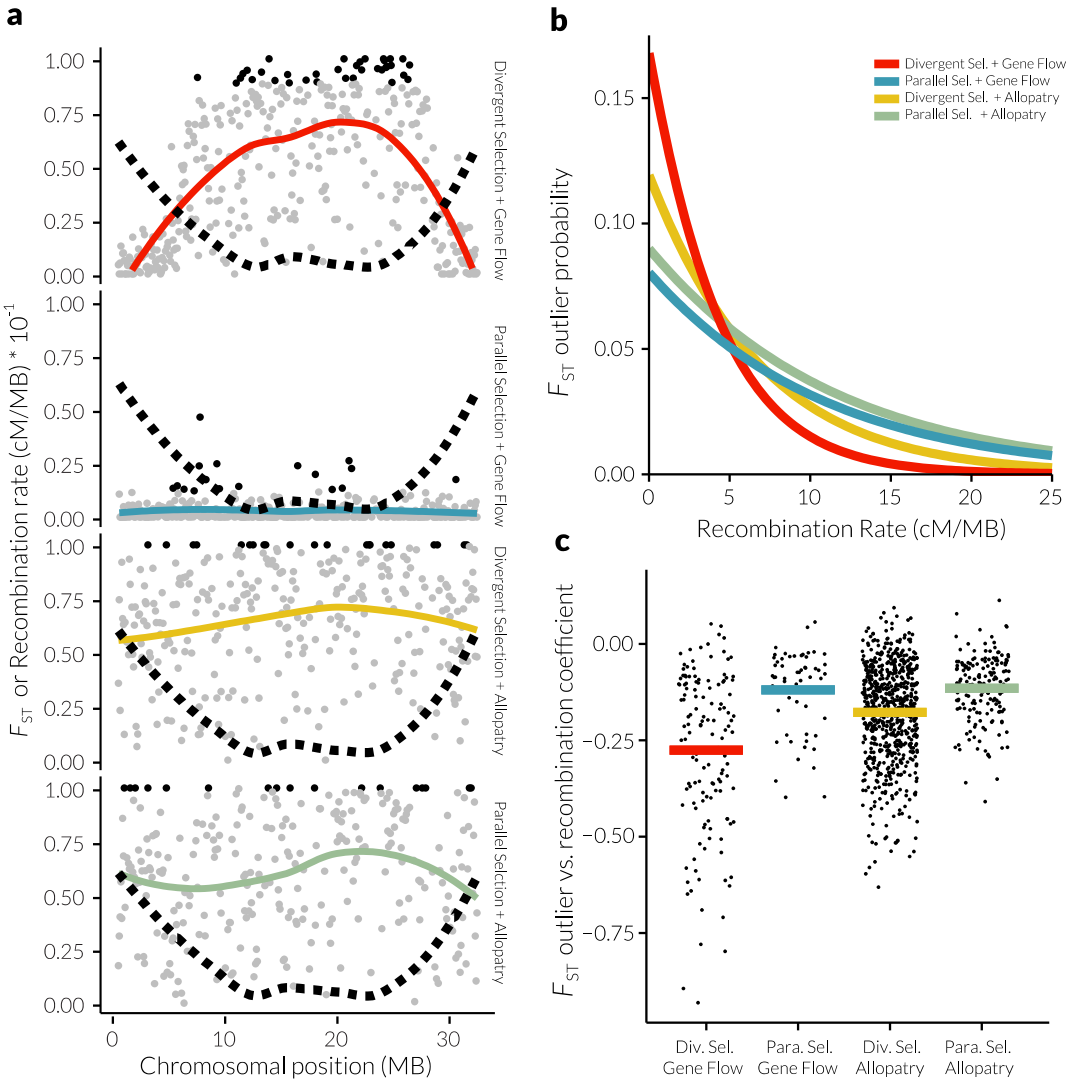

Figure 4.1 | Depiction of the relationship between gene flow, selection regime and the clustering of outliers in regions of low recombination. (a) Divergence (F_{ST}) profiles for chromosome IV from four representative pairs of stickleback populations. Each coloured line is a loess smooth of F_{ST} to show the trend with chromosomal position. Line

color indicates gene flow and selection regime (legend in Fig 1b). In each panel the first regime (red) is DS-GF. Dashed lines show the trends in rate of recombination with chromosomal position (in centimorgans per megabase, divided by 10 to achieve a common scale). From top to bottom, the pairwise comparisons are: Priest Lake (benthic, British Columbia) vs. Priest Lake (limnetic, British Columbia); Captain's Pond (marine, Nova Scotia) vs. Sky River (marine, Nova Scotia); Mariager (marine, Denmark) vs. Joes Lake (stream, British Columbia); Lake Constance (lake, Switzerland) vs. Joes Lake (lake, British Columbia). **(b)** Logistic regressions of outlier status against recombination rate, averaged over all pairwise comparisons within the four gene flow and selection regimes. Regressions are corrected for variation in mutation rate and gene density. **(c)** Individual logistic regression coefficients for each pairwise comparison (points) in each gene flow / selection regime. Colored horizontal lines indicate means. The more negative the coefficient, the more rapidly outlier status declines with increasing recombination rate. The curves shown in (b) are based on the mean values shown in (c).

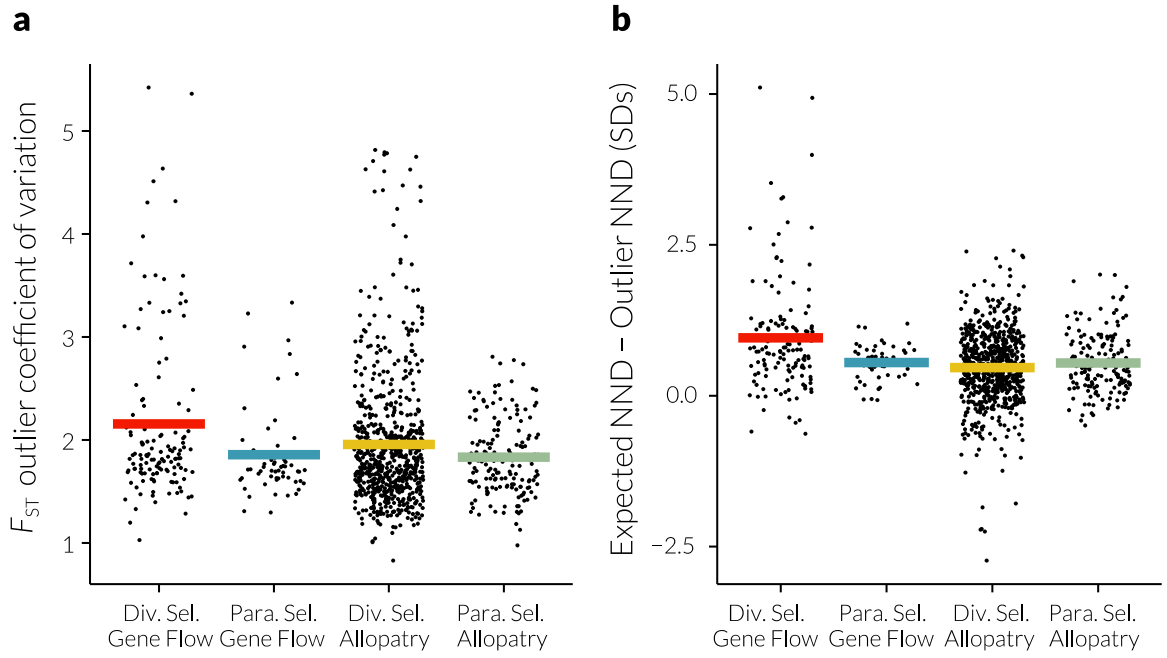

Figure 4.2 | Metrics of F_{ST} outlier clustering from across four gene flow and selection regimes. In each panel the first regime (red) is DS-GF. Each point is the mean of a cluster metric across all chromosomes for a single population comparison (n = 1128). **(a)** The coefficient of variation of genetic positions of outliers. Higher values are indicative of more clustering. **(b)** The difference between the expected average nearest-neighbor genetic distance between all SNPs and the observed mean distance between outlier SNPs, in units of standard deviations of the expected distribution. A higher value indicates that observed distance between outlier SNPs is smaller than the expected distance.

DS-GF population pairs also showed more clustering of F_{ST} outlier SNPs than population pairs in other gene flow/selection regimes. Outlier SNPs in DS-GF comparisons had substantially

shorter genetic distances between them compared to outlier SNPs from other regimes. These outliers were one standard deviation closer together in the genome on average than expected on the basis of overall SNP density (Figure 4.2B, Appendix Figure C.3, permutation test: $p < 0.0001$). Coefficients of variation of distance between F_{ST} outlier SNPs showed similar results (Fig 4.2A, Appendix Figure C.4, permutation test: $p < 0.0001$), again indicating highest levels of clustering in DS-GF comparisons. Consistent with the outlier window analysis, clusters of outlier SNP in DS-GF comparisons were entirely localized in regions of low recombination (Fig 4.1A). Although theory predicts that under DS-GF alleles should cluster even in the absence of variation in recombination rate (Yeaman & Whitlock 2011), strong linkage in regions of low recombination may allow clusters to build more easily, and between a larger possible set of alleles.

In sum, the evolution of clusters of adaptive alleles in regions of low recombination appears to result from divergent natural selection with gene flow, as predicted by theory (Noor *et al.* 2001a; Navarro & Barton 2003; Yeaman & Whitlock 2011). This is consistent with the finding that quantitative trait loci between stickleback populations evolving under DS-GF preferentially map to regions of low recombination (Arnegard *et al.* 2014; Conte *et al.* 2015, but see Noor et al. 2001b). This implies that *where* divergence can occur in the genome is constrained by gene flow, leading to more heterogeneous adaptation across the genome than would otherwise be the case. In effect, the "usable area" of the genome is smaller under DS-GF than other regimes. Interestingly, by limiting the where adaptation can occur, DS-GF may indirectly increase the probability that the same loci will be reused during parallel phenotypic evolution in general. Thus, we predict that pairs of DS-GF populations (perhaps even ones where selective pressures are different) should display unusual levels of concordance in the loci involved in divergence, and that these loci will occur in regions of low recombination. This may also explain why even allopatric populations show some clustering of

divergent alleles, since each population of an allopatric pair may nevertheless have opportunity for gene flow with other populations in its vicinity.

Other studies have found mixed empirical support for the pattern we identify here (Carneiro *et al.* 2009; Feder *et al.* 2012; Renaut *et al.* 2013; Burri *et al.* 2015; Holliday *et al.*). One reason may simply be power: our results are based on large numbers of individuals and populations, and balanced representation of population pairs experiencing different gene flow and selection regimes. Another possible explanation is that clustered genetic architectures may require long temporal scales and/or recurrent bouts of gene flow in order to develop. Although most stickleback populations are less than 10, 000 years old, stickleback have repeatedly cycled between adapting to freshwater during interglacial periods, followed by extinction of these populations during glacial periods (Schluter & Conte 2009; Jones *et al.* 2012). However, gene flow between freshwater and marine populations has likely allowed ancient freshwater haplotypes to persist in marine populations throughout this process. This recurrent process, coupled with large effective population sizes of marine stickleback, may have increased the opportunity for clustered "cassettes" of divergently selected alleles to arise and be maintained (Schluter & Conte 2009; Jones *et al.* 2012).

There are two important caveats to the work we present here. One is that we were not able to include explicit estimates of the strength of divergent selection into our models. If strong selection results in higher concentrations of outliers in regions of low recombination (e.g. due to stronger hitchhiking effects) (Charlesworth *et al.* 1993), this may help explain some of the increased clustering we see in the divergent-allopatry regime (Figure 4.1, yellow lines). Secondly, we make the assumption that recombination rates are invariant among populations. Recombination-altering structural variants, such as inversions and translocations, are known to segregate among populations included in the analysis (Jones *et al.* 2012). However, assuming that such variants decrease recombination rate

and increase divergence in random genomic locations, this would merely add noise to our data. Thus, our assumption of a homogenous genetic map should not bias our results in any particular direction.

In sum our findings suggest that the co-localization of adaptive alleles in the genome is a product of both extrinsic forces (selection and gene flow) and properties of the genome itself (recombination rate). The limitation to regions of low recombination further suggests that gene flow places specific limits on the usable area of the genome, which represents a previously unappreciated constraint on adaptive evolution. This finding also holds keys implications for our ability to predict the outcome of adaptive evolution.

4.4 Detailed materials and methods

4.4.1 Github repository

The code used to generate our dataset and perform the analyses described here is available on Github at https://github.com/ksamuk/gene_flow_linkage. Additional raw data is also hosted on Dryad (Dryad accession, to be made available).

4.4.2 Data sources

The stickleback population genomic datasets used in this study came from two sources: online databases, and new data from two of the authors. During the period from May to July 2014, we periodically searched the Short Read Archive (SRA), the European Nucleotide Archive (ENA)

and the Databank of Japan Sequence Read Archive (DRA) for "threespine/three-spined/threespine/three-spine stickleback", "stickleback", "Gasterosteous aculeatus". We also searched for stickleback population genetic studies on Google Scholar using the same terms as above, with the inclusion of "genomic", "genome scan", "population genetic", and "genetics", and examined them for SRA/ENA/DRA accession numbers. All information for populations included in the study is shown in Appendix Table C.1.

The two unpublished datasets – benthic/limnetic populations from British Columbia and white/marine populations from Nova Scotia – were prepared by two of the authors using the GBS method of Elshire et al. (Weir & Cockerham 1984; Elshire *et al.* 2011). The collection locations are listed in Table S1. The resultant libraries were sequenced at the UBC Biodiversity Sequencing Centre on an Illumina Hi-Seq 2000. These datasets will be made available on the SRA (accession # to be made available).

4.4.3 Variant identification and processing

We identified variants using a standard, reference-based bioinformatics pipeline (see Github code repository for details). After demultiplexing, we used Trimmomatic (Nei & Miller 1990; Cruickshank & Hahn 2014; Bolger *et al.* 2014) v0.32 to filter low quality sequences and adapter contamination. We then aligned reads to the stickleback reference genome (McKinnon & Rundle 2002; Jones *et al.* 2012; Catchen *et al.* 2013) using BWA v0.7.10 (Li & Durbin 2009; Vavrek), followed by realignment with STAMPY v1.0.23 (Lunter & Goodson 2011; Roesti *et al.* 2013). We then followed the GATK v3.3.0 (Yang & Bielawski 2000; Noor *et al.* 2001b; McKenna *et al.* 2010) best practices workflow (DePristo *et al.* 2011) except that we skipped the MarkDuplicates step when

reads were derived from reduced representation libraries (RAD and GBS). We realigned reads around indels using RealignTargetCreator, and IndelRealigner, identified variants in individuals using the HaplotypeCaller, and each dataset using GenotypeGVCFs. The results were sent to a VCF file containing *all* genotyped sites (variant and invariant), and converted to tabular format. All datasets were combined for processing.

4.4.4 Calculation of divergence metrics

Our final dataset included individuals from 48 unique populations. As there was no *a priori* reason to select only a subset pairs of populations in the analysis, we instead performed all possible pairwise comparisons. We employ an unbiased significance testing method to overcome the issue of redundancy of pairs (see permutation test).

