THE GENOMIC BASIS FOR FITNESS AND ECOMORPHOLOGICAL VARIATION IN RECOVERING POPULATIONS OF LAKE TROUT (SALVELINUS NAMAYCUSH) IN THE GREAT LAKES

By

Seth Robert Smith

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Integrative Biology—Doctor of Philosophy
Ecology, Evolutionary Biology & Behavior — Dual Major

2021

ABSTRACT

THE GENOMIC BASIS FOR FITNESS AND ECOMORPHOLOGICAL VARIATION IN RECOVERING POPULATIONS OF LAKE TROUT (SALVELINUS NAMAYCUSH) IN THE GREAT LAKES

By

Seth Robert Smith

Here I describe the development of novel genomic resources that will be fundamental for advancing a new generation of genomic research on Lake Trout (*Salvelinus namaycush*) including a high-density linkage map, an annotated chromosome-anchored genome assembly, and three high-throughput genotyping panels. We used these resources to identify genomic regions exhibiting signals of adaptive divergence between Lake Trout hatchery strains, some of which were found to underlie differences in fitness (survival and reproduction) between strains in the contemporary Lake Huron environment. Loci associated with differences in fitness between Seneca and Great Lakes origin strains were localized using local ancestry inference and local ancestry outlier tests. By evaluating locus specific allelic contributions of the ancestral Seneca Lake and Great Lakes-derived hatchery strains to naturally-produced wild Lake Huron populations across the genomes of F2 wild born individuals, we were able to determine that a subset of 7 genomic regions contributed to differences in fitness between Seneca Lake and Great Lakes origin individuals during the re-emergence of wild populations in Lake Huron. We also identified multiple regions associated with elevated Great Lakes strain fitness, hybrid vigor, and hybrid inferiority. These admixture outlier regions contained a significant excess of genes related to swimming behavior and negative regulation of vascular wound healing, which strongly suggests that differences in fitness between strains are due to behavioral and physiological

factors associated with the ability to avoid and survive predation by Sea Lamprey. Additionally, we carried out two studies seeking to identify genetic variation associated with habitat occupancy and phenotypic variation in Lake Trout. First, we carried out a quantitative trait locus (QTL) mapping study in which we identified loci associated with length and condition related traits, skin pigmentation, and body shape. We produced a linkage map for Lake Trout as a prerequisite for this work. The information on locus order obtained from the linkage map was also used to produce a chromosome anchored genome assembly for Lake Trout. This study also allowed us to determine the location of the Lake Trout sex determination locus, determine centromere locations, and characterize structural differences (i.e., chromosomal inversions and translocations) between Lake Trout and other salmonid species. Second, we performed a genome-wide scan for loci associated with ecomorphological divergence in Lake Superior Lake Trout (specifically between lean, siscowet, and humper forms), and identified numerous regions with abnormally high levels of divergence between forms. These loci likely underlie variation in traits that differentiate forms, as well as traits that contributed to reproductive isolation historically. For example, the genomic region most strongly associated with length and condition (from our QTL mapping study) was also associated with ecomorphological divergence in Lake Superior and this region also contains a putative chromosomal inversion. Interestingly, we find that hybridization primarily occurred between humpers and siscowets and humpers and leans immediately preceding a genetic homogenization event that occurred in the late 1990s or early 2000s. Using a collection of samples over a multi-decade time series collected from the Apostle Islands, we show that levels of hybridization with humpers increased substantially starting in the 1980s.

This dissertation is dedicated to my wife Annie– my best friend; most helpful critic; and an endless source of motivation, ideas, and excitement about science. I would also like to thank my parents, Doug and Sheila, for always encouraging me to think and teaching me that mountains, rivers, trees, and fish are something worth caring about.

ACKNOWLEDGEMENTS

I would like to acknowledge the training, assistance, guidance, and collaboration provided by a number of people. First, Stephen Amish – who helped channeled my excitement about genomics and conservation into a set of meaningful bioinformatic, analytic, and laboratory skills. I would also like to thank those who collaborated with me on the research contained herein. These people include Louis Bernatchez, Eric Normandeau, Chris C. Wilson, Chantelle Penney, Haig Djambazian, Pubudu M. Nawarathna, Pierre Berube, Andrew Muir, Jiannis Ragoussis. Jeremy Le Luyer, and Olivia Boeberitz. I would also like to thank my guidance committee members: Mariah Meek, Ingo Braasch, and Juan Steibel. Finally, I would like to thank Gordon Luikart and Kim Scribner for supporting me throughout this endeavor and providing me with ample opportunities to pursue my research interests.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Lake Trout (*Salvelinus namaycush*) were historically an abundant top predator in the Great Lakes of North America and represent an important component of the native aquatic community and economy of the Great Lakes region (Zimmerman and Krueger 2009). Lakes were inhabited by multiple Lake Trout forms (referred to as morphotypes, ecomorphotypes, or ecotypes hereafter) with distinct morphology, physiology, and behaviors that allowed for the exploitation of resources and habitats at varying depths (Muir et al., 2014; Muir et al., 2016). These forms include lean, siscowet, and humper Lake Trout, which are known for inhabiting shallow water habitats, deep-water habitats, and off-shore shoals, respectively. Forms varied with respect to multiple traits including, swim bladder morphology, fin size, tissue lipid content, pigmentation, body shape, spawning time, age at maturity, visual acuity, and patterns of diel migration (Muir et al., 2014, Muir et al., 2016l Zimmerman et al., 2006; Zimmerman et al., 2007; Harrington et al., 2015; Moore & Bronte, 2001; Khan & Qadri, 1970; Hansen et al. 2016; Sitar et al., 2008; Ahrenstorff et al., 2011; Burnham-Curtis & Bronte, 1996; Rahrer, 1965)

Lake Trout in the Great Lakes experienced severe reductions in abundance, distribution, and ecomorphological diversity over the course of the 20th century which were primarily the result of overfishing and high levels of predation by invasive Sea Lamprey (Petromyzon marinus; Hansen, 1999). Lake Trout were ultimately extirpated from Lakes Michigan, Ontario, and Erie. A single isolated population of lean Lake Trout avoided extirpation in Georgian Bay in eastern Lake Huron. Populations of lean, siscowet, and humper Lake Trout in Lake Superior also experienced severe reductions in abundance (Hansen, 1999). Ecomorphological variation was lost from all lakes except Lake Superior.