For each of the 1128 pairwise comparisons, we calculated two divergence metrics: Weir and Cockerham's F_{ST} (Weir & Cockerham 1984) and d_{XY} (Nei & Miller 1990; Cruickshank & Hahn 2014). We calculated F_{ST} at two scales: first, at each individual shared SNP; and second, averaged across 75 kilobase pair (kbp) windows. Window-averaged F_{ST} values were calculated by dividing the sum of the numerators of all SNP-wise F_{ST} estimates within a given window by the sum of their denominators. For all SNPs, we required a minor allele frequency of > 0.05, and coverage (after GATKs best practices filters) in at least 5 individuals per population. We calculated d_{XY} in 75-kbp windows, including all shared variant and invariant sites in the window (Cruickshank and Hahn 2014). We required d_{XY} windows to contain more than 500 sequenced sites, because we found that the variance in d_{XY} greatly increases below this threshold. After calculating these metrics, we classified SNPs and windows exhibiting extreme values as 'outliers', defined as those in the 95th

percentile or higher of F_{ST} or d_{XY}. Note, only d_{XY} window 'outliers' were used because individual site d_{XY} scores are uninformative. All calculations were performed using custom Perl and R scripts (see code repository).

4.4.5 Classification of populations

We classified each pairwise comparison of populations into one of four classes according to whether they inhabit areas with divergent ("divergent") or similar ("parallel") ecology and whether they had the opportunity for gene flow ("gene flow") or not ("allopatric"). Populations were considered "divergent" if they inhabited different ecosystems or ecological niches or had been directly identified by previous authors as ecologically divergent: benthic and limnetic lake stickleback; lake and stream stickleback; marine and freshwater stickleback; Sea of Japan and Pacific marine stickleback; white and Atlantic marine stickleback (McKinnon & Rundle 2002; Catchen *et al.* 2013). Populations were considered to have the opportunity for gene flow if they were within 500 km of one another. We calculated geographic distance (great circle distance) between all pairs of populations using the function *earth.dist* from the R package *fossil* (Vavrek 2010).

4.4.6 Addition of genomic variables

We measured three genomic variables in each 75-kbp window in the divergence dataset with: recombination rate, mutation rate and gene density. Recombination rates (cM/MB) were obtained from a recently published high-density genetic map (Roesti *et al.* 2013). Where windows overlapped

regions with different estimates of recombination rate, we assigned them an average of the two rates weighted by the degree of overlap.

We obtained estimates of mutation rate by estimating the synonymous substitution rate (*dS*) in a phylogenetic framework. For neutral sites, *dS* is an estimator of the primary mutation rate (Yang & Bielawski 2000; Noor *et al.* 2001b). To do this, we used the R (version 3.2.2) (Charlesworth *et al.* 1993; Elshire *et al.* 2011; Team 2015) package *biomaRt* (Durinck *et al.* 2009; Bolger *et al.* 2014) to obtain a list of all annotated *G. aculeatus* coding DNA sequences (CDS) from ENSEMBL (Jones *et al.* 2012; Cunningham *et al.* 2015). For each *G. aculeatus* CDS, we queried ENSEMBL for all homologous CDS from three other fish species: *Xiphophorous maculatus*, *Poecilia formosa*, and *Oreochromis niloticus*. These species all have identical estimated divergence times from *G. aculeatus* (150 MYA) (Li & Durbin 2009; Hedges *et al.* 2015). We aligned each set of homologous coding sequences using PRANK (Lunter & Goodson 2011; Löytynoja 2013), and analyzed the output using PAML (Yang 2007; McKenna *et al.* 2010) (Branch model 2) to estimate *dS* trees. We excluded trees with fewer than three species, in order to ensure that lineage-specific artefacts did not bias dS estimates. We also excluded any individual branches where *dS* exceeded 5 standard deviations of the distribution of the *dS* values from all branches of every tree (values exceeding this threshold were categorically the result of bad alignments). After filtering *dS* trees, we used the R package *ape* to calculate the mean pairwise branch distance between *G. aculeatus* and each other species in the tree. Because the other three species all have identical divergence times from *G. aculeatus,* this results in a single normalized value of *dS* for each coding sequence. After obtaining all the mutation rate estimates, we assigned them to 75 kbp windows in the divergence datasets by averaging the *dS* estimates for genes in each window (if any), weighted by the degree of overlap for each gene.

Estimates of gene density (number of genes overlapping the window) were calculated by querying ENSEMBL for the physical position of all genes in the stickleback genome using *biomaRt*. We then wrote a custom R script (see Github repository) to count the number of genes in each 75-kbp window along the reference genome.

4.4.7 Tendency for adaptive divergence in regions of low recombination

To test the hypothesis that adaptation with gene flow favors divergence in regions of low recombination, we employed a linear modeling approach. Using the 75-kbp windows as data points, we fit a logistic regression model to each comparison dataset using the following form: Outlier status = Recombination rate + mutation rate + gene density, where outlier status is 0 if no outliers are present, and 1 if at least one outlier was present. We performed separate model fits for F_{ST} and D_{XY} outliers.

We fit these models in R (version 3.2.2) (DePristo *et al.* 2011; R Core Team 2015) using the generalized linear model function *glm*. Prior to model fitting, we filtered out pairwise population comparisons with fewer than 100 75-kbp windows represented to ensure convergence of the linear models. To assess statistical significance of the model fits, we extracted the regression coefficient for the recombination rate term from each model, representing the slope of the relationship between outlier occurrence and recombination rate. The steepness of the slope coefficients estimates the tendency for outliers to occur in regions of low recombination, controlling for the effects of mutation rate and gene density.

We then performed a permutation test to assess whether the slopes differed significantly between populations differing in divergent selection and gene flow. To do this, we randomly

shuffled regime assignments of all the populations and estimated the mean low recombination outlier tendency (regression coefficient from above) for each regime in 10,000 permutations. This generated a null distribution of mean slopes for each regime, accounting for sample size differences between categories (Appendix Figure C.2). We then calculated a *P* value for each empirical mean by the computing the fraction of samples in the null distribution greater than the observed value and multiplying by two.

4.4.8 Clustering vs. geographic distance and overall divergence

To ensure our results were not influenced by our discrete categorization scheme, we examined how the tendency for F_{ST} outliers to occur in regions of low-recombination varied with pairwise geographic distance and overall genetic divergence. To do this, we regressed the low recombination outlier tendency (regression coefficients from above) on geographic distance between populations. As expected, the tendency for outliers to occur in regions of low recombination increased with decreasing geographic distance, but only when populations exhibited divergent adaptation (DS-GF, DS-Allopatry) (Appendix Figure C.4 A, ANOVA, distance x selection interaction, $F_{1,920} = 10.579$, $p = 0.0011$). Further, the tendency for F_{ST} outliers to occur in regions of low-recombination was positively associated with overall genetic divergence, indicating that differences in overall divergence between gene flow/selection regimes did not influence our results (Appendix Figure C.4 B, ANOVA, overall F_{ST} x regime interaction, $F_{3,916} = 53.983$, $p < 2.2 \times 10^{-16}$). Interestingly, D_{XY} showed very similar overall patterns to F_{ST} when analyzed in this way (Appendix Figure C.5). Note the pairwise nature of the comparisons results in interdependence among

observations in these analyses, although this is unlikely to systematically bias the direction of the results.

Finally, we also explored analyses using two other metrics of gene flow: a measure of the amount of non-aquatic terrain between populations (amount of land separating a lake and the ocean); and 'swimmable distance,' a measure of distance between populations following coastlines instead of great circles. The results of these analyses were similar to those obtained using great circle distance, and we omitted them here for clarity.

4.4.9 Increased clustering of outlier loci

To test the hypothesis that adaptation with gene flow favors clustering (reduced map distance) between outlier loci, we used two metrics of clustering: nearest neighbor distance (NND) and the coefficient of variation. Both of these metrics were calculated using the SNP-level data.

We first asked: do map distances between nearest-neighbour outlier loci differ significantly from the expected map distances of identical numbers of nearest-neighbour SNPs? This approach was designed to explicitly account for disparities in SNP density that might occur due to differences in sequencing outcomes between our various datasets. To do this, we first partitioned each SNP data set by chromosome. Then, for each chromosome we identified the number of outlier loci using the previously described method. We then drew 10 000 samples of random SNPs from each chromosome equal to the number of outliers on that chromosome, and calculated the mean map distance between each SNP and its nearest neighbor in the random sample. We then compared the empirical mean nearest neighbor map distance of outliers to this null distribution for each chromosome within each individual comparison dataset. Significant over clustering (or under

clustering) of outliers was determined using the P values calculation method described previously. We then used permutations tests to compare (a) the proportion of chromosomes that were significantly over-clustered and (b) the difference between the average NND between outliers and the average NND expected between SNPs, in units of standard deviations, between the four selection and gene flow regimes.

Another commonly used metric of spatial clustering is the coefficient of variation, the ratio of the standard deviation in distances between objects and the mean of these distances. In our one-dimensional (chromosomal) case, these are the distances between each SNP and the next on the chromosome, moving in one direction, and including the start and end of the chromosome in the calculation of distances. Values exceeding one are indicative of over-dispersion (clustering), whereas values below one suggest under-dispersion (uniformity of distances). We calculated the coefficient of variation for outliers on each chromosome, and computed the mean for all chromosomes containing outliers for each comparison. We then used a permutation test to compare the means of this quantity among gene flow/selection regimes.

Chapter 5: Conclusion

Biological diversity is shaped by several core evolutionary processes. Of these processes, adaptation and speciation are arguably the most important. In this thesis, I presented three studies that advance our understanding of adaptation and speciation. Here, I connect each chapter to the broader literature, discuss key limitations of each study, and suggest avenues for future work.

5.1 A new system for studying early speciation

In Chapter 2, I showed that a new study system – the white stickleback – is an excellent model for studying the role of gene flow in speciation. Using a variety of population genetic methods, I found that white stickleback diverged very recently from marine common stickleback, and this likely occurred face of gene flow.

Implications

The results of Chapter 2 are important for three reasons. First, the majority of speciation study systems are pairs of sister taxa with substantial amounts of reproductive isolation and genomic divergence – in other words, near the midpoint of the speciation continuum (Hendry *et al.* 2009; Nosil & Feder 2011). In contrast, white stickleback provide a critical data point very early in the speciation continuum (in terms of genomic divergence). Further, the white stickleback is only the second case in which two forms of stickleback have been found to coexist in complete sympatry (McKinnon & Rundle 2002). The only other example is the benthic-limnetic species pairs found in

British Columbia (McKinnon & Rundle 2002). Secondly, our work suggests that incipient species of stickleback are not limited to post-glacial lakes and streams (McKinnon & Rundle 2002). This raises the potential for discovering additional young species of stickleback in the marine environment. It also highlights the importance of taking variation in the marine population into account when using it as a 'living ancestral' population in comparative studies (Jones *et al.* 2012) – the white stickleback is obviously not the ancestor of contemporary freshwater populations, for example. Lastly, the white stickleback provides a logistically accessible system for stickleback speciation research in eastern North America. To date, stickleback research has focused largely on glacial lakes on the west coast of North America (or Europe) (McKinnon & Rundle 2002; Hendry *et al.* 2009), creating logistical challenges for stickleback workers in the eastern parts of Canada and the United States. White stickleback provide many of the benefits of established stickleback systems (genomic and ecological knowledge, divergence with gene flow) without the need for long-distance travel.

Limitations

The analyses presented in Chapter 2 have a number of important limitations. A number of these arise because methods for fitting demographic models to genomic data are still in their infancy. For example, dadi has difficultly confidently distinguishing long divergence times and large amounts of gene flow from recent divergence and small amounts of gene flow (Gutenkunst *et al.* 2009). Further, dadi and TREEMIX both do not incorporate the effects of natural selection on genetic variation (Gutenkunst *et al.* 2009; Pickrell & Pritchard 2012). We chose to remove loci likely under natural selection (F_{ST} outliers), but ideally the selective effects of these loci could be incorporated into the dadi and TREEMIX models explicitly.

Another issue with Chapter 2 was our inability to examine genetic divergence on the male sex chromosome. Male sex chromosomes are thought to be important in speciation (Kitano *et al.* 2009; Yoshida *et al.* 2014), and it seems likely that the male-biased nature of white stickleback traits is connected to differences on the male sex chromosome. Indeed, some of our experimentation with crossing white and common stickleback suggests that the white coloration may map to the male sex chromosome. For example, in crosses where a male white stickleback was crossed to a female common stickleback, all the male ancestors of the original white male (F1, F2 and F3) displayed white breeding colors (Samuk, unpublished observation). Addressing the role of the male sex chromosome in reproductive isolation between white and common sticklebackwill be eventually possible when a reference Y chromosome is assembled and included in the stickleback reference genome.

5.2 The role of trophic and sexual divergence in speciation

In Chapter 3, I used the white stickleback system to examine the role of trophic versus sexual divergence in the early phases of speciation. I found evidence for weak trophic differentiation between white and common stickleback, and that divergent regions of the genome are enriched for male-biased genes.

Implications

These findings have several important implications. First, the results of Chapter 3 imply that speciation with gene flow need not occur via the evolution of trophic differences. Many authors

have suggested that speciation with gene flow might require this type of ecological differentiation (Hendry *et al.* 2007; Nosil 2008; Wolf *et al.* 2010) – our results appear to provide a counter example. Secondly, the fact that divergent regions of the genome harbour more male-biased genes suggests that male-linked traits probably played a key role in the evolution of reproductive isolation between white and common stickleback. This supports the idea that sexually-linked traits can form strong barriers early in the process of speciation (Edwards *et al.* 2005; Servedio & Kopp 2012).

That said, non-trophic ecological differences may still play a role in reproductive isolation between white and common stickleback. For example, the sexual differences we see between the two types could very well be environment dependent. Thus, the suite of male characters displayed by the white stickleback may only be more fit than the common male suite of characters in the presence of filamentous algae. Indeed, Blouw's original work on white stickleback shows that that the presence of filamentous algae strongly correlates with the occurrence of white stickleback (Blouw & Hagen 1990). Further, the lack of white stickleback outside of Nova Scotia suggests that some specific combination of environmental variables might be required for the white stickleback to colonize and persist (Blouw & Hagen 1990). Thus, ecology (in this case nesting habitat) may yet play a role in reproductive isolation between white and common stickleback.

Limitations

Chapter 3 integrated many different types of data, each with its own limitations. For example, while stable isotopic abundances did not differ between white and common stickleback, this does not necessarily mean that they are trophically equivalent (Hobson & Wassenaar 1999). White and common stickleback could be eating different sizes (or even species) of zooplankton that

happen to have the same isotopic signatures. Further, our panel of morphological traits was limited – for example, we did not examine the bones of the head and mouth, which are known to be important for feeding (Arnegard *et al.* 2014). Thus, white and common stickleback may differ in other ecological axes that we failed to capture in our study, including perhaps those related to body size differences.

Another limitation of Chapter 3 was the use of sex-biased expression as a proxy for 'sexual' traits – those involved in the acquisition of mates or mating. We followed the assumption common in the sexual selection literature that genes with sex-biased expression will tend to be enriched for those involved in mating (Ellegren & Parsch 2007; Perry *et al.* 2014). However, some (perhaps many) sex biased traits in stickleback likely have nothing to do with mating *per se* – for example, many of the genes involved in male parental care are probably male biased in expression (Páll *et al.* 2002; Hoffmann *et al.* 2008). Thus, the presence of more divergence in sex-biased genes may be not as strictly informative of the role of 'sexual' traits as we assumed. More work on the functional roles of different genes in mating and parental care in stickleback will help alleviate this issue.