Regional population collapse was followed by an intensive recovery effort that primarily focused on controlling lamprey populations, creating aquatic refuges, and stocking juvenile lean Lake Trout originating from multiple domesticated hatchery populations (Muir et al., 2012). Hatchery populations used for supplementation were initially derived from extant populations of lean Lake Trout from Lake Superior (Isle Royale, Apostle Islands, and Marquette Strains), populations with Lake Michigan ancestry that were planted in other locations prior to population collapse (Green Lake and Lewis Lake Strains), and brood stock derived from the Lake Trout population in Seneca Lake, New York (the Seneca Strain; Page et al., 2003). Populations in Lake Superior rebounded relatively quickly and multiple trophic ecomorphotypes still exist in this locale; however, levels of phenotypic and genetic variation are declining between extant ecomorphotypes in Lake Superior, potentially due to increased levels of hybridization (Baillie et al. 2016; Muir et al. 2014).

Restoration proceeded more slowly in other Great Lakes. Evidence for the re-emergence of natural recruitment was first observed in the U. S waters of Lake Huron starting in the early 1990s (Riley et al. 2007; Scribner et al., 2008) and Lake Michigan in 2005 (Hansen et al., 2003). Evidence for reproduction in the wild was also observed in Lake Ontario more recently (Gatch et al., 2021). Given that multiple strains were stocked in these locales, recent studies have sought to determine whether certain strains are contributing disproportionately to wild recruitment in certain locales. For instance, work by Scribner et al. (2018) found that the Seneca Lake hatchery strain, originating from the Finger Lakes in New York, contributed disproportionately to natural recruitment in Lake Huron. Additionally, recent work by Larson et al. (2021) came to a similar conclusion after

evaluating strain contributions to wild recruitment in Lake Michigan. Although the re-emergence of natural recruitment is an important milestone in population recovery, the re-establishment of viable, productive and diverse populations of Great Lakes Lake Trout requires an improved understanding of the genetic basis for variation in survival and reproductive success within these recovering populations. The recovery of self-sustaining populations of Lake Trout is a critical federal, state, tribal and international management goal (Muir et al., 2012).

Until recently, Great Lakes Lake Trout restoration programs have exclusively made use of lean-form brood stocks for reintroduction. Lean Lake Trout are only able to utilize a small proportion of habitats and niches that Lake Trout historically occupied (Edsall and Kennedy 1995) and the continued absence of siscowet and humper Lake Trout is a direct impediment to the long-term goal of restoring the deep-water food web in the Great Lakes. A strain of humper Lake Trout, originating from Klondike Reef in Lake Superior, has been stocked in Lake Michigan in recent years and proposals have been made to reintroduce siscowet Lake Trout to locations outside of Lake Superior (Muir et al., 2012; Kornis et al., 2019). The success of these endeavors will require an improved understanding of the biological basis for ecomorphological variation and the genetic and environmental factors that contributed to genetic homogenization among extant Lake Superior ecomorphotypes during the 1990s and early 2000s (Baillie et al. 2016).

Genomic methods could help to address multiple important Lake Trout management questions, uncover the genetic basis for ecomorphological variation, and identify genomic regions associated with variation in fitness in re-emerging wild populations; however, progress is hindered by a lack of genomic resources for the species.

The work described herein documents the creation of a research system that will be broadly useful for addressing questions related to Lake Trout ecology, evolution, and conservation in the Great Lakes and elsewhere. The components of this research system include a linkage map, a physical genomic map, and multiple genotyping panels for Lake Trout that were used to address two fundamental questions related to Lake Trout conservation and restoration in the Great Lakes. First, 1) What is the genetic basis for variation in fitness (survival and reproductive success) between Lake Trout hatchery strains that were used to restore Great Lakes populations? Specifically, can elevated performance of the Seneca strain be attributed to adaptive differences between this strain and others? If so, which genes and biological processes are involved? Additionally, are certain loci associated with hybrid vigor or outbreeding depression in re-emerging wild populations? Second, 2) What is the genetic basis for phenotypic and ecomorphological variation in Great Lakes Lake Trout? How many loci are associated with adaptive differences between ecomorphotypes and how are these loci distributed across the genome?

In Chapter 1, I describe the creation of a high-density linkage map for Lake Trout containing 15,740 restriction site associated DNA (RAD) markers. Linkage mapping uses the observed distribution of 2-locus genotypes within families to infer recombination frequencies between loci (Rastas, 2017). Given that the frequency of recombination between loci is a function of the physical distance separating them, this information on recombination is valuable for assigning loci to linkage groups (e.g., chromosomes) and determining the relative order of loci along chromosomes. Linkage maps have been foundational to the creation of most model systems in the field of genomics (e.g., Humans,

Murray et al., 1994; Mice, Copeland et al., 1993; Zebrafish, Postlethwait et al., 1994; Medaka, Wada et al., 1995; Three-Spine Stickleback, Peichel et al., 2001) and recent technological advances now allow for the creation of high-density linkage maps for non-model species at comparatively low cost (Baird et al., 2008; Miller et al., 2012). After generating a linkage map for Lake Trout, I use this resource to localize quantitative trait loci (QTL) associated with traits that vary within and between ecomorphotypes in the Great Lakes. These include multiple ecologically relevant traits including body shape, growth and condition related traits, and traits related to skin pigmentation. This resource was also used determine the approximate location of the Lake Trout sex determination locus and identify structural genetic differences between Lake Trout and other salmonids. Half tetrad analysis was also used to determine centromere locations for 41 of 42 Lake Trout chromosomes.

In Chapter 2, I use a combination of long read sequencing, short read sequencing, chromatin conformation capture (Hi-C), and the previously constructed linkage map to produce a chromosome-level genome assembly for Lake Trout. Self-vs-self synteny analysis was also performed for the purpose of identifying homeologs resulting from the Salmonid Specific Autotetraploid genome duplication event and characterizing levels of sequence divergence between these duplicated loci across the genome. Synteny analysis was also performed between the Lake Trout genome and genomes for closely related taxa for the purpose of identifying structural genomic differences between Lake Trout and other salmonids. A Lake Trout specific repeat library, genome annotation, and interpolated recombination map were also generated. The genome assembly and these resources will be foundational to all future genomic research on Lake Trout and enabled analyses conducted

in Chapters 3 and 4. Most importantly, the availability of a standardized physical map of the Lake Trout genome will greatly simplify the process of comparing results from genome-wide association analyses, scans for evidence of selection, and genotype-environment association analyses across studies.