5.3 The effects of gene flow during adaptation

In Chapter 4, my co-authors and I used a large genomic dataset to test the idea that divergent selection with gene flow favors clustering of adaptive alleles in the genome. We found that that was indeed the case – adaptive alleles tend be more tightly linked when both divergent selection and gene flow occur. Other studies have attempted to look at this relationship, but generally lacked the statistical power to do so – often having only a few comparisons between population pairs with and without gene flow (Renaut *et al.* 2013).

Implications

Of the results presented in this thesis, those of Chapter 4 are probably the most important. For one, the study in Chapter 4 specifically integrated genomic data from a wide variety of natural populations. Thus, the pattern we describe is probably not due to the idiosyncrasies of a single pair of populations or ecological contrast. Secondly, there is no reason to think that the general pattern of divergent selection with gene flow favoring clustered architectures is stickleback-specific. Other than having an unusually large geographic range, stickleback appear to be a fairly typical vertebrate in terms of life history, population connectivity, population sizes, etc. (McKinnon & Rundle 2002; Hendry *et al.* 2009). Together, these features suggest that the results of Chapter 4 are likely applicable to other systems.

If the results of Chapter 4 are indeed general, the implications could be quite major. First, our results imply that gene flow can potentially dramatically alter genomics of adaptation by constraining the usable area of the genome. This information could be of great use to those interested in explaining (or predicting) the genomics of adaptive evolution. The connection between reduced recombination and adaptation has been discussed in the theoretical literature, but is not generally appreciated by empiricists (Charlesworth *et al.* 1993; Barton 2010; Yeaman & Whitlock 2011). Secondly, our results may help explain the cause of the highly controversial "genomic islands" – genomic regions of unusually high divergence of varying sizes --described in many population genomic studies (Noor & Bennett 2009; Nadeau *et al.* 2011; Renaut *et al.* 2013). The biological meaning of these has been a matter of vigorous debate (Noor & Bennett 2009; Cruickshank & Hahn 2014). Our results suggest that genomic islands may be partially caused by selection favoring

divergence in low recombination regions in the presence of gene flow (i.e. making these regions large "islands").

Finally, from a methodological perspective, our study in Chapter 4 demonstrates that comparative population genomics can be a highly effective tool for studying evolutionary processes. This is an important point, as some authors have argued that the deluge of genomic studies over the last ten years has left us awash in sea of uninterpretable data (Feder *et al.* 2012; Seehausen *et al.* 2014). Hopefully, our study will inspire others to perform similar data-rich, but hypothesis-driven, studies in other taxa.

Limitations

Our exploration of the role of gene flow in shaping the genomic architecture of adaptation also has a number of key limitations. First, although we integrated as much genomic data as possible, we had few population pairs in the 'parallel selection with gene flow' category. Thus, we are less confident about the correlation between divergence and recombination rate in this category. Why were so few population pairs in this category? The most likely explanation is that populations exhibiting parallel selection and connected gene flow are less likely to be represented in the literature because they are generally not biologically interesting – in these cases, there is (by definition) very limited scope for both local adaptation and speciation (Slatkin 1987; Lenormand 2002). Patterns of genetic differentiation between such populations should be driven largely by isolation by distance, which is generally not considered an 'exciting' process compared to adaptation or gene flow (Marko & Hart 2011; Bradburd *et al.* 2013). However, pairs of populations experiencing parallel selection and gene flow provide a very useful point of contrast for other more interesting comparisons. For

example, in Chapter 3 we used a parallel selection + gene flow comparison (mainland common vs. Bras d'Or common) to control for factors that might cause elevated divergence in the absence of reproductive isolation.

The second major limitation of the Chapter 4 was that we had to rely on geographic distance as a proxy for gene flow. For stickleback, this is probably a reasonable assumption (McKinnon & Rundle 2002; Spoljaric & Reimchen 2007). However, explicit estimates of gene flow (e.g. via an isolation with migration model) between pairs of populations would likely be a more sensitive approach. In the case of Chapter 4, fitting thousands of isolation with migration models was not computationally feasible, but in the future this may be possible. A similar problem is that we were unable to account for *variation* in gene flow over the course of divergence between populations. The expectations for the evolution of clustered architectures are likely different between short bursts of gene flow with long periods of allopatry and constant gene flow (Yeaman & Whitlock 2011). Unfortunately, population genetic data cannot be used to infer complex time-courses of gene flow (Marko & Hart 2011). An experimental evolution approach could better address this particular issue.

5.4 Future work

The results I presented here suggest a number of important new lines of inquiry. For one, the white stickleback is now ready for "prime time", and can be used to probe many new evolutionary questions. For example, one could test the role of divergent natural selection in maintaining reproductive isolation between white and common stickleback – i.e. testing the hypothesis of ecological speciation (Schluter 2001). This could involve a reciprocal transplant to test the importance of filamentous algae in maintain reproductive isolation, or testing if female white

stickleback display concomitant preferences for white males as predicted by sexual selection theory (Ritchie 2007). Another possibility would be to use the white stickleback examine the genetic basis of the loss of parental care. This could be a major contribution to our understanding of the genetic basis of behavioral traits and, for example, if this basis differs from morphological traits in effect size, frequency regulatory vs. coding mutations, etc. (Boake *et al.* 2002).

With respect to Chapter 4, a key next step in understanding the role of gene flow in shaping the genomic architecture of adaptation would be to replicate our study in other systems. This replication would require systems with (a) large amounts of population genomic data, (b) a wide geographic range and (c) key knowledge of ecological adaptation and (d) pairs of populations in which to study this phenomenon. Some possibilities include Annual sunflowers, *Timema* walking sticks, *Arabidopsis,* or *Drosophila* (Rieseberg *et al.* 2006; Nosil *et al.* 2008; Stapley *et al.* 2010; Langley *et al.* 2012).

Finally, now that we know that there is a connection between the genomic architecture of adaptation and gene flow, we can test this relationship a variety of other ways. For example, we could examine how the correlation between adaptive molecular evolution (e.g. dN/dS) in regions of high vs. low recombination changes as a function of gene flow and selection (Presgraves 2005). This would avoid the problem of linked selection have a broader effect in regions of low recombination, and thus inflating the number of highly diverged loci.

5.5 Conclusion

Evolutionary biology has entered the genomic era. New datasets and tools abound, allowing us to examine the core processes of evolution with unprecedented resolution. In this thesis, I used a mix of genomic and phenotypic data to explore various aspects of adaptation and speciation. I

fleshed out a new system for studying speciation, which will be a key tool in future research. I also showed that early speciation need not involve strong trophic differences – sexual divergence appears to be enough. Finally, myself and my co-authors showed that gene flow can actually shape the genomic architecture of adaptation – a potentially far-reaching result. Together, I believe these studies provide significant advances in our understanding of adaptation and speciation, and hopefully will inspire future work on these key processes.

References

Adams DC, Otárola-Castillo E, Sherratt E (2014) *geomorph: Software for geometric morphometric analyses*. R package version 2.0.

Albert AYK, Sawaya S, Vines TH *et al.* (2007) The Genetics Of Adaptive Shape Shift In Stickleback: Pleiotropy And Effect Size. *Evolution*, **62**, 76–85.

Andrew RL, Kane NC, Baute GJ, Grassa CJ, Rieseberg LH (2012) Recent nonhybrid origin of sunflower ecotypes in a novel habitat. *Molecular Ecology*, **22**, 799–813.

Arnegard ME, McGee MD, Matthews B *et al.* (2014) Genetics of ecological divergence during speciation. *Nature*, **511**, 307–311.

Barton NH (2010) Genetic linkage and natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 2559–2569.

Beaumont MA (2005) Adaptation and speciation: what can Fst tell us? *Trends in Ecology & Evolution*, **20**, 435–440.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Bell MA, Aguirre WE (2013) Contemporary evolution, allelic recycling, and adaptive radiation of the threespine stickleback. *Evolutionary Ecology Research*, **15.4**, 377-411.

Bell MA, Foster SA (1994) *The evolutionary biology of the threespine stickleback*. Oxford University Press.

Benton MJ (2009). The Red Queen and the Court Jester: species diversity and the role of biotic and abiotic factors through time. *Science*, **323**(5915), 728-732.

Blouw D, Hagen D (1990) Breeding ecology and evidence of reproductive isolation of a widespread stickleback fish (Gasterosteidae) in Nova Scotia, Canada. *Biological Journal of the Linnean Society*, **39**, 195–217.

Blouw M, Blouw DM (1996) Evolution of offspring desertion in a stickleback fish. *Ecoscience*, **3**, 18–24.

Boake CRB, Arnold SJ, Breden F *et al.* (2002) Genetic Tools for Studying Adaptation and the Evolution of Behavior. *The American Naturalist*, **160**, S143–S159.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bolnick DI, Lau OL (2008) Predictable Patterns of Disruptive Selection in Stickleback in Postglacial Lakes. *The American Naturalist*, **172**, 1–11.

Boughman JW, Rundle HD, Schluter D (2005) Parallel evolution of sexual isolation in stickleback. *Evolution*, **59**, 361–373.

Bradburd GS, Ralph PL, Coop GM (2013) Disentangling The Effects Of Geographic And Ecological Isolation On Genetic Differentiation. *Evolution*, **67**, 3258–3273.

Brockmann HJ (2001) The evolution of alternative strategies and tactics. *Advances in the Study of Behavior* , **30**, 1-51.

Burri R, Nater A, Kawakami T, Mugal CF (2015) 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**, 1656–1665.

Butlin RK (2010) Population genomics and speciation. *Genetica*, *138*(4), 409-418.

Butlin RK, Saura M, Charrier G *et al.* (2014) Parallel Evolution Of Local Adaptation And Reproductive Isolation In The Face Of Gene Flow. *Evolution*, **68**, 935–949.

Carneiro M, Ferrand N, Nachman MW (2009) Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (Oryctolagus cuniculus). *Genetics*, **181**, 593–606.

Catchen J, Bassham S, Wilson T *et al.* (2013) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology*, **22**, 2864–2883.

Charlesworth B (2012) The Role of Background Selection in Shaping Patterns of Molecular Evolution and Variation: Evidence from Variability on the Drosophila X Chromosome. *Genetics*, **191**, 233–246.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Conte GL, Arnegard ME, Best J *et al.* (2015) Extent of QTL Reuse During Repeated Phenotypic Divergence of Sympatric Threespine Stickleback. *Genetics*, **201**, 1189–1200.

Conte GL, Schluter D (2013) Experimental Confirmation That Body Size Determines Mate Preference Via Phenotype Matching In A Stickleback Species Pair. *Evolution*, **67**, 1477–1484.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer, Sunderland, MA.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to

reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Cunningham F, Amode MR, Barrell D *et al.* (2015) Ensembl 2015. *Nucleic acids research*, **43**, D662–9.

Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics*, **14**, 262–274.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Dasmahapatra KK, Walters JR, Briscoe AD *et al.* (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**(7405), 94-98.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Dobzhansky T (1951). Genetics and the Origin of Species, edn 3. New York: Columbia University Press.

Durinck S, Spellman PT, Birney E, Huber W (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, **4**, 1184–1191.

Edwards SV, Kingan SB, Calkins JD *et al.* (2005) Speciation in birds: genes, geography, and sexual selection. *Proceedings of the National Academy of Sciences*, **102 Suppl 1**, 6550–6557.

Egan SP, Nosil P, Funk DJ (2008) Selection And Genomic Differentiation During Ecological Speciation: Isolating The Contributions Of Host Association Via A Comparative Genome Scan Of Neochlamisus Bebbianae Leaf Beetles. *Evolution*, **62**, 1162–1181.

Ellegren H, Parsch J (2007) The evolution of sex-biased genes and sex-biased gene expression. *Nature Reviews Genetics*, **8**, 689–698.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, **6**, e19379.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.

Feder JL, Flaxman SM, Egan SP, Comeault AA, Nosil P (2013) Geographic Mode of Speciation and Genomic Divergence. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 73–97.

Feder JL, Roethele JB, Filchak K, Niedbalski J, Romero-Severson J (2003) Evidence for Inversion Polymorphism Related to Sympatric Host Race Formation in the Apple Maggot Fly, Rhagoletis

pomonella. *Genetics*, **163**, 939–953.

Ferchaud A-L, Pedersen SH, Bekkevold D *et al.* (2014) A low-density SNP array for analyzing differential selection in freshwater and marine populations of threespine stickleback (Gasterosteus aculeatus). *BMC Genomics*, **15**, 867.

Feulner PGD, Kirschbaum F, Tiedemann R (2008) Adaptive radiation in the Congo River: An ecological speciation scenario for African weakly electric fish (Teleostei; Mormyridae; Campylomormyrus). *Journal of Physiology-Paris*, **102**, 340–346.

Flaxman SM, Feder JL, Nosil P (2013) Genetic Hitchhiking And The Dynamic Buildup Of Genomic Divergence During Speciation With Gene Flow. *Evolution*, **67**, 2577–2591.

Foster S, Robert K, Shaw K, Baker J (2008) Benthic, limnetic and oceanic threespine stickleback: profiles of reproductive behaviour. *Behaviour*, **145**, 485–508.

Fry, B (2006). *Stable Isotope Ecology*. Springer, Berlin

Glazer AM, Killingbeck EE, Mitros T, Rokhsar DS, Miller CT (2015) Genome Assembly Improvement and Mapping Convergently Evolved Skeletal Traits in Stickleback with Genotyping-by-Sequencing. *G3: Genes|Genomes|Genetics*, **5**, 1463–1472.

Gow JL, Peichel CL, TAYLOR EB (2007) Ecological selection against hybrids in natural populations of sympatric threespine stickleback. *Journal of Evolutionary Biology*, **20**, 2173–2180.

Grace JL, Shaw KL (2011) Coevolution Of Male Mating Signal And Female Preference During Early Lineage Divergence Of The Hawaiian Cricket, Laupala Cerasina. *Evolution*, **65**, 2184–2196.

Greenwood, AK, Cech JN., & Peichel CL (2012) Molecular and developmental contributions to divergent pigment patterns in marine and freshwater sticklebacks. *Evolution & development*, **14**(4), 351-362.

Gross MR (1996). Alternative reproductive strategies and tactics: diversity within sexes. *Trends in Ecology & Evolution*, **11**(2), 92-98.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data (G McVean, Ed,). *PLoS Genetics*, **5**, e1000695.

Haglund TR, Buth DG, Blouw DM (1990) Allozyme variation and the recognition of the "white stickleback." *Biochemical systematics and ecology*, **18**(7), 559-563.

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S (2015) Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, **32**, 835–845.

Hendry AP, Bolnick DI, Berner D, Peichel CL (2009) Along the speciation continuum in stickleback. *Journal of Fish Biology*, **75**, 2000–2036.

Hendry AP, Nosil P, Rieseberg LH (2007) The speed of ecological speciation. *Functional Ecology*.

Hendry AP, Taylor EB, McPhail JD (2002) Adaptive divergence and the balance between selection and gene flow: lake and stream stickleback in the Misty system. *Evolution*, **56**(6), 1199-1216.

Hennig C (2013) *fpc: Flexible procedures for clustering. R package version 2.1-5*.

Hobson KA, Wassenaar LI (1999) Stable isotope ecology: an introduction. *Oecologia* , **120**(3), 312-313.

Hoffmann E, Mayer I, Österman A, Borg B (2008) 11-ketotestosterone is not responsible for the entire testicular effect on male reproductive behaviour in the threespine stickleback. *Behaviour*, **145**, 509–525.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags (DJ Begun, Ed,). *PLoS Genetics*, **6**, e1000862.

Holliday, J. A., Zhou, L., Bawa, R., Zhang, M., & Oubida, R. W. (2016). Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in Populus trichocarpa. *New Phytologist*, *209*(3), 1240-1251.

Hyun S (2013) Body size regulation and insulin-like growth factor signaling. *Cellular and Molecular Life Sciences*, **70**, 2351–2365.

Jain AK, Dubes RC (1988) *Algorithms for clustering data*. Prentice-Hall, Inc.

Jamieson IG, Blouw DM, Colgan PW (1992a) Parental care as a constraint on male mating success in fishes: a comparative study of threespine and white stickleback. *Canadian Journal of Zoology*, **70**, 956.

Jamieson IG, Blouw DM, Colgan PW (1992b) Field observations on the reproductive biology of a newly discovered stickleback (Gasterosteus). *Canadian Journal of Zoology*, **70**, 1057–1063.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine stickleback. *Nature*, **484**, 55–61.

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

Kitano J, Ross JA, Mori S *et al.* (2009) A role for a neo-sex chromosome in stickleback speciation. *Nature*, **461**, 1079–1083.

Kopp M, Hermisson J (2006). The evolution of genetic architecture under frequency-dependent

disruptive selection. *Evolution*, **60**(8), 1537-1550.

Küpper C, Stocks M, Risse JE, dos Remedios N, Farrell LL, McRae SB, Morgan TC, Karlionova N, Pinchuk P, Verkuil YI, Kitaysky AS (2015). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics* **48**, 79–83

Langley CH, Stevens K, Cardeno C *et al.* (2012) Genomic Variation in Natural Populations of Drosophila melanogaster. *Genetics*, **192**, 533–598.

Leder EH, McCairns RJS, Leinonen T *et al.* (2015) The evolution and adaptive potential of transcriptional variation in sticklebacks--signatures of selection and widespread heritability. *Molecular Biology and Evolution*, **32**, 674–689.

Lenormand T (2002) Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.

Liou LW, Price TD (1994) Speciation by Reinforcement of Premating Isolation. *Evolution*, **48**, 1451.

Löytynoja A (2013) Phylogeny-aware alignment with PRANK. *Multiple Sequence Alignment Methods*, 155-170.

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.

M'Gonigle LK, Mazzucco R, Otto SP, Dieckmann U (2012) Sexual selection enables long-term coexistence despite ecological equivalence. *Nature*, **484**, 506–509.

Maan ME, Seehausen O (2011) Ecology, sexual selection and speciation. *Ecology Letters*, **14**, 591–602.

Macdonald JF, Bekkers J, Macisaac SM, Blouw DM (1995) Intertidal Breeding and Aerial Development of Embryos of a Stickleback Fish (Gasterosteus). *Behaviour*, **132**, 1183–1206.

Marchinko KB (2009) Predation's Role In Repeated Phenotypic And Genetic Divergence Of Armor In Threespine Stickleback. *Evolution*, **63**, 127–138.

Marie Curie SPECIATION Network, Butlin R, Debelle A *et al.* (2012) What do we need to know about speciation? *Trends in Ecology & Evolution*, **27**, 27–39.

Marko PB, Hart MW (2011) The complex analytical landscape of gene flow inference. *Trends in Ecology & Evolution*, **26**, 448–456.

Marques DA, Lucek K, Meier JI *et al.* (2016) Genomics of Rapid Incipient Speciation in Sympatric

Threespine Stickleback. *PLoS Genetics*, **12**, e1005887.

Matthews B, Marchinko KB, Bolnick DI, Mazumder A (2010) Specialization of trophic position and habitat use by stickleback in an adaptive radiation. *Ecology*, **91**, 1025–1034.

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

McKinnon JS (2002) Aquatic hotspots: speciation in ancient lakes III. *Trends in Ecology & Evolution*, **17**, 542–543.

McKinnon JS, Rundle HD (2002) Speciation in nature: the threespine stickleback model systems. *Trends in Ecology & Evolution*, *17*(10), 480-488.

Mendelson TC, Shaw KL (2005) Sexual behaviour: Rapid speciation in an arthropod. *Nature*, **433**, 375–376.

Muschick M, Indermaur A, Salzburger W (2012). Convergent evolution within an adaptive radiation of cichlid fishes. *Current biology*, **22**(24), 2362-2368.

Nachman MW, Payseur BA (2012) Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philosophical Transactions Of The Royal Society B-Biological Sciences*, **367**, 409–421.

Nadeau NJ, Whibley A, Jones RT *et al.* (2011) Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 343–353.

Nagel L, Schluter D (1998) Body Size, Natural Selection, and Speciation in Stickleback. *Evolution*, **52**, 209.

Narum SR, Hess JE (2011) Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11**, 184–194.

Navarro A, Barton N (2003) Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution*, *57*(3), 447-459.

Nei M, Miller JC (1990) A Simple Method for Estimating Average Number of Nucleotide Substitutions within and between Populations from Restriction Data. *Genetics*, **125**, 873.

Noor M, Cunningham AL, Larkin JC (2001b) Consequences of Recombination Rate Variation on Quantitative Trait Locus Mapping Studies: Simulations Based on the *Drosophila melanogaster* Genome. *Genetics*,**159**(2), 581-588.

Noor M, Feder JL (2006) Speciation genetics: evolving approaches. *Nature Reviews Genetics*, *7*(11),

851-861.

Noor MAF, Bennett SM (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, **103**, 439–444.

Noor MAF, Grams KL, Bertucci LA, Reiland J (2001a) Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences*, **98**, 12084–12088.

Nosil P (2008) Speciation with gene flow could be common. *Molecular Ecology*, **17**, 2103–2106.

Nosil P, Egan SP, Funk DJ (2008) Heterogeneous Genomic Differentiation Between Walking-Stick Ecotypes: "Isolation By Adaptation" And Multiple Roles For Divergent Selection. *Evolution*, **62**, 316–336.

Nosil P, Feder JL (2011) Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 332–342.

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

Nosil P, Vines TH, Funk DJ (2005) Reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution*,**59**(4), 705-719.

Oleksyk TK, Smith MW, O'Brien SJ (2009) Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 185–205.

Orr HA (2005) The genetic basis of reproductive isolation: insights from Drosophila. *Proceedings of the National Academy of Sciences*, ***102***(suppl 1), 6522-6526.

Páll MK, Mayer I, Borg B (2002) Androgen and Behavior in the Male Three-Spined Stickleback, Gasterosteus aculeatus. *Hormones and Behavior*, **42**, 337–344.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289–290.

Pavlidis P, Jensen JD, Stephan W, Stamatakis A (2012) A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. *Molecular Biology and Evolution*, **29**, 3237–3248.

Peichel CL, Nereng KS, Ohgi KA *et al.* (2001) The genetic architecture of divergence between threespine stickleback species. *Nature*, **414**, 901–905.

Peichel CL, Ross JA, Matson CK *et al.* (2004) The Master Sex-Determination Locus in Threespine Stickleback Is on a Nascent Y Chromosome. *Current Biology*, **14**, 1416–1424.

Perry JC, Harrison PW, Mank JE (2014) The ontogeny and evolution of sex-biased gene expression

in Drosophila melanogaster. *Molecular Biology and Evolution* , **31**(5), 1206-1219.

Pickrell JK, Pritchard JK (2012) Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data (H Tang, Ed,). *PLoS Genetics*, **8**, e1002967.

Pitcher TE, Dunn PO, Whittingham LA (2005) Sperm competition and the evolution of testes size in birds. *Journal of Evolutionary Biology*, **18**, 557–567.

Presgraves DC (2005) Recombination Enhances Protein Adaptation in Drosophila melanogaster. *Current Biology*, **15**, 1651–1656.

Price T (2008) *Speciation in birds*. Roberts and Co.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, **81**, 559–575.

R Core Team (2015) *R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, 2012)*. URL: http://www.R-project.org. Retrieved April 15th 2016.

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, **197**, 573–589.

Rasband WS (2012) ImageJ: Image processing and analysis in Java. *Astrophysics Source Code Library*.

Rattray A, Santoyo G, Shafer B, Strathern JN (2015) Elevated Mutation Rate during Meiosis in Saccharomyces cerevisiae. *PLoS Genetics*, **11**, e1004910.

Reimchen TE, Ingram T, Hansen SC (2008) Assessing niche differences of sex, armour and asymmetry phenotypes using stable isotope analyses in Haida Gwaii stickleback. *Behaviour*, **145**, 561–577.

Reimchen TE, Nosil P (2004) Variable Predation Regimes Predict The Evolution Of Sexual Dimorphism In A Population Of Threespine Stickleback. *Evolution*, **58**, 1274.

Reimchen, TE, Nosil P (2004) Variable predation regimes predict the evolution of sexual dimorphism in a population of threespine stickleback. *Evolution*, **58**(6), 1274-1281.

Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.

Rieseberg L (2001) Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, **16**, 351–358.

Rieseberg LH, Kim S-C, Randell RA *et al.* (2006) Hybridization and the colonization of novel habitats by annual sunflowers. *Genetica*, **129**, 149–165.

Rieseberg LH, Willis JH (2007) Plant Speciation. *Science*, **317**, 910–914.

Ritchie MG (2007) Sexual selection and speciation. *Annual Review of Ecology*.

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

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

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

Rogers SM, Bernatchez L (2006) The genetic basis of intrinsic and extrinsic post-zygotic reproductive isolation jointly promoting speciation in the lake whitefish species complex (Coregonus clupeaformis). *Journal of Evolutionary Biology*, **19**, 1979–1994.

Rundle HD, Schluter D (2004) Natural selection and ecological speciation in stickleback. *Adaptive speciation*, *19*(3), 192-209.

Safran RJ, Scordato ESC, Symes LB, Rodríguez RL, Mendelson TC (2013) Contributions of natural and sexual selection to the evolution of premating reproductive isolation: a research agenda. *Trends in Ecology & Evolution*, **28**, 643–650.

Samuk K, Iritani D, Schluter D (2014) Reversed brain size sexual dimorphism accompanies loss of parental care in white sticklebacks. *Ecology and Evolution*. doi: 10.1002/ece3.1175.

Schluter D (2001) Ecology and the origin of species. *Trends in Ecology & Evolution*. ***16***(7), 372-380.

Schluter D (2009). Evidence for ecological speciation and its alternative. *Science*, **323**(5915), 737-741.

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

Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proceedings of the National Academy of Sciences*, **106 Suppl 1**, 9955–9962.

Schluter D, McPhail JD (1992) Ecological Character Displacement and Speciation in Stickleback. *American Naturalist*, **140**(1): 85-108.

Seehausen O, Butlin RK, Keller I *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192.

Servedio MR, Bürger R (2014) The counterintuitive role of sexual selection in species maintenance and speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 8113–8118.

Servedio MR, Kopp M (2012) Sexual selection and magic traits in speciation with gene flow. *Curr. Zool*, *58*(3), 507-513.

Shaw J, Piper DJW, Fader GBJ *et al.* (2006) A conceptual model of the deglaciation of Atlantic Canada. *Quaternary Science Reviews*, **25**, 2059–2081.

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

Smith C (1999) Parental energy expenditure of the male three-spined stickleback. *Journal of Fish Biology*, **54**, 1132–1136.

Soulsbury CD (2010) Genetic Patterns of Paternity and Testes Size in Mammals (A Dornhaus, Ed,). *PLoS ONE*, **5**, e9581.

Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, **14**, 404–414.

Spoljaric MA, Reimchen TE (2007) 10 000 years later: evolution of body shape in Haida Gwaii three-spined stickleback. *Journal of Fish Biology*, **70**, 1484–1503.

Stapley J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, **25**, 705–712.

Stasko AD, Swanson H, Majewski A, Atchison, S, Reist J, Power M (2016) Influences of depth and pelagic subsidies on the size-based trophic structure of Beaufort Sea fish communities. Marine *Ecology Progress Series*, **549**, 153-166.

Taborsky, M., Oliveira, RF, Brockmann HJ. (2008). The evolution of alternative reproductive tactics: concepts and questions. *Alternative reproductive tactics: an integrative approach*. Cambridge University Press, Cambridge, 1-21.

Taylor EB, Gerlinsky C, Farrell N, Gow JL (2011) A Test Of Hybrid Growth Disadvantage In Wild, Free-Ranging Species Pairs Of Threespine Stickleback (Gasterosteus Aculeatus) And Its Implications For Ecological Speciation. *Evolution*, **66**, 240–251.

Taylor, EB, McPhail DJ. (2000) Historical contingency and ecological determinism interact to prime speciation in stickleback, Gasterosteus. *Proceedings of the Royal Society B: Biological Sciences*, **267**, 2375–2384.

Vavrek MJ (2010) fossil: Palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, **14**, 1T.

Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Sciences*,

106, 9939–9946.

Viitaniemi HM, Leder EH (2011) Sex-Biased Protein Expression in Threespine Stickleback, Gasterosteus aculeatus. *Journal of Proteome Research*, **10**, 4033–4040.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, **38**, 1358.

Williams GC (1966) *Adaptation and natural selection: a critique of some current evolutionary thought*. Princeton University Press, Princeston.

Wolf JBW, Lindell J, Backström N (2010) Speciation genetics: current status and evolving approaches. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 1717–1733.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, **15**, 496–503.

Yeaman S, Whitlock MC (2011) The Genetic Architecture Of Adaptation Under Migration-Selection Balance. *Evolution*, **65**, 1897–1911.

Yoshida K, Makino T, Yamaguchi K *et al.* (2014) Sex Chromosome Turnover Contributes to Genomic Divergence between Incipient Stickleback Species (J Zhang, Ed,). *PLoS Genetics*, **10**, e1004223.

Zhang Y, Sturgill D, Parisi M, Kumar S, Oliver B (2007) Constraint and turnover in sex-biased gene expression in the genus Drosophila. *Nature*,**450**(7167), 233-237.

Zheng X, Levine D, Shen J *et al.* (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.