In Chapter 3, I utilize the genome assembly to characterize the genetic architecture of ecomorphological divergence in Lake Superior Lake Trout and explore patterns of gene flow and population structure during the decades preceding the genetic homogenization event documented by Baillie et al., (2016). I find evidence for an increase in rates of hybridization between lean and humper and siscowet and humper Lake Trout during the 1980s and 1990s, suggesting that homogenization was initially driven by hybridization between these forms. Additionally, I identify multiple islands of divergence that are likely associated with adaptive differences or reproductive isolation mechanisms between ecomorphotypes and determine that some of these regions are likely associated with structural variants.

In Chapter 4, I use the genome assembly and an interpolated recombination map to determine the ancestral origins of haplotype segments within wild F2 hybrid Lake Trout born in Lake Huron. The distribution of haplotype ancestries and runs of hybridity across the genomes of these individuals were used to identify genomic regions associated with differences in fitness between strains, hybrid vigor, and outbreeding depression in the contemporary Lake Huron environment. Signals of adaptive differentiation between hatchery strains were also identified. We then evaluated whether or not adaptively diverged genes within fitness associated regions were disproportionately associated with biological processes related to the ability to survive or avoid lamprey predation.

In Chapter 5, I describe two new genotyping panels for Lake Trout that will be useful for routine Lake Trout management and monitoring activities including genetic stock identification, parentage assignment, estimation of individual inbreeding coefficients, estimation of the effective number of breeders, evaluation of population structure, the identification of inter-strain and inter-morphotype hybrids, and mixed stock assessments.

Overall, this work provides resources that will be foundational to future genomic research on Lake Trout, improves our understanding of the genetic basis for phenotypic and ecomorphological variation in Lake Trout, and sheds light on the biological basis for variation in fitness between hatchery strains that were used to restore Great Lakes populations. Of particular importance, the availability of a standardized, publicly available, genome assembly for Lake Trout will allow the results presented here to be easily compared with those of future studies; will allow researchers to address qualitatively and quantitatively unique questions related to Lake Trout ecology, evolution, and conservation; and will likely quicken the pace of scientific discovery.

CHAPTER 1: MAPPING OF ADAPTIVE TRAITS ENABLED BY A HIGH-DENSITY LINKAGE MAP FOR LAKE TROUT (SALVELINUS NAMAYCUSH)

ABSTRACT

Understanding the genomic basis of adaptative intraspecific phenotypic variation is a central goal in conservation genetics and evolutionary biology. Lake Trout (Salvelinus namaycush) are an excellent species for addressing the genetic basis for adaptive variation because they express a striking degree of ecophenotypic variation across their range; however, necessary genomic resources are lacking. Here we utilize recently-developed analytical methods and sequencing technologies to (1) construct a high-density linkage and centromere map for Lake Trout, (2) identify loci underlying variation in traits that differentiate Lake Trout ecophenotypes and populations, (3) determine the location of the Lake Trout sex determination locus, and (4) identify chromosomal homologies between Lake Trout and other salmonids of varying divergence. The resulting linkage map contains 15,740 single nucleotide polymorphisms (SNPs) mapped to 42 linkage groups, likely representing the 42 Lake Trout chromosomes. Female and male linkage group lengths ranged from 43.07 to 134.64 centimorgans, and 1.97 to 92.87 centimorgans, respectively. We improved the map by determining coordinates for 41 of 42 centromeres, resulting in a map with 8 metacentric chromosomes and 34 acrocentric or telocentric chromosomes. We use the map to localize the sex determination locus and multiple quantitative trait loci (QTL) associated with intraspecific phenotypic divergence including traits related to growth and body condition, patterns of skin pigmentation, and two composite geomorphometric variables quantifying body shape. Two QTL for the presence of vermiculations and spots mapped with high certainty to an arm of linkage group Sna3,

growth related traits mapped to two QTL on linkage groups Sna1 and Sna12, and putative body shape QTL were detected on six separate linkage groups. The sex determination locus was mapped to Sna4 with high confidence. Synteny analysis revealed that Lake Trout and congener Arctic char (*Salvelinus alpinus*) are likely differentiated by three or four chromosomal fissions, possibly one chromosomal fusion, and 6 or more large inversions. Combining centromere mapping information with putative inversion coordinates revealed that the majority of detected inversions differentiating Lake Trout from other salmonids are pericentric and located on acrocentric and telocentric linkage groups. Our results suggest that speciation and adaptive divergence within the genus Salvelinus may have been associated with multiple pericentric inversions occurring primarily on acrocentric and telocentric chromosomes. The linkage map presented here will be a critical resource for advancing conservation oriented genomic research on Lake Trout and exploring chromosomal evolution within and between salmonid species.

INTRODUCTION

Maintaining adaptive phenotypic diversity is a central tenet of conservation biology. In many taxa, diversity is produced through selective pressures that favor reduced intraspecific competition and trophic specialization (Skúlason and Smith 1995; Robinson and Schluter 2000; Whiteley 2007). The evolution of trophically specialized morphotypes has been observed in multiple fish species including Arctic char (Snorrason et al. 1994), Lake Trout (Eschenroder 2008; Muir et al. 2016), multiple coregonid species (Lu and Bernatchez 1999; Thomas et al. 2019), and African cichlids (Ruber et al. 1999), and represents an important pathway by which phenotypic diversity is generated and maintained in nature (Pfennig and Pfennig 2012). Intraspecific diversity can promote

community and ecosystem stability (Schindler et al. 2010); however, the genomic basis for this variation is often poorly understood for non-model species. Advancement of our understanding is largely limited by a lack of genomic resources.

Lake Trout (*Salvelinus namaycush*) are a salmonid fish species endemic to North America with substantial cultural, ecological, and economic importance. Across their range, Lake Trout are often the keystone predator of lentic ecosystems (Ryder et al. 1981) and historically supported valuable commercial and subsistence fisheries (Waters 1987; Hansen 1999; DFO 2012; Brenden et al. 2013). Lake Trout express a large degree of sympatric phenotypic variation (Muir et al. 2016) making them a useful species for exploring the genomic basis for phenotypic diversity. Multiple morphotypes exist across the species range (Muir et al. 2016; Marin et al. 2017), with diversification largely associated with the ability to exploit resources and habitats at varying depths in large post-glacial lakes (Zimmerman et al. 2006; Stafford et al. 2013; Muir et al. 2014; Marin et al. 2017). In the Great Lakes, trophic specialization has resulted in the evolution of three widely recognized morphotypes — leans, siscowets, and humpers — that are differentiated by patterns of skin pigmentation, size-at-age, body shape, tissue lipid content, habitat use, and diet (Thurston 1962; Eschmeyer and Phillips 1965; Burnham-Curtis 1994; Harvey et al. 2003; Alfonso 2004; Zimmerman et al. 2007; Zimmerman et al. 2009; Goetz et al. 2013). Similar patterns of divergence exist in other Lake Trout populations (Blackie et al. 2003; Zimmerman et al. 2006; Hansen et al. 2012; Marin et al. 2016; Chavarie et al. 2015), with some degree of morphological and phenological variation existing among individuals of the same morphotype (Bronte 1993; Bronte and Moore 2007).

Previous studies have evaluated differences in gene expression and signals of adaptive divergence between Lake Trout morphotypes (Goetz et al. 2010; Bernatchez et al. 2016; Perreault-Payette et al. 2017). However, no study has explicitly evaluated which loci control variation in specific traits that underly morphotype divergence. Additionally, these studies have relied on de novo assembled markers distributed anonymously across the genome. Although these approaches can be powerful (Davey et al. 2013), fully interpreting results requires some knowledge of how loci are ordered along chromosomes. All scans for adaptively significant loci and genotype-phenotype associations inherently take advantage of linkage disequilibrium between genotyped markers and causal loci. Without knowing the relative locations of loci, it can be difficult to determine if genotype-phenotype associations or signals of selection are associated with a single genomic region or multiple regions distributed widely across the genome. Information on the order of loci along chromosomes can be readily attained via linkage mapping or assembly of a reference genome; however, linkage maps are often needed a priori to produce chromosome-scale genome assemblies.

Linkage maps have been used to map loci associated with disease resistance (Houston et al. 2008; Moen et al. 2009), life history and physiological trait variation (Rogers et al. 2007; Miller et al. 2012; Gagnaire et al. 2013a; Sutherland et al. 2017; Pearse et al. 2019), and commercially valuable traits (Gonzalez-Pena et al. 2016) in salmonids and have been instrumental in the assembly of salmonid reference genomes (Lien et al. 2016, Christensen et al. 2018a, 2018b; Pearse et al. 2019; Sävilammi et al. 2019). A linkage map for Lake Trout would enable the application of cutting-edge genomic tools to questions in Lake Trout management and evolution and would aid in the identification of loci

underlying phenotypic variation and local adaptation. Specifically, a linkage map would increase the strength of inference from genome-wide association studies and scans for selection (Bradbury et al. 2013; Gagnaire et al. 2013b; McKinney et al. 2016) and allow for the localization of quantitative trait loci (Peichel et al. 2001; Qiu et al. 2018) and tracts of admixture and homozygosity, and the estimation of historical effective population sizes and admixture dynamics (Hollenbeck et al. 2016; Leitwein et al. 2018). This information would be valuable for selecting stocks for reintroduction and translocation and for estimating the adaptive potential of intact populations under changing climate and abiotic conditions (Leitwein et al. 2016; Bay et al. 2017).

Comparative analysis of linkage maps and genome assemblies from related species can also shed light on chromosomal evolution and speciation (Rastas et al. 2015; Sutherland et al. 2016; Hale et al. 2017). Chromosomal inversions appear to have played an important role in speciation and adaptive divergence within the salmonid lineage (Miller et al. 2012; Sutherland et al. 2016, Pearse et al. 2019) and within other taxa (Lowry and Willis 2010; Aylala et al. 2013; Kupper et al. 2016; for review see Wellenruether and Bernatchez 2018). Instances of reduced hybrid fitness and hybrid inviability are widespread within the family Salmonidae (Leary et al, 1993; Fugiwara et al. 1997; Muhlfeld et al. 2009). Information on the locations of inversions differentiating species and phenotypically divergent populations could shed light on the genetic basis for these phenomena. Inversions can contribute to isolation between species and populations because they can suppress recombination over large chromosomal regions, allowing for adaptive differences to accumulate between inverted and non-inverted haplotypes even in the presence of gene flow (Berg et al. 2017; Wellenruether and Bernatchez 2018). Inversions can also produce

post-zygotic isolation between incipient species if crossing over within heterozygous individuals results in formation of abnormal or inviable gametes (Wellenreuther and Bernatchez 2018). An improved understanding of the extent to which pericentric (including the centromere) and paracentric (outside the centromere) inversions can accumulate between salmonid species over varied evolutionary time scales, could provide clues about pre- and post-zygotic isolation mechanisms that contributed to adaptive divergence and incipient speciation within salmonids.

Linkage maps have been constructed for multiple salmonid species including rainbow trout (Miller et al. 2012; Palti et al. 2012; Gonzalez-Pena et al. 2016), chinook salmon (Brieuc et al. 2014; McKinney et al. 2016; McKinney et al. 2019), coho salmon (Kodama et al. 2014), sockeye salmon (Everett, Miller and Seeb 2012; Larson et al. 2015; Limborg et al. 2015), chum salmon (Waples et al. 2016); pink salmon (Spruell et al 1999; Lindner et al. 2000), Atlantic salmon (Moen et al. 2008; Lien et al. 2011; Brenna-Hansen et al. 2012; Gonen et al. 2014), Arctic char (Nugent et al. 2017; Christensen et al. 2018a), brook trout (Sauvage et al. 2012; Sutherland et al. 2016; Hale et al. 2017), brown trout (Leitwein et al. 2017), European grayling (Sävilammi et al. 2019), lake whitefish (Rogers et al. 2007; Gagnaire et al. 2013a), and European whitefish (De-Kayne et al. 2018). No linkage map has been constructed for Lake Trout (but see May et al. 1979, Johnson et al. 1987, for work on segregation patterns in Lake Trout x brook trout hybrids), although the Lake Trout karyotype has been characterized in multiple previous studies (Phillips and Zajicek 1982; Reed and Phillips 1995) providing a reference for the number of expected chromosomes.