Appendices

Appendix A - Chapter 2 Supplementary Material

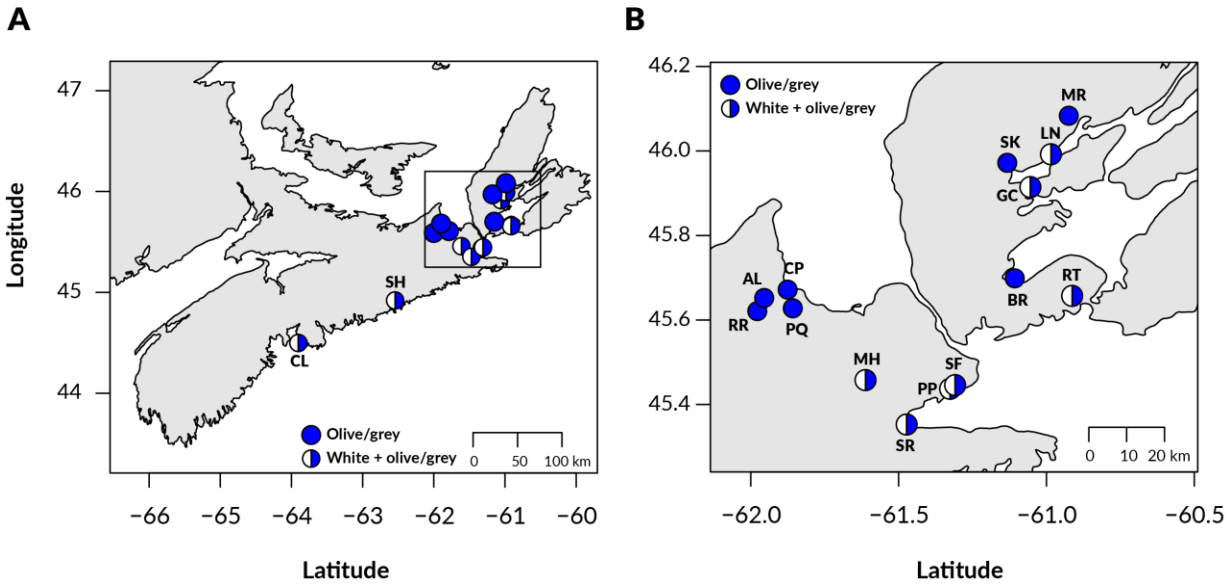

Figure A.1| Stickleback samples sites in Nova Scotia, Canada. Symbols represent the presence of putative common and "white" (pearlescent white dorsal colors) nesting male stickleback. Two-colored symbols represent sites where both types were observed nesting in the same general location. **A** Provincial view, **B** Detail view of Guysborough / Bras d'Or Lake sample sites. Site labels: CL = Canal Lake, SH = Sheet Harbour, RR = Rights River, AL = Antigonish Landing, CP = Captain's Pond, PQ = Pomquet, MH = Milford Haven River, SF = St. Francis Harbour, PP = Porper Pond, SR = Salmon River, RT = River Tillard, BR = Black River, GC = Gillies Cover, SK = Skye River, LN = Little Narrows, MR = Middle River.

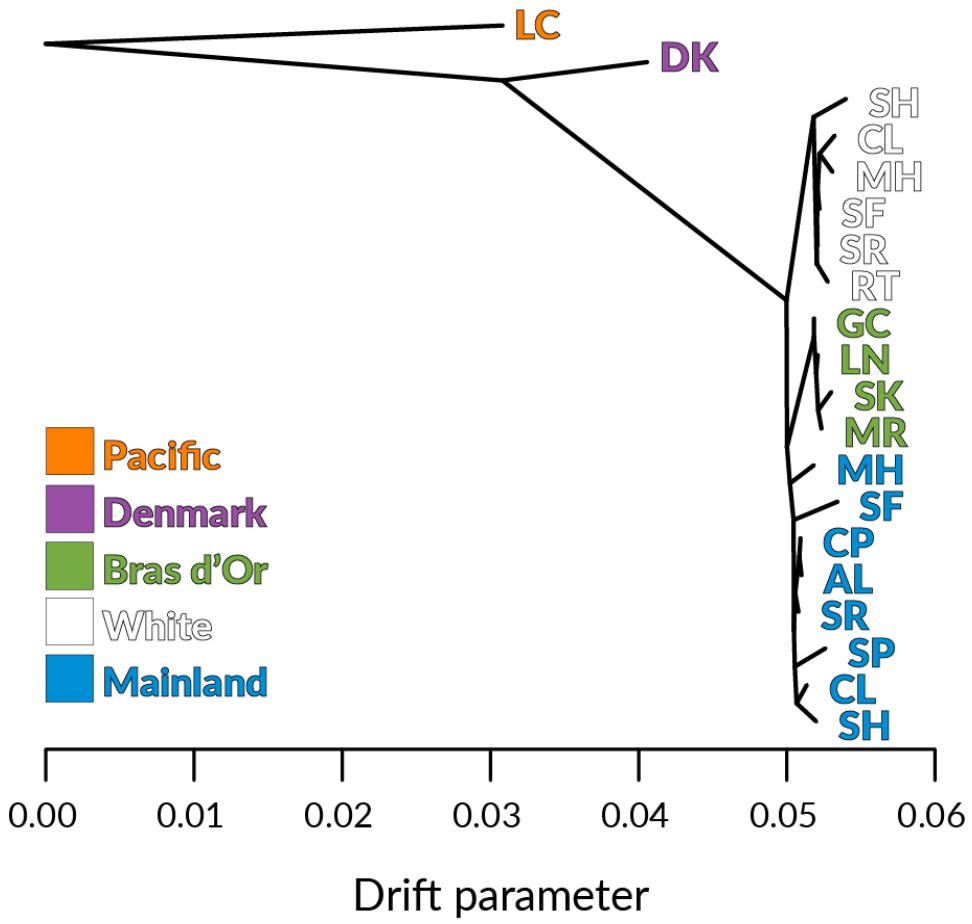

Figure A.2| TREEMIX maximum likelihood tree of stickleback populations from Nova Scotia, sampled in 2014. The tree was fit by designating the Pacific population (Little Campbell River marine) as the outgroup. The drift parameter corresponds to the estimated amount of genetic drift that has occurred between populations.

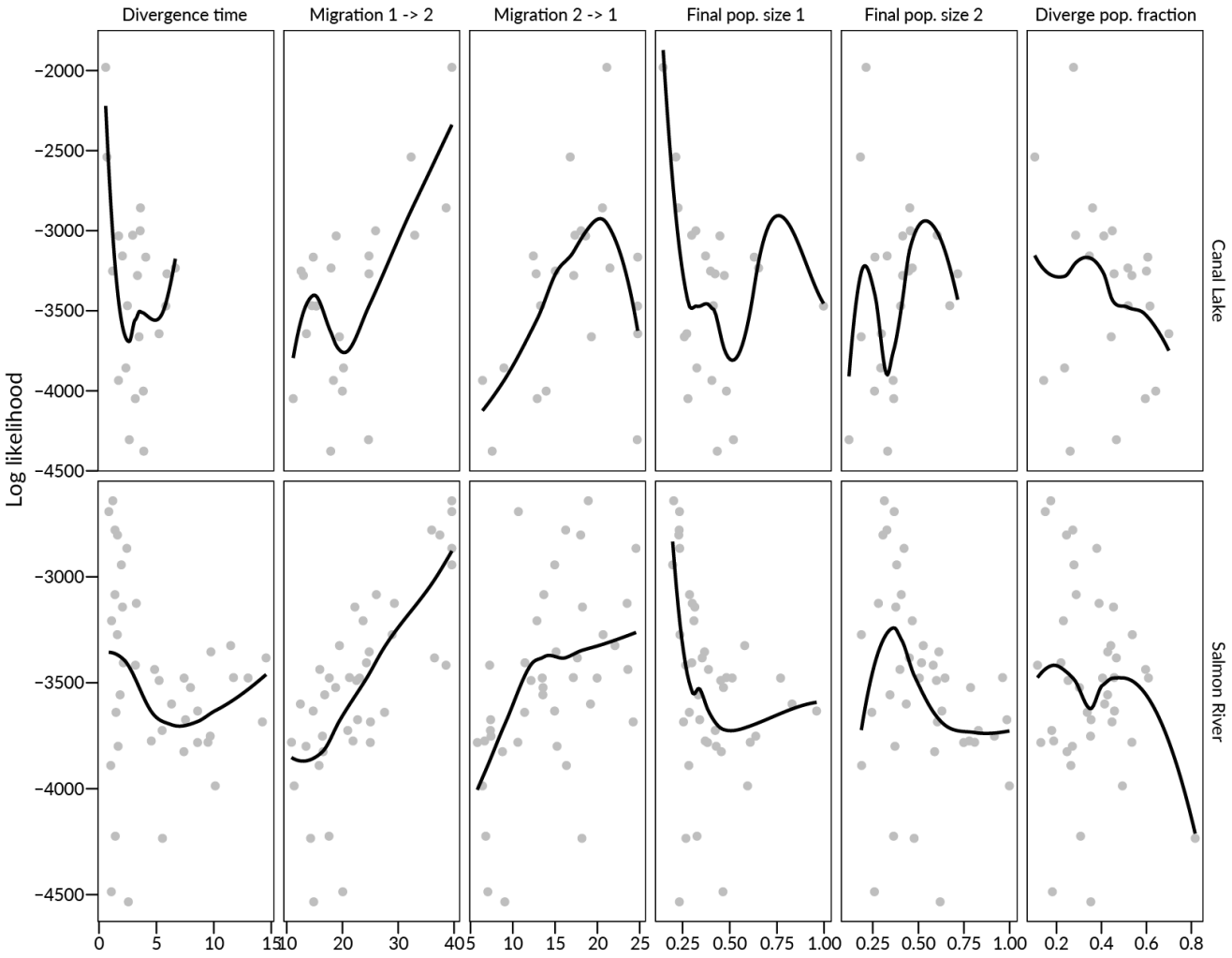

Figure A.3| Model log likelihoods and corresponding parameter estimates for isolation with migration models fit using dadi. Each point represents a parameter estimate from an optimized model resulting from a single run of dadi. Lines are LOESS smoothed conditional means. Because all parameters are optimized simultaneously, log likelihood values correspond to the total log likelihood of a complete model (including all six parameters, plus an estimate of the population genetic parameter theta). Parameter estimates for the Canal Lake and Salmon River opulation pairs are shown in the top and bottom rows respectively.

Appendix B - Chapter 3 Supplementary Material

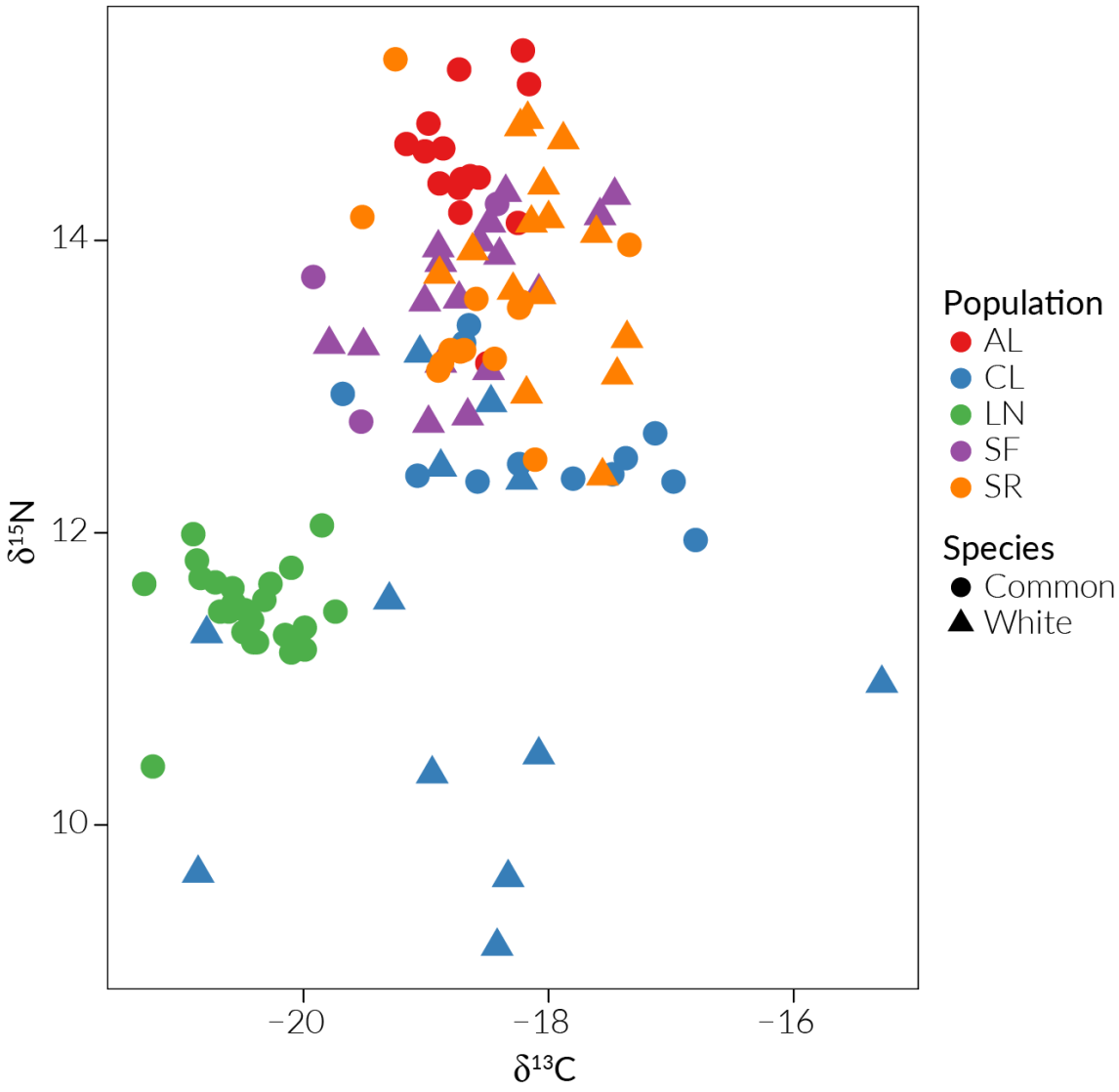

Figure B.4| Raw ^{13}C and ^{15}N stable isotope abundances in white and common stickleback. Colors represent the following geographic sampling locations: AL, Antigonish Landing; CL, Canal Lake; LN, Little Narrows; SF, St. Francis Harbour; SR, Salmon River.

Table B.1 | Genomic windows found to contain more outlier (>95th percentile) SNPs than expected by chance. Location columns specify the chromosome and genomic coordinates of each window. Average F_{ST} refers to the average F_{ST} of all SNPs (outlier and otherwise) in each window. F_{ST} values from three separate pair-wise comparisons are provided. Sex bias refers to the direction (Male or Female) and magnitude of sex-biased gene expression. ENSEMBL identifiers are provided along with short and long gene names.

Location			Average F_{ST}			Sex Bias		Gene information		
Chr	Start	End	White vs. Common	White vs. Bras d'Or	Common vs. Bras d'Or	Direction	Magnitude	ENSEMBL ID	Name	Description
1	12225001	12300000	0.25831	0.33179	0.009444	M	0.22	ENSGACG00000011157	traf4a	tnf receptor-associated factor 4a
1	12225001	12300000	0.25831	0.33179	0.009444	F	0.062	ENSGACG00000011119	tada2a	transcriptional adaptor 2A
1	19500001	19575000	0.252154	0.137594	0.024919	F	0.211	ENSGACG00000013633	akap1b	A kinase
1	19500001	19575000	0.252154	0.137594	0.024919	M	0.1	ENSGACG00000013619	dhrs13b	dehydrogenase/reductase
1	21600001	21675000	0.121909	0.063195	0.016371	M	0.161	ENSGACG00000014296	inha	inhibin, alpha
1	22650001	22725000	0.076106	0.135709	0.039491	M	0.345	ENSGACG00000014752		
1	22650001	22725000	0.076106	0.135709	0.039491	F	0.108	ENSGACG00000014735	lancl1	
4	13800001	13875000	0.126784	0.080881	0.02978	F	0.114	ENSGACG00000018408	CDHR2	cadherin related family member 2
4	24600001	24675000	0.1401	0.018164	0.090391	M	0.19	ENSGACG00000019373	snx4	sorting nexin 4
4	24600001	24675000	0.1401	0.018164	0.090391	M	0.035	ENSGACG00000019383	ppp6r2a	
4	25350001	25425000	0.118378	0.19312	0.053497	M	0.184	ENSGACG00000019480		Novel gene
4	25500001	25575000	0.006243	0.216117	0.143218	F	0.151	ENSGACG00000019497	apex1	APEX nuclease
4	25500001	25575000	0.006243	0.216117	0.143218	F	0.047	ENSGACG00000019505	si:dkey-14k9.3	si:dkey-14k9.3
5	3900001	3975000	0.059459	0.102522	0.03371	M	0.471	ENSGACG00000003652	g6pca.2	glucose-6-phosphatase a
5	3900001	3975000	0.059459	0.102522	0.03371	M	0.172	ENSGACG00000003660	g6pca.1	glucose-6-phosphatase a
5	8400001	8475000	0.108313	0.114622	0.007942	F	0.05	ENSGACG00000006563	dclre1a	DNA cross-link repair 1A
5	8400001	8475000	0.108313	0.114622	0.007942	F	0.008	ENSGACG00000006575	nhlrc2	NHL repeat containing 2
5	8475001	8550000	0.15172	0.17183	0.003367	M	0.224	ENSGACG00000006599	afap1l2	
6	2250001	2325000	0.173455	0.054981	0.047949	M	0.107	ENSGACG00000002949	mar5	membrane-associated ring finger
6	2325001	2400000	0.22576	0.092915	0.049091	M	0.053	ENSGACG00000002988	cdc42ep3	
6	15225001	15300000	0.114103	0.15436	0.011205	M	0.066	ENSGACG00000011492		Novel gene
6	15225001	15300000	0.114103	0.15436	0.011205	F	0.035	ENSGACG00000011471	cox20	
7	4200001	4275000	0.174866	0.06592	0.036535	M	0.445	ENSGACG00000019365	shbg	sex hormone-binding globulin

Location			Average F_{ST}			Sex Bias		Gene information		
Chr	Start	End	White vs. Common	White vs. Bras d'Or	Common vs. Bras d'Or	Direction	Magnitude	ENSEMBL ID	Name	Description
7	4200001	4275000	0.174866	0.06592	0.036535	M	0.098	ENSGACG00000019361	hdlbpb	high density lipoprotein binding protein b
7	4200001	4275000	0.174866	0.06592	0.036535	F	0.065	ENSGACG00000019350	TRIP6	thyroid hormone receptor interactor 6
7	6075001	6150000	0.118441	0.147622	0.006567	M	0.071	ENSGACG00000019537	sec24d	
7	22350001	22425000	0.094202	0.120042	0.024214	F	0.496	ENSGACG00000020692	slc25a5	
7	22350001	22425000	0.094202	0.120042	0.024214	F	0.019	ENSGACG00000020693	zbtb20	zinc finger and BTB domain containing 20
8	6900001	6975000	0.077812	0.066111	0.033393	M	0.414	ENSGACG00000006545	si:ch211-262h13.5	si:ch211-262h13.5
8	6900001	6975000	0.077812	0.066111	0.033393	F	0.092	ENSGACG00000006562	MCOLN3	mucolipin 3
8	6900001	6975000	0.077812	0.066111	0.033393	F	0.021	ENSGACG00000006551	gpsm2l	G-protein signaling modulator 2, like
9	10050001	10125000	0.156382	0.076105	0.025303	M	1.251	ENSGACG00000018120	IGFBP4	insulin like growth factor binding protein 4
9	10050001	10125000	0.156382	0.076105	0.025303	F	0.009	ENSGACG00000018126	CISD3	CDGSH iron sulfur domain 3
9	10050001	10125000	0.156382	0.076105	0.025303	M	0.008	ENSGACG00000018128	PCGF2	polycomb group ring finger 2
9	10425001	10500000	0.187077	0.115175	0.032595	F	0.255	ENSGACG00000018210	trim25	tripartite motif containing 25
9	10425001	10500000	0.187077	0.115175	0.032595	M	0.185	ENSGACG00000018206	rasal3	RAS protein activator like 3
9	10425001	10500000	0.187077	0.115175	0.032595	M	0.174	ENSGACG00000018208	trim25	tripartite motif containing 25
9	10425001	10500000	0.187077	0.115175	0.032595	M	0.155	ENSGACG00000018202	SEC14L1	SEC14-like lipid binding 1
9	10425001	10500000	0.187077	0.115175	0.032595	M	0.044	ENSGACG00000018216	cbx8b	
9	10650001	10725000	0.121478	0.12817	0.013648	F	0.485	ENSGACG00000018233	pdcd11	programmed cell death 11
9	10650001	10725000	0.121478	0.12817	0.013648	M	0.102	ENSGACG00000018230	CEP95	
10	7350001	7425000	0.055612	0.147526	0.033291	M	0.224	ENSGACG00000004817	rprd1a	
10	7350001	7425000	0.055612	0.147526	0.033291	M	0.065	ENSGACG00000004837	si:ch211-13c6.2	si:ch211-13c6.2
10	7350001	7425000	0.055612	0.147526	0.033291	F	0.055	ENSGACG00000004846	nfyc	nuclear transcription factor Y
10	13125001	13200000	0.189017	0.060273	0.049051	M	0.126	ENSGACG00000008805	ST3GAL1	ST3 beta-galactoside
10	13125001	13200000	0.189017	0.060273	0.049051	M	0.014	ENSGACG00000008798	si:ch1073	si:ch1073-296d18.1

Location			Average F_{ST}			Sex Bias		Gene information		
Chr	Start	End	White vs. Common	White vs. Bras d'Or	Common vs. Bras d'Or	Direction	Magnitude	ENSEMBL ID	Name	Description
11	10425001	10500000	0.756388	0.717484	0	M	0.317	ENSGACG00000011601	Novel gene	
11	10425001	10500000	0.756388	0.717484	0	M	0.198	ENSGACG00000011578	DXO	decapping exoribonuclease
11	10425001	10500000	0.756388	0.717484	0	M	0.168	ENSGACG00000011550	si:ch211-256e16.3	si:ch211-256e16.3
11	10425001	10500000	0.756388	0.717484	0	F	0.127	ENSGACG00000011592	stk19	
12	7875001	7950000	0.256928	0.230369	0.000216	F	0.363	ENSGACG00000006447	ybx1	Y box binding protein 1
12	7875001	7950000	0.256928	0.230369	0.000216	M	0.309	ENSGACG00000006452	arhgef16	Rho guanine nucleotide exchange
15	8175001	8250000	0.145391	0.081074	0.02338	M	0.546	ENSGACG00000009968	crlf1a	cytokine receptor-like factor 1a
15	8175001	8250000	0.145391	0.081074	0.02338	M	0.214	ENSGACG00000009976	keap1a	kelch-like ECH-associated
16	9300001	9375000	0.081652	0.078257	0.006642	F	0.123	ENSGACG00000004189	uggt2	UDP-glucose glycoprotein glucosyltransferase 2
16	9300001	9375000	0.081652	0.078257	0.006642	F	0.009	ENSGACG00000004166	dnajc3a	DnaJ
16	9975001	10050000	0.096806	0.121849	0.020575	M	0.281	ENSGACG00000004687	gpr155a	G protein-coupled receptor 155a
16	9975001	10050000	0.096806	0.121849	0.020575	M	0.186	ENSGACG00000004749	si:ch73-167c12.2	
16	9975001	10050000	0.096806	0.121849	0.020575	M	0.132	ENSGACG00000004721	scrn3	secernin 3
16	17625001	17700000	0.087822	0.174921	0.02715	M	0.2	ENSGACG00000009132	armc8	armadillo repeat containing 8
16	17625001	17700000	0.087822	0.174921	0.02715	M	0.116	ENSGACG00000009047	dnajb11	DnaJ
16	17625001	17700000	0.087822	0.174921	0.02715	M	0.096	ENSGACG00000009102	rab5b	RAB5B, member RAS oncogene family
16	17625001	17700000	0.087822	0.174921	0.02715	F	0.054	ENSGACG00000009122	dbr1	debranching RNA lariats 1
17	4875001	4950000	0.130357	0.067325	0.025469	M	0.348	ENSGACG00000006223	lmod1b	leiomodin 1b
17	4875001	4950000	0.130357	0.067325	0.025469	F	0.14	ENSGACG00000006233	ipo9	
17	4875001	4950000	0.130357	0.067325	0.025469	M	0.058	ENSGACG00000006187	timm17a	translocase of inner mitochondrial membrane
17	4875001	4950000	0.130357	0.067325	0.025469	F	0.043	ENSGACG00000006153	rap1ab	RAP1A, member of RAS oncogene family b
17	8025001	8100000	0.092204	0.050912	0.019746	M	0.042	ENSGACG00000008661	tnks1bp1	tankyrase 1 binding protein 1
17	8025001	8100000	0.092204	0.050912	0.019746	F	0.032	ENSGACG00000008663	Novel gene	
18	6225001	6300000	0.107588	0.215214	0.024027	M	0.263	ENSGACG00000007459	gareml	GRB2 MAPK1-like regulator

Location			Average F_{ST}			Sex Bias		Gene information		
Chr	Start	End	White vs. Common	White vs. Bras d'Or	Common vs. Bras d'Or	Direction	Magnitude	ENSEMBL ID	Name	Description
18	6300001	6375000	0.200318	0.290753	0.00745	F	0.125	ENSGACG00000007478	Novel gene	
18	6300001	6375000	0.200318	0.290753	0.00745	F	0.064	ENSGACG00000007514	esr2a	estrogen receptor 2a
18	6300001	6375000	0.200318	0.290753	0.00745	F	0.061	ENSGACG00000007483	syne2b	spectrin repeat containing, nuclear envelope
18	6450001	6525000	0.084937	0.124306	0.016623	M	0.666	ENSGACG00000007631	clu	clusterin
18	6450001	6525000	0.084937	0.124306	0.016623	M	0.2	ENSGACG00000007571	heca	
18	6450001	6525000	0.084937	0.124306	0.016623	F	0.195	ENSGACG00000007565	mtif3	mitochondrial translational initiation factor 3
18	6450001	6525000	0.084937	0.124306	0.016623	M	0.186	ENSGACG00000007582	abracl	
18	8550001	8625000	0.069055	0.213069	0.068178	M	0.208	ENSGACG00000008921	snx3	sorting nexin 3
18	8550001	8625000	0.069055	0.213069	0.068178	M	0.069	ENSGACG00000008950	ostm1	
18	11175001	11250000	0.03976	0.084893	0.049158	F	0.278	ENSGACG00000011113	si:ch211-286f9.2	
18	11175001	11250000	0.03976	0.084893	0.049158	F	0.239	ENSGACG00000011129	LINC00116	
18	11625001	11700000	0.094429	0.018638	0.039109	F	0.101	ENSGACG00000011399	si:ch211-225h24.2	
18	11625001	11700000	0.094429	0.018638	0.039109	F	0.085	ENSGACG00000011452	tmem18	
18	11625001	11700000	0.094429	0.018638	0.039109	M	0.054	ENSGACG00000011402	fbxo25	
18	11625001	11700000	0.094429	0.018638	0.039109	F	0.022	ENSGACG00000011427	acp1	
19	7050001	7125000	0.019806	0.015755	0.0183	F	1.031	ENSGACG00000005659	gtf2h1	
19	7050001	7125000	0.019806	0.015755	0.0183	F	0.986	ENSGACG00000005590	INCENP	
19	7050001	7125000	0.019806	0.015755	0.0183	F	0.775	ENSGACG00000005561	athl1	ATH1, acid trehalase-like 1
19	7050001	7125000	0.019806	0.015755	0.0183	F	0.676	ENSGACG00000005632	hps5	Hermansky-Pudlak syndrome 5
19	11550001	11625000	0.030804	0.042761	0.050858	F	0.872	ENSGACG00000010024	AEBP2	AE binding protein 2
20	7875001	7950000	0.021005	0.163616	0.086679	M	0.563	ENSGACG00000007038	Novel gene	
20	7875001	7950000	0.021005	0.163616	0.086679	M	0.331	ENSGACG00000007040	Novel gene	
20	7875001	7950000	0.021005	0.163616	0.086679	M	0.175	ENSGACG00000007061	mrpl13	mitochondrial ribosomal protein
20	11025001	11100000	0.13778	0.13564	0.002271	F	0.708	ENSGACG00000009439	znf574	zinc finger protein 574

Location			Average F_{ST}			Sex Bias		Gene information		
Chr	Start	End	White vs. Common	White vs. Bras d'Or	Common vs. Bras d'Or	Direction	Magnitude	ENSEMBL ID	Name	Description
20	11025001	11100000	0.13778	0.13564	0.002271	F	0.379	ENSGACG00000009451	Novel gene	
20	11025001	11100000	0.13778	0.13564	0.002271	F	0.237	ENSGACG00000009439	znf574	zinc finger protein 574
20	11025001	11100000	0.13778	0.13564	0.002271	F	0.019	ENSGACG00000009449	znf526	zinc finger protein 526
21	2925001	3.00E+06	0.217365	0.176361	0.005635	M	0.506	ENSGACG00000002155	RAMP3	receptor
21	9675001	9750000	0.159078	0.092084	0.022575	M	0.458	ENSGACG00000004347	SOX17	
21	9675001	9750000	0.159078	0.092084	0.022575	M	0.086	ENSGACG00000004365	vps41	vacuolar protein sorting 41 homolog
21	9675001	9750000	0.159078	0.092084	0.022575	F	0.009	ENSGACG00000004363	Novel gene	
21	9675001	9750000	0.159078	0.092084	0.022575	M	0.005	ENSGACG00000004359	esco1	
21	10650001	10725000	0.118973	0.056434	0.046541	M	0.094	ENSGACG00000004967	ralaa	
21	10650001	10725000	0.118973	0.056434	0.046541	M	0.077	ENSGACG00000004953	fbxl7	F-box and leucine-rich repeat protein 7
21	10650001	10725000	0.118973	0.056434	0.046541	F	0.034	ENSGACG00000004982	cdk13	cyclin-dependent kinase 13
Un	5550001	5625000	0.117639	0.135304	0.014323	M	0.452	ENSGACG00000000869	CLEC19A	cyclin-dependent kinase 13

Table B.2| Statistical summaries of morphological trait distributions in white and common stickleback, with all statistical tests including sex as a covariate. The mean and standard deviation (sd) are reported for each type along with sample size (n). Cohen's D is the difference in means (white minus common) in units of the pooled standard deviation. Tests of statistical significance either took the form of an ANOVA (F test), or a chi-squared test following an analysis of deviance (D statistic). The "no covariate" and "standard length as covariate" columns refer to tests of significance without and with body size correction. All p-values were corrected for multiple comparisons via the false discovery rate method.