Here we present a high-density linkage map for Lake Trout generated using restriction site associated DNA (RAD) capture (Rapture; Ali et al. 2016), a modified RAD sequencing protocol that allows variable loci to be preferentially genotyped. The map was used to characterize the Lake Trout karyotype, estimate recombination rates, determine centromere locations, map the sex determination locus, and identify chromosomal inversions and translocations differentiating Lake Trout from other salmonids. We demonstrate the utility of the linkage map by using available phenotype data to map quantitative trait loci (QTL) associated with pigmentation patterns, growth and condition related traits, and variation in body shape — all traits hypothesized to be adaptive in Lake Trout and other salmonids.

MATERIALS AND METHODS

LINKAGE MAPPING FAMILIES

Two F1 full-sibling families were created by crossing Seneca Lake hatchery strain females with Parry Sound strain males (Table 1.1, Figure 1.1). The Seneca Lake strain was founded using individuals from Seneca Lake, New York and this strain has contributed disproportionately to restoring Lake Trout populations in the Great Lakes (Scribner et al. 2018). The Parry Sound strain was founded by wild individuals collected from Georgian Bay in Lake Huron. The Seneca and Parry Sound strains are genetically divergent (FST = 0.089) based on a previous study using microsatellites (Scribner et al. 2018). Crosses were produced in 2017 using adult Lake Trout and housed at Pendills Creek National Fish Hatchery (U.S. Fish and Wildlife Service, Figure 1.1). Eggs were fertilized, incubated in Heath trays at ambient temperature, and raised until swim-up phase. Offspring were then euthanized using a lethal dose of MS-222 and preserved in 95% ethanol. Genetic sex was

determined for offspring using a sdY presence-absence quantitative PCR (qPCR) assay designed using the approach of Anglès d'Auriac et al. (2014; see Trait Mapping methods below). These families were ultimately used for constructing the linkage map and localizing the Lake Trout sex determination locus.

An additional F2 half-sibling family was created using adult Lake Trout from the Killala Lake hatchery strain and wild individuals from Kingscote Lake, Ontario (Table 1.1, Figure 1.1). The Killala Lake strain was founded by individuals from Killala Lake, Ontario, which is within the Lake Superior drainage. This hatchery strain is most similar to lean form hatchery strains derived from Lake Superior based on a previous allozyme genotyping study (Marsden et al. 1993). Individuals from the Kingscote Lake strain also resemble lean Lake Trout; however, they are small bodied and lack spots and vermiculations (Wilson and Evans 2010). Examination of F2 offspring at age 3 revealed substantial variation in pigmentation, weight and length at age, and body shape among individuals. These traits are commonly recognized as being adaptively differentiated between Lake Trout populations and ecophenotypes (Eshenroder 2008; Muir et al. 2016). Body shape and early growth rate in particular have been recognized as important traits for differentiating lean, siscowet, and humper ecophenotypes (Moore and Bronte 2001; Hansen et al. 2016). The observation that skin pigmentation patterns vary between ecophenotypes and across depth strata in some Lake Trout populations also suggests that pigmentation traits might be an important axis of ecophenotypic divergence within Lake Trout (Zimmerman et al. 2006). The F2 Kingscote x Killala family was used for linkage map construction, localization of the sex determination locus, and QTL mapping. Crosses, culture conditions, and phenotyping procedures are described below.

Initial Kingscote x Killala F1 crosses were produced using adult Lake Trout using a 2x2 factorial mating design. In 2012, mature adults from initial crosses were mated to produce F2 families. Eggs from each family were incubated in Heath trays at ambient temperature (2-5oC). Prior to swim-up, hatched sac fry from families were transferred to 36L laundry tubs (200 fry per tub) where they remained until age 1+. Families were manually fed 1% of tank biomass twice-daily and family sizes were periodically reduced by culling to avoid overcrowding. At age 1+, families were transferred to 700L circular tanks with ambient lighting and fed to satiation on an EWOS pellet diet. At age 3, fork length and weight were determined and lateral photographs were collected using the protocol from Bernatchez et al. (2016). Fish were photographed using a Nikon Coolpix P7700 digital camera with a focal length of 50mm mounted on a tripod in fixed position. Fish were photographed with the head facing to the left and were cradled in a stretched mesh net as in Zimmerman et al. (2006) in order to avoid distorting body shape. Fin clips were collected and preserved in 95% ethanol. Photographs were later used for morphometric analysis and scoring individuals for presence-absence of spots and vermiculations (see Trait Mapping methods section below).

An additional gynogenetic diploid family was created using a female F1 resulting from initial Kingscote x Killala crosses using a protocol similar to that of Thorgaard et al. (1983). This family was used for mapping centromeres using half-tetrad analysis (Thorgaard et al. 1983; Limborg et al. 2016). Sperm from a male Lake Trout was diluted 10:1 using sperm extender (9.2 g Tris buffer, 1.05 g citric acid, 4.81 g glycine, 2.98 g KCl, 100g PVP-40, and 1 liter of distilled water), mixed thoroughly in a 9x13x2 inch glass pyrex dish, placed on ice, and irradiated for 2 minutes using a 25-watt germicidal UV lamp placed

20 centimeters from the dish. Eggs and sperm were then mixed and sperm was activated by adding water. Ten minutes after fertilization, eggs were heat shocked at 26 ºC for 10 minutes, water hardened, transferred to Heath trays for incubation, and raised using the same conditions described for diploid families. All Kingscote X Killala families were produced at the Codrington Fisheries Research Facility (Ontario Ministry of Natural Resources and Forestry; Figure 1.1; Codrington, Ontario). This facility has a surface water supply which undergoes seasonal and diel temperature variation ranging between 2-5^{0}C in winter and 9-16^{0}C in summer.

SAMPLE PREPARATION

For all Kingscote x Killala families, DNA from offspring and parents (Table 1.1) was extracted using the high-throughput SPRI bead-based extraction protocol described in Ali et al. (2016) with Serapure beads (described in Rohland and Reich 2012) substituted for Ampure XP beads. For the Seneca x Parry Sound crosses, DNA was extracted using Qiagen DNeasy Blood and Tissue extraction kits (69506, Qiagen, Hilden, Germany) using manufacturer recommendations. DNA quality was initially assessed using a Nanodrop 2000 Spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts) by evaluating 260/230 and 260/280 absorbance ratios. Samples were diluted to less than 100ng/ul based on Nanodrop readings, then diluted 10-fold before determining double-stranded DNA (dsDNA) concentrations using Quantit Picogreen assays (Thermo Fisher Scientific, Waltham, Massachusetts).

SEQUENCING LIBRARY PREPARATION

DsDNA concentrations were normalized to 10ng/ul using an Eppendorf epMotion 2750 TMX liquid handling robot (Eppendorf, Hamburg, Germany) before proceeding with

the bestRAD protocol and RAD-capture using 100ng of total input DNA (Ali et al. 2016). Modifications to the protocol are noted below with a detailed description of methods provided as supplementary material (Supplementary Methods 1.1; Appendix A). First, the enzyme PstI was substituted for SbfI and PstI was heat-killed at 80ºC rather than 65 ºC. After ligating bestRAD adapters and pooling samples, shearing was carried out using a Covaris E220 Ultrasonicator (Covaris Inc., Woburn, Massachusetts) using the recommended settings for a 300bp mean fragment length. Finished libraries were amplified for 10 cycles, pooled equally in sets of two, and bead cleaned twice using a 0.9:1 bead-to-DNA ratio Ampure XP cleanup (A63881, Beckman Coulter, Brea, California). The two resulting pools were then enriched for 58,889 RAD loci that were previously found to be variable in Lake Trout populations in the Great Lakes, Seneca Lake, Ontario, Montana, and Alaska using the RAD-capture protocol (Ali et al. 2016). Target enrichment reactions were carried out using a MyBaits Custom Target Enrichment kit using manufacturer recommendations (MycroArray, Ann Arbor, Michigan; Protocol Version 3; for more information on capture and bait selection see Supplemental Methods 1.1; Appendix A). Finished capture reactions were amplified for an additional 9 cycles, pooled, and sequenced in three lanes of an Illumina HiSeq X instrument (2 X 150 bp paired end reads; Illumina, San Diego, California) by the Novogene Corporation (Novogene, Sacremento California).

BIOINFORMATICS AND GENOTYPING

Read quality was initially assessed using FastQC v0.11.5 (Andrews 2014), and a custom script was used to re-orient paired end reads such that individual specific barcodes and restriction enzyme overhang sequences were always located at the beginning of the

first read. Reads were demultiplexed using process_radtags v2.2, duplicate reads were removed using clone_filter v2.2 (Catchen et al. 2013; Rochette et al. 2019), and adapter sequences were clipped from reads using Trimmomatic v0.36 (Bolger et al. 2014). At this point, we produced two sets of fastq files: one conservatively filtered dataset used for de novo assembly of RAD loci and a slightly less conservatively filtered dataset used for calculating genotype likelihoods that would ultimately be used for linkage mapping and other analyses. For the de novo assembly dataset, reads were trimmed whenever the mean base quality across a sliding window of 4bp dropped below Q20, read pairs were removed if one or both reads in a pair were less than 140bp in length after trimming, and reads were cropped to a length of 140bp such that all reads were of identical length. For the dataset used to calculate genotype likelihoods, reads were trimmed whenever the mean base quality across a sliding window of 4bp dropped below Q15 and excluded if one or both reads in the pair were less than 50bp after trimming.

The stringently filtered dataset (read length =140bp, trimming threshold of Q20) was used to assemble RAD loci de novo using modules available in Stacks v2.2 (Rochette et al. 2019). RAD loci were identified for individuals using ustacks v2.2, which was run with a minimum depth of coverage of 3 (-m 3), a maximum distance between stacks of 3 (-M 3), a maximum distance to align secondary reads to primary stacks of 2 (-N 2), a minimum of 2 stacks at each de novo locus (--max_locus_stacks 2), and disabling calling haplotypes from secondary reads (-H). We then created a catalog of RAD loci for the parents of crosses using cstacks v2.2, allowing for up to two mismatches between sample loci when building the locus catalog (-n 2). Putative RAD loci alleles for all individuals were matched to this catalog using sstacks v2.2, converted to bam format using tsv2bam v2.2, and then

assembled using gstacks v2.2. Consensus sequences for RAD loci were obtained by passing the "--fasta-loci" flag to the populations v2.2 module. The fasta file containing RAD locus consensus sequences was normalized using Picard NormalizeFasta v2.8 (http://broadinstitute.github.io/picard/), indexed using bwa index v7.15 (Li 2013) and samtools faidx v1.3 (Li et al. 2009), and used as a de novo reference for subsequent analysis.

Next, the larger set of variable length paired end reads that were trimmed using a Q-threshold of 15 were mapped to the de novo assembly using bwa mem v7.15 (Li 2013) with default setting. Genotype likelihoods were calculated for single nucleotide polymorphisms (SNPs) within RAD loci using Lepmap3 v0.2 and associated modules (Rastas et al. 2015). SAM files produced by bwa-mem were converted to bam format and sorted using samtools v1.3, then converted to mpileup format using a minimum mapping quality of 30 and a minimum base quality of 20. The resulting file was filtered using the script pileupParser2.awk using a minimum read depth of 3 and a missingness threshold of 0.3. Genotype likelihoods were calculated using the pileup2posterior.awk script distributed with LepMap3 v0.2 (Rastas et al. 2015, Rastas 2017). We opted to use pileup2posterior.awk to calculate genotype likelihoods because the LepMap3 pipeline was originally validated using likelihoods calculated using this program (Rastas 2017).

LINKAGE MAP CONSTRUCTION

Linkage mapping and additional data filtering were carried out using various programs distributed with LepMap3 v0.2 (Rastas 2017). First, any missing parental SNP genotypes were imputed using ParentCall2. Second, SNPs showing evidence of segregation distortion were removed using Filtering2 with a p-value (--dataTolerance) threshold of

0.01. We required that SNPs be informative for linkage mapping in at least 1 family and removed SNPs with minor allele frequencies less than 0.05.