			Common		White			No covariate		Standard length as covariate	
Trait	**Function**	**Unit**	**Mean (sd)**	**n**	**Mean (sd)**	**n**	**Cohen's D**	**Test statistic**	**p-value (FDR)**	**Test statistic**	**p-value (FDR)**
Body depth	Trophic	cm	1.23 (0.12)	161	0.95 (0.12)	73	-2.29	F1,231 = 275.06	0.00006	F1,230 = 0.71	0.48
Egg number	Life history	eggs	86.28 (25.43)	36	50.52 (17.46)	23	-1.58	NA	NA	NA	NA
Standard length	Trophic, mating	cm	4.86 (0.84)	230	3.76 (0.58)	132	-1.45	F1,231 = 311.18	0.00001	NA	NA
Pelvic spine	Predation	cm	0.83 (0.13)	157	0.68 (0.12)	72	-1.14	F1,226 = 63.71	0.00003	F1,225 = 7.59	0.015
Testis weight	Mating	mg	0.78 (0.24)	53	0.55 (0.13)	31	-1.12	FNA,NA = NA	NA	NA	NA
2nd dorsal spine	Predation	cm	0.56 (0.14)	226	0.49 (0.09)	131	-0.55	F1,225 = 42.24	0.0001	F1,224 = 2.98	0.13
1st dorsal spine	Predation	cm	0.51 (0.16)	226	0.44 (0.12)	131	-0.52	F1,225 = 54.38	0.0002	F1,224 = 4.92	0.06
Body lightness	Mating?	Intensity	474.81 (38.91)	166	491.7 (49.99)	73	0.4	F1,232 = 7.7	0.01	F1,228 = 1.16	0.36
Long gill rakers	Trophic	Rakers	20.26 (1.5)	70	20.83 (1.42)	59	0.39	NA	NA	NA	NA
Armor plate count	Predation	Plates	31.41 (1.01)	70	31.02 (1.12)	59	-0.37	NA	NA	NA	NA
Egg diameter	Life history	mm	1.23 (0.28)	36	1.32 (0.17)	23	0.35	NA	NA	NA	NA
Short gill rakers	Trophic	Rakers	15.14 (1.09)	70	15.44 (0.93)	59	0.29	NA	NA	NA	NA
Egg weight	Life history	mg	2.83 (1.39)	36	2.53 (0.57)	23	-0.26	NA	NA	NA	NA
3rd dorsal spine	Predation	cm	0.16 (0.06)	160	0.15 (0.05)	72	-0.23	F1,229 = 2.64	0.1	F1,228 = 0.48	0.53
Body shape PC2	Trophic?	-	0 (0.01)	122	0 (0.01)	70	0.22	F1,189 = 10.44	0.0053	F1,188 = 8.91	0.0096
Body shape PC3	?	-	0 (0.02)	122	0 (0.01)	70	0.19	F1,189 = 14.69	0.00071	F1,188 = 0.23	0.65
Body shape PC6	?	-	0 (0.01)	122	0 (0.01)	70	-0.12	F1,189 = 4.36	0.0734	F1,188 = 9.34	0.0085
Body shape PC5	?	-	0 (0.02)	122	0 (0.01)	70	0.07	F1,189 = 3.07	0.13	F1,188 = 2.77	0.14
Body shape PC4	?	-	0 (0.01)	122	0 (0.01)	70	0	F1,189 = 3.99	0.082	F1,188 = 1.96	0.22

Appendix C - Chapter 4 Supplementary Material

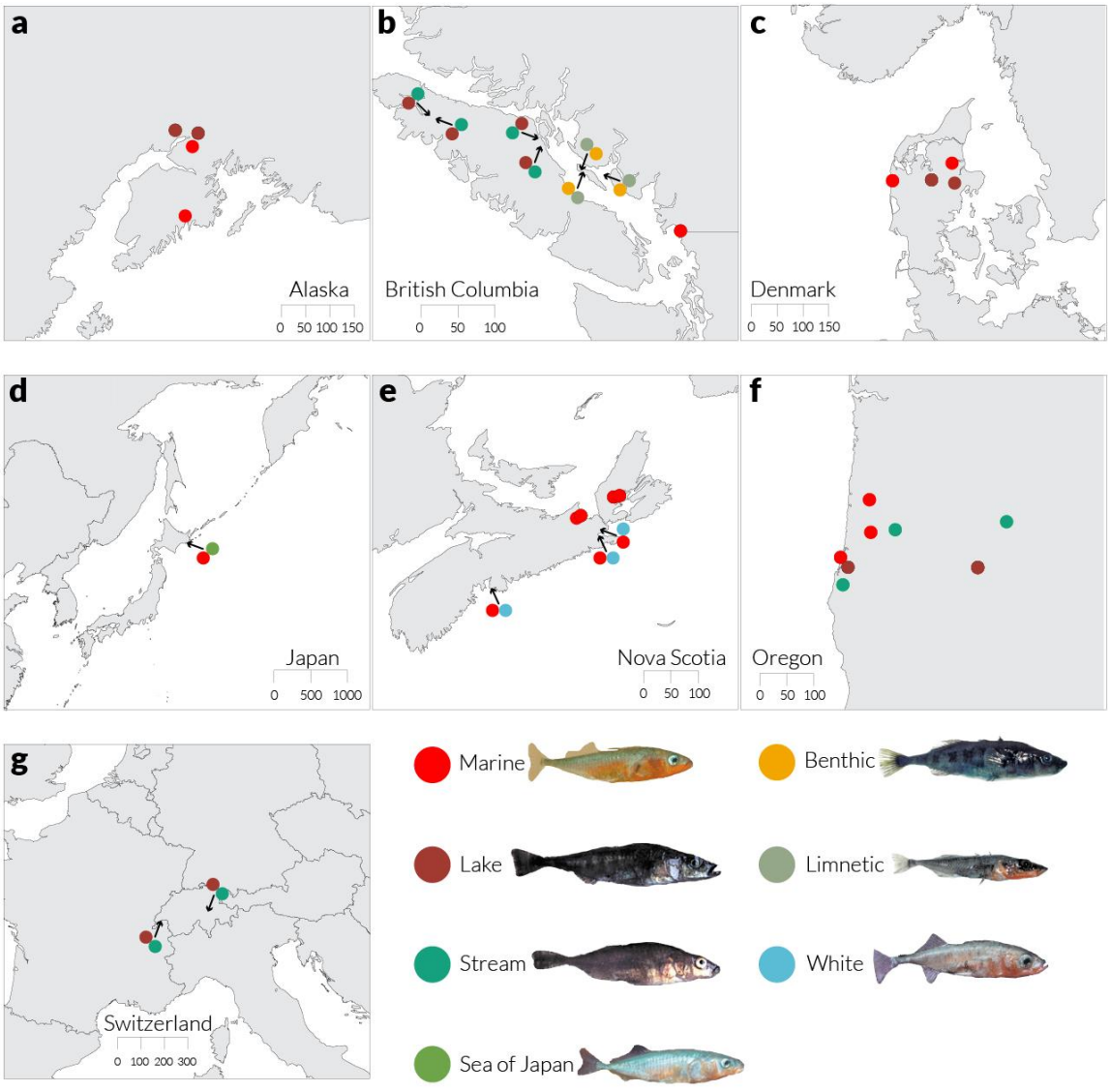

Figure C.1 | Collection locations of all stickleback populations used in the study**.** Ecotypes are color-coded. Arrows indicate locations where two ecotypes/populations are found in near or complete sympatry. Scale bars indicate distances in kilometers. Populations shown as inland (e.g. the Oregon populations, sixth panel) are found in lake or streams. Photographs obtained from previously published figures (Blouw & Hagen 1990; McKinnon & Rundle 2002). See Appendix

Table C.1 for further population sample information.

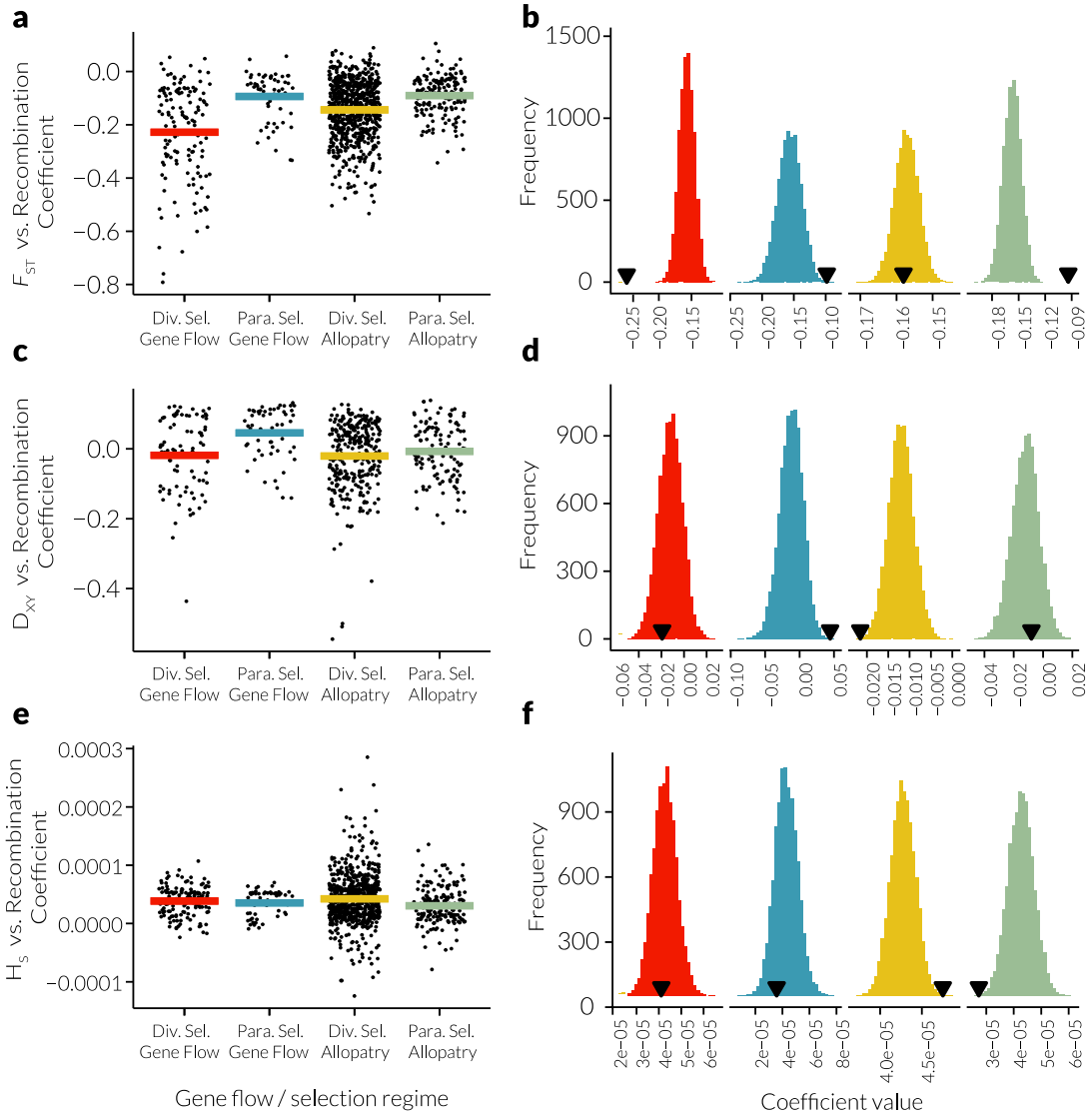

Figure C.2| Low recombination tendency estimates and permutation significance tests for three population genetic parameters across the four categories of population pairs, differing in gene flow and selection regimes. (a,c,e) Each dot represents a coefficient derived from a single pairwise comparison, measuring F_{ST}, D_{XY} or H_S (mean intra-population heterozygosity), with the colored line representing the category mean. (b,d,f) Null expectations (histograms) and observed values (black arrows) for permutation tests of the significance of the differences in mean recombination tendency seen in a, c, and e respectively. Observed means in the tails of the distributions indicate significance (see main text for *P* values).

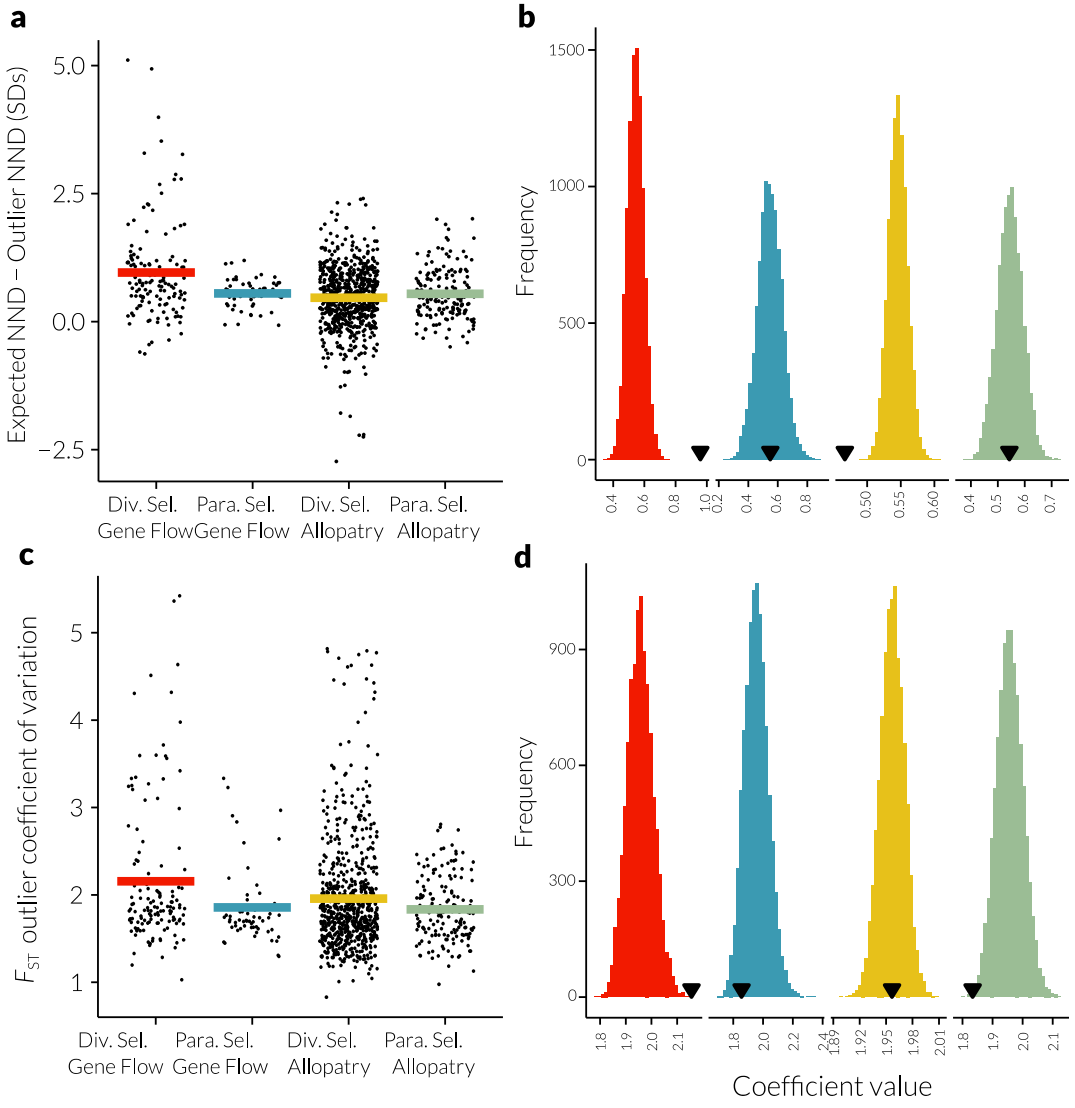

Figure C.3| Clustering estimates and permutation significance tests for two metrics of clustering across different gene flow / selection regimes. (a,c) Each dot represents a clustering metric averaged across all chromosomes in a single comparison, with the colored line representing the mean estimate for that regime. (b,d) Null expectations (histograms) and observed values (black arrows) for permutation tests of the significance of the differences in mean recombination tendency seen in a, c, and e respectively. Observed means in the tails of the distributions indicate significance (see main text for *P* values).

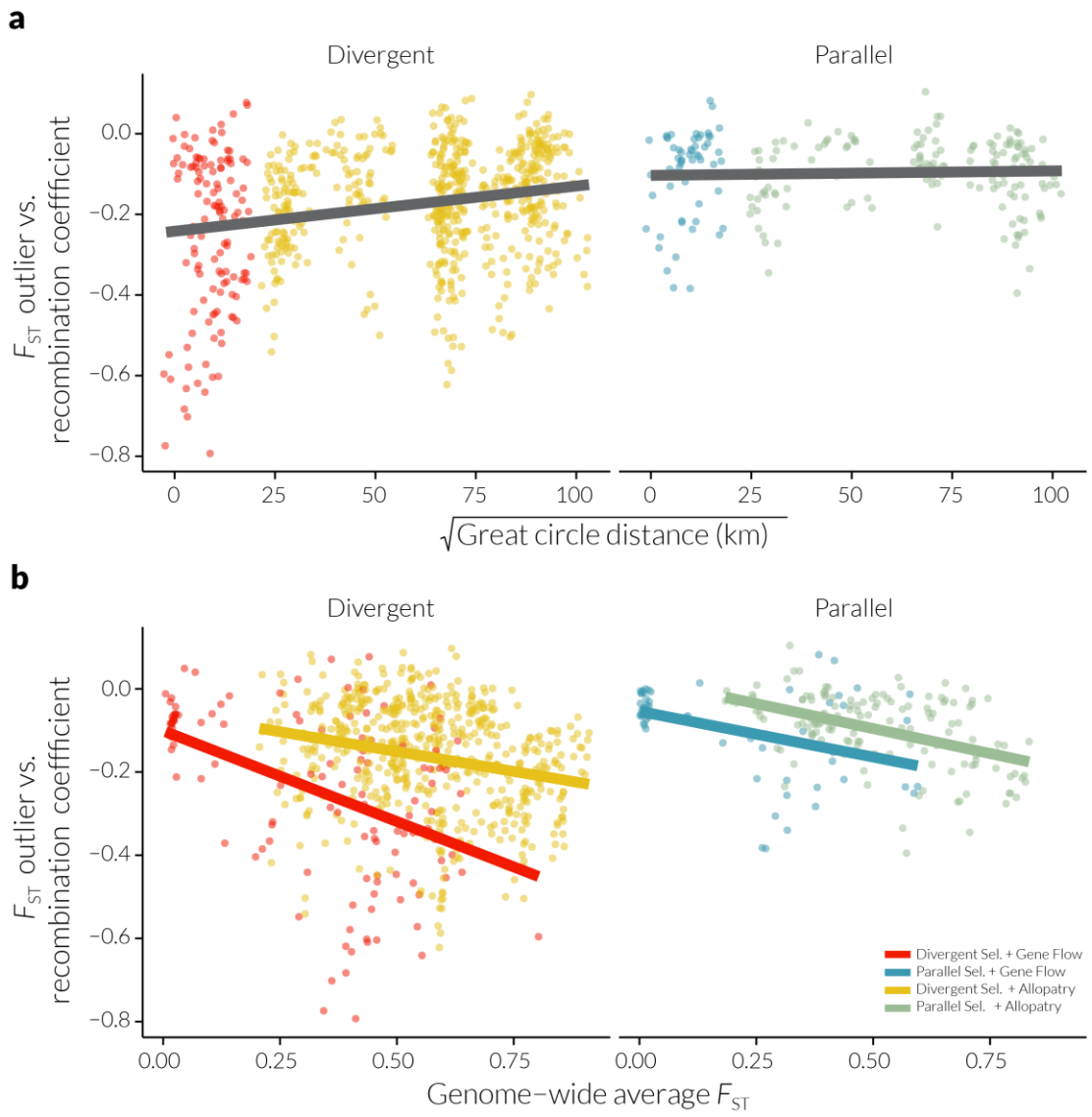

Figure C.4| Correlation between low recombination tendency (F_{ST} outliers) and **(a)** pairwise geographic distance or **(b)** overall genetic divergence (average genome-wide F_{ST}). Individual dots represent bias/F_{ST}/pairwise distance values from single comparisons. Points (a & b) and lines (b) are colored according to their gene flow / selection regime.

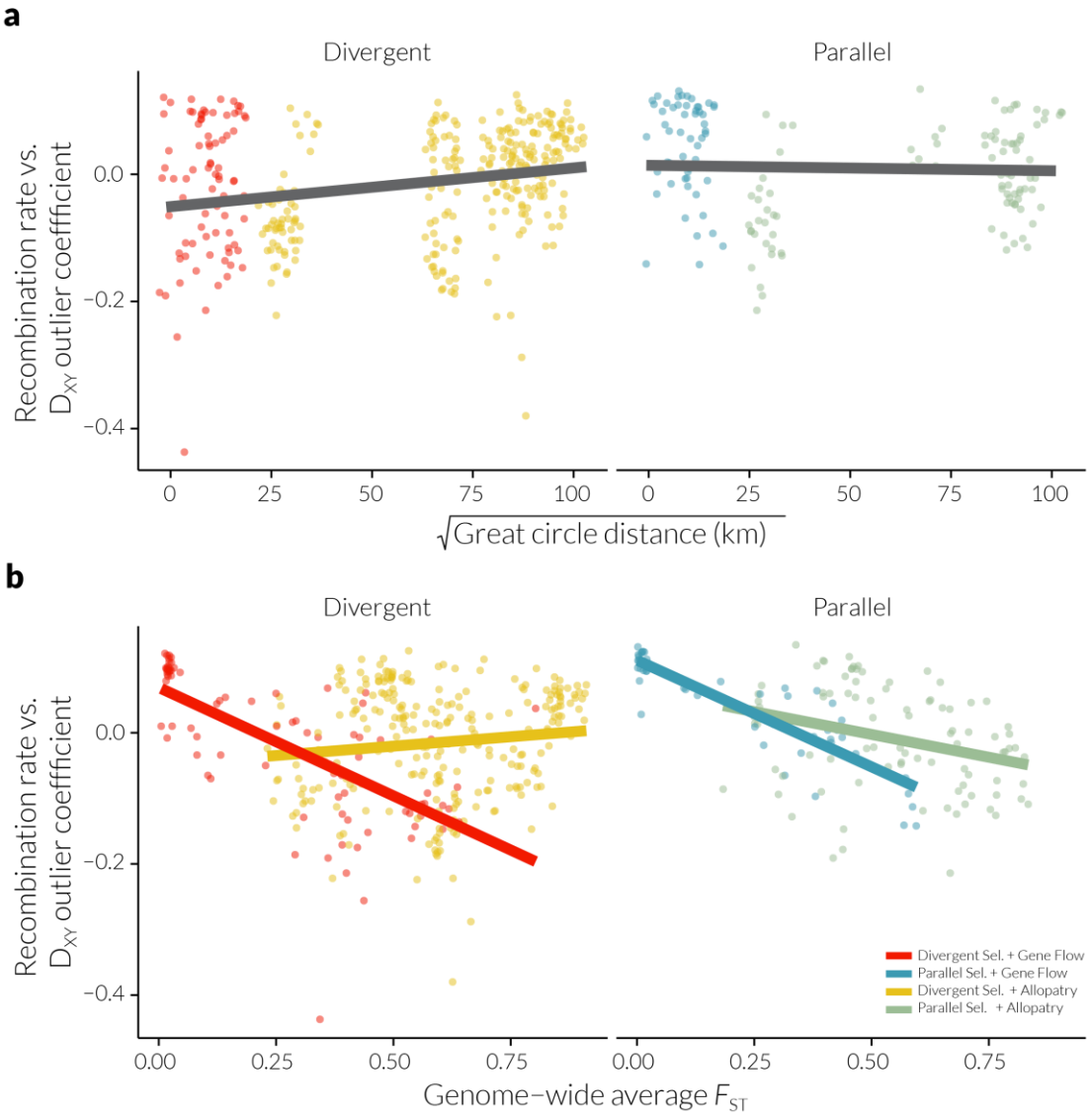

Figure C.5| Correlation between recombination rate tendency (D_{XY} outliers) and **(a)** pairwise geographic distance or **(b)** overall genetic divergence (average genome-wide F_{ST}). Individual dots represent tendency/F_{ST}/pairwise distance values from single comparisons. Points (a & b) and lines (b) are colored according to their gene flow / selection regimes.

Table C.1 | Collection locations, names and metadata for all samples included in the study. Citations for each study are noted for the first occurrence of the study only. Location abbreviations: Japan (JP), Oregon (OR), Nova Scotia (NS), British Columbia (BC), Europe (EU), Alaska (AL). SOJ refers to the Sea of Japan Stickleback. Data source abbreviations: Short Read Archive (SRA), the European Nucleotide Archive (ENA) and the Databank of Japan Sequence Read Archive (DRA). Sequencing technology abbreviations: Whole Genome Sequencing (WGS), Restriction Amplified Digest (RAD), Genotyping-by-Sequencing (GBS).

Study	Lat.	Long.	Reg.	Population Name	Ecotype	Source	Acc. No.	Technology	*n*
Yoshida[23]	43.054	144.894	JP	Japan	SOJ	DRA	DRA001136	WGS	8
Yoshida	43.054	144.894	JP	Japan	Marine	DRA	DRA001136	WGS	8
Catchen[11]	43.145	-124.190	OR	Winchester Creek	Stream	SRA	SRA070979	RAD	22
Catchen	43.424	-121.153	OR	Pony Creek Reservoir	Lake	SRA	SRA070979	RAD	68
Catchen	43.427	-121.153	OR	Paulina Lake	Lake	SRA	SRA070979	RAD	22
Catchen	43.430	-124.076	OR	South Twin Lake	Lake	SRA	SRA070979	RAD	50
Catchen	43.592	-124.243	OR	Cushman Slough	Marine	SRA	SRA070979	RAD	98
Catchen	44.000	-123.563	OR	South Jetty	Marine	SRA	SRA070979	RAD	96
Catchen	44.043	-123.012	OR	Riverbend	Stream	SRA	SRA070979	RAD	140
Catchen	44.172	-120.504	OR	Crooked River	Stream	SRA	SRA070979	RAD	24
Catchen	44.531	-123.593	OR	Millport Slough	Marine	SRA	SRA070979	RAD	68
Samuk	44.499	-63.903	NS	Canal Lake	Marine	UBC	NA	GBS	12
Samuk	44.499	-63.903	NS	Canal Lake	White	UBC	NA	GBS	15
Samuk	45.353	-61.473	NS	Salmon River Estuary	Marine	UBC	NA	GBS	14
Samuk	45.353	-61.473	NS	Salmon River Estuary	White	UBC	NA	GBS	17
Samuk	45.458	-61.612	NS	Milford Haven	Marine	UBC	NA	GBS	9
Samuk	45.458	-61.612	NS	Milford Haven	White	UBC	NA	GBS	7
Samuk	45.632	-61.960	NS	Antigonish Landing	Marine	UBC	NA	GBS	16
Samuk	45.672	-61.861	NS	Captain's Pond	Marine	UBC	NA	GBS	30
Samuk	45.970	-61.119	NS	Skye River	Marine	UBC	NA	GBS	15
Samuk	45.992	-60.985	NS	Little Narrows	Marine	UBC	NA	GBS	25
Jones[3]	49.013	-122.778	BC	Little Campbell River	Marine	SRA	PRJNA247503	WGS	5
Rennison	49.663	-124.109	BC	Little Quarry Lake	Benthic	UBC	NA	GBS	20
Rennison	49.663	-124.109	BC	Little Quarry Lake	Limnetic	UBC	NA	GBS	10
Rennison	49.709	-124.525	BC	Paxton Lake	Limnetic	UBC	NA	GBS	20
Rennison	49.709	-124.525	BC	Paxton Lake	Benthic	UBC	NA	GBS	20
Rennison	49.745	-124.566	BC	Priest Lake	Limnetic	UBC	NA	GBS	20
Rennison	49.745	-124.566	BC	Priest Lake	Benthic	UBC	NA	GBS	20
Roesti[24]	46.205	6.544	EU	Lake Geneva	Stream	SRA	SRP007695	RAD	27
Roesti	46.313	6.344	EU	Lake Geneva	Lake	SRA	SRP007695	RAD	27
Roesti	47.332	9.225	EU	Lake Constance	Lake	SRA	SRP007695	RAD	27
Roesti	47.333	9.164	EU	Lake Constance	Stream	SRA	SRP007695	RAD	27
Roesti	50.022	-125.336	BC	Boot Lake	Stream	SRA	SRP007695	RAD	27
Roesti	50.030	-125.323	BC	Boot Lake	Lake	SRA	SRP007695	RAD	26
Roesti	50.134	-125.331	BC	Roberts Lake	Lake	SRA	SRP007695	RAD	27
Roesti	50.143	-125.352	BC	Roberts Lake	Stream	SRA	SRP007695	RAD	27
Roesti	50.363	-127.156	BC	Misty Lake	Lake	SRA	SRP007695	RAD	27
Roesti	50.365	-127.322	BC	Joes Lake	Stream	SRA	SRP007695	RAD	26
Roesti	50.366	-127.170	BC	Misty Lake	Stream	SRA	SRP007695	RAD	27
Roesti	50.373	-127.291	BC	Joes Lake	Lake	SRA	SRP007695	RAD	27
Feulner[25]	56.369	8.182	EU	Atlanic Ocean	Marine	ENA	PRJEB2954	WGS	6
Ferchaud[26]	56.330	10.048	EU	Hadsten Lake	Lake	SRA	SRX437379	RAD	20
Ferchaud	56.383	9.354	EU	Hald Lake	Lake	SRA	SRX437379	RAD	20
Ferchaud	56.663	9.969	EU	Mariager	Marine	SRA	SRX437379	RAD	20
Hohenlohe[27]	60.127	-149.406	AL	Resurrection Bay	Marine	SRA	SRP001747	RAD	20
Hohenlohe	61.330	-149.151	AL	Rabbit Slough	Marine	SRA	SRP001747	RAD	16
Hohenlohe	61.563	-148.949	AL	Mud Lake	Lake	SRA	SRP001747	RAD	19
Hohenlohe	61.614	-149.756	AL	Bear Paw Lake	Lake	SRA	SRP001747	RAD	28