SNPs were assigned to linkage groups (LGs) using SeparateChromosomes2 run with logarithm of odds ratios (LOD) thresholds ranging from 8 to 60 and a minimum LG size of 50 SNPs. No single LOD threshold produced the expected number of LGs (n=42; Phillips and Zajicek 1982; Reed and Phillips 1995). Beginning with the map produced using a universal LOD threshold of 10, we determined the LOD thresholds needed to further split each LG by running SeparateChromosomes2 using all LOD thresholds between 10 and 60 and specifying the LG targeted for additional splitting using the "lg" and "map" flags (similar to Christensen et al. 2018a).

We determined that the largest 8 of the initial 30 LGs could be split using LOD thresholds ranging from 11-52, with the remaining 22 LGs remaining intact for all LOD thresholds between 10 and 60. The 8 largest LGs were split using the maximum LOD threshold that resulted in a new LG containing more than 50 SNPs, resulting in 42 LGs. Unassigned singleton SNPs were then joined to this map using JoinSingles2All run iteratively with a LOD threshold of 10 and a minimum LOD difference of 5.

The order of SNPs was initially determined by running 20 iterations of OrderMarkers2 and selecting the order with the highest likelihood for each LG. LGs were further refined by evaluating LOD matrices (output using computeLODscores=1). For each SNP, the vector of LOD scores corresponding to possible map positions was normalized such that values ranged from 0 to 1. SNPs were removed if the maximum LOD score was less than 1 standard deviation from the mean or if more than one LOD 'peak' was observed for any given SNP, indicating the existence of multiple mapping positions of similar

likelihood. LOD peaks were identified using the findPeaks function from the R package pracma v2.2.5, a minimum normalized peak height of 0.95 and a minimum distance between peaks greater than 25% of the length of the vector of mapping positions. RAD loci were removed from the data set if associated SNPs mapped to more than one LG. Finally, the dataset was thinned to include a single SNP for each RAD locus, with preference given to the SNP closest to the PstI restriction cut site. We opted to thin SNPs after determining which loci could be effectively mapped in order to maximize the number of unique RAD loci on the map. Maps for each LG were then reconstructed using the evaluateOrder and improveOrder=1 options from OrderMarkers2, with SNPs that failed the above filtering criteria flagged for removal using the removeMarkers option.

Finally, LGs were inspected for possible mis-ordering using LMPlot and any LG marked with possible errors were reordered using OrderMarkers2 for an additional 60 iterations. The linkage map was further improved by trimming SNPs from the ends of LGs based on manual inspection of LOD matrix plots and alignment to rainbow trout, Arctic char, and Atlantic salmon (Salmo salar) genome sequences (see Homology section below). An additional 10 iterations of ordering were conducted after removing potential erroneously placed SNPs from the ends of LGs. Final LGs were sorted based on their number of mapped SNPs and named as Sna1-Sna42. Both male and female linkage maps were output by the program.

CENTROMERE MAPPING

We identified centromeres by estimating the frequency of second division segregation (γ) across linkage groups using half-tetrad analysis conducted on gynogenetic diploid offspring from family G1 (Thorgaard et al 1983). Cells of gynogenetic diploid

offspring contain two of the four possible meiosis II products (a half-tetrad) and the frequency of heterozygous offspring can be used to estimate the frequency of recombination events between the locus in question and the centromere (Thorgaard et al 1983). Reads for these individuals were aligned to de novo assembled RAD loci, sorted and indexed using samtools v1.3 (Li et al. 2009), and variable positions within RAD loci were genotyped using freebayes v1.1.0 (Garrison and Marth 2012). Genotypes were called without applying population or binomial observation priors, an assumed contamination probability of 1%, a minimum base quality of 20, and a minimum mapping quality of 20. Called loci were then converted to their simplest representation using vcfallelicprimatives (https://github.com/vcflib/vcflib; vcflib v1.0.0) and loci with more than 2 alleles and indels were removed, such that only SNPs remained. Genotypes were set to missing if there was less than 1 order of magnitude difference in genotype likelihoods between the called genotype and the second most likely genotype using vcftools v0.1.16 (GQ >10; Danecek et al. 2011). SNPs were removed from the dataset if more than 30% of individuals were missing genotypes or if the frequency of the minor allele was less than 0.05. SNPs were further excluded if they were not placed on the linkage map, not called heterozygous in the mother, or if both possible homozygous genotypes were not observed in offspring. The mother was removed from the dataset at this point, and observed heterozygosity for the offspring (y) was calculated using the hwe function from SeqVarTools v1.20.2 (Gogarten et al. 2017; https://github.com/smgogarten/SeqVarTools). Centromeric regions were delineated as the region between the first and last markers with y-values less than 0.1 (as in Limborg et al. 2016).

Results were cross-validated and improved upon using the RFm method (Limborg et al. 2016) applied to the phased genotypes of progeny from families S1, S2, P1, and P3. Counts of maternal recombination events were reported using OrderMarkers2 with outputPhasedData=1 and used to calculate RFm across all maternal haplotypes and identify putative centromeric regions using a cut-off value of 0.45 as suggested in Limborg et al. (2016). The correct centromeric locations for acrocentric and telocentric chromosomes were identified by selecting the region containing, or neighboring, the lowest y-values from half-tetrad analysis.

HOMOLOGY

RAD loci were aligned to the reference genomes for Arctic char (RefSeq Accession: GCF_002910315.2), Rainbow trout (Oncorhynchus mykiss; RefSeq Accession: GCF_002163495.1) and Atlantic salmon (RefSeq Accession: GCF_000233375.1) using bwa mem v7.15 (Li 2013). RAD loci were assigned to their respective linkage map positions, and male and female linkage maps were visualized relative to their order along homologous chromosomes using ggplot2 v3.2.1 (Wickham and Chang 2008). Chromosomes were considered homologous if 50 or more mapped RAD loci aligned to a chromosome with mapping qualities greater than MQ60. The map was also compared with a linkage map for brook trout (Salvelinus fontinalis; Sutherland et al. 2016) using the program MapComp (Sutherland et al. 2016; https://github.com/enormandeau/mapcomp) and the Arctic char genome as an intermediate reference in order to detect large structural variants differentiating the two species.

Putative chromosomal inversions were detected by manually inspecting plots produced by mapping the Lake Trout linkage map to divergent references. Inversion

breakpoints were defined by the coordinates with the greatest discrepancy between the divergent physical map and the female linkage map we constructed. Inversions were classified as pericentric if putative inversion coordinates overlapped centromere mapping positions. Inversions differentiating Lake Trout and brook trout were detected by manually inspecting dot plots produced by MapComp.

TRAIT MAPPING

Offspring from diploid Kingscote x Killala crosses were phenotyped for fork length (FL), weight (WT), and condition factor (CF) at age 3. Additionally, photographs collected at age 3 were used to score individuals for presence-absence of spots and vermiculations (VPA) and two composite variables (PCA1 and PCA2) summarizing variation in body shape. Body shape variables were derived by performing a principal-components analysis (PCA) on the coordinates of morphometric landmarks that were normalized for slight differences in fish position and rotation using generalized Procrustes analysis. Using available photographs from families P1 and P3, we placed landmarks using tpsDIG v2 (Rohlf 2005) consistent with those described in Muir et al. (2014). Landmark coordinates were normalized and rotated using generalized Procrustes analysis conducted using the function gpagen from the R-package geomorph v3.1.1 (Adams and Otárola-Castillo 2013). Four of 20 landmarks could not be consistently placed using available images (1,6,7,10) and were therefore excluded from the analysis. Synthetic variables PCA1 and PCA2 were calculated by performing PCA on the resulting normalized coordinates and extracting scores for the first two axes. PCA was carried out using the function prcomp from the R-package stats v3.5.3. VPA, PCA1, and PCA2 phenotypes were available for 143 of 179 individuals. Fork length, condition factor, and weight phenotypes were collected for 179 of 179 individuals.

Phased SNP genotypes for offspring were extracted from the final map files reported by OrderMarkers2 using the script map2genotypes.awk from LepMap3 v0.2. QTL mapping was then carried out for traits of interest using the R-package qtl2 v0.2 and associated functions (Broman et al. 2019). All traits were mapped to sex-averaged linkage map coordinates. Prior to QTL mapping, pseudo-markers were added to the map using insert_pseudomarkers with a step size of 1cM and genotype probabilities were calculated using calc_genoprob. A kinship matrix was calculated using the calc_kinship function using genotype probabilities. A thinned subset of markers obtained using calc_grid (step=3) and probs_to_grid was used as input for the calc_kinship function. QTL scans were carried out using scan1 and suggestive QTL peaks were identified using find_peaks (drop=2, peakdrop=2, threshold = 3). Traits with approximately normal distributions (FL, CF, WT, PCA1, PCA2) were mapped using a mixed linear model with the kinship matrix included as a random effect (model = "normal" and kinship options in qtl2). Presence-absence of vermiculations and spots (VPA) was mapped as a binary trait (model = "binary" in qtl2). The kinship matrix was not included as a random effect in the binary trait mapping model because this option was not available in qtl2. For each identified LOD peak, 95% credible intervals were calculated using the function find_peaks (prob = 0.95, peakdrop=2, threshold=3). Finally, p-values were calculated by comparing observed LOD scores for each peak with a null distribution obtained from permuting the data 1000 times. Permutations were carried out using the function scan1perm using the same settings as the original tests and p-values were calculated using the ecdf function. The proportion of phenotypic variation explained (PVE) by each QTL peak was calculated from LOD scores and sample sizes using the equation PVE=1-10^(-(2/n)*LOD) (Broman and Sen 2009). Candidate genes

for significant LOD peaks ($p <= 0.05$) were identified by mapping RAD loci within 95% credible intervals to the Arctic char genome and determining the three genes closest to each mapping position using the program bedtools closest v2.26 (Quinlan and Hall 2010). Genes were considered candidates if they were within 50Kb of the mapping position of a RAD locus falling within the identified QTL mapping interval.

We also mapped the sex determination region using the binary trait model using qtl2 and assessed significance using the same methodology described above. The sexually dimorphic on the Y chromosome gene (sdY) is believed to underly sex determination in Lake Trout and some other salmonids (Yano et al. 2013). We designed a sdY presence-absence melt curve qPCR assay (similar to Anglès d'Auriac et al. 2014) using the Lake Trout sdY and 18S primers described in Yano et al. (2013). 18S served as an internal amplification control. Each reaction was carried out using a 0.4 uM concentration of primers sdYE2S1 (CCCAGCACTGTTTTCTTGTCTCA) and sdYE2AS1 (TGCTCTCTGTTGAAGAGCATCAC), and a 0.04uM concentration of primers 18SS (GTYCGAAGACGATCAGATACCGT) and 18SAS (CCGCATAACTAGTTAGCATGCCG). Reaction volumes were 20uL and contained 10ul of Forget-Me-Not EvaGreen qPCR mastermix (31045, Biotium, Fremont, California), 2.5 uL of template DNA, and 7.5uL of primers eluted in water. 18S and sdY were amplified in a two-step multiplex reaction using a 2-minute heat activation step at 95oC followed by 40 cycles of denaturation at 95C for 5 seconds and annealing/extension at 60oC for 30 seconds. Melt curve analysis was carried out on PCR product for temperatures between 60oC and 95oC using 0.1oC temperature shifts and a 3 second pause between temperature shifts. We first tested the assay on a subset of 32 individuals of known sex (16 males and 16 females), including the parents used for crosses,

in order to verify that males and females could consistently be differentiated based on the presence of a male specific sdY peak in the derivative of the melt curve. Offspring from all diploid families were subsequently genotyped using the same reaction conditions described above. At least one known male and one known female were included on each plate as a control. Sex locus mapping was carried out with sdY presence being coded as 1 and sdY absence coded as 0.

RECOMBINATION RATE ESTIMATION

We estimated sex averaged recombination rates for each chromosome by performing a simple linear regression of pairwise physical distance (base pairs) against genetic distance (cMs) and requiring the intercept to pass through 0. In order to evaluate a pair of RAD loci, we required that they map to the same chromosome on the Arctic char, rainbow trout, or Atlantic salmon genome assemblies and only retained scaffolds and chromosomes with greater than 50 mapped RAD loci for which the mapping quality was 60. For each LG, 100 pairs of RAD loci mapping to the same chromosome were randomly sampled from all possible pairs and recombination rate (cM/MB) was estimated using the slope of the resulting regression. This process was repeated 100 times using alignments against the Arctic char, rainbow trout, and Atlantic salmon genomes. The mean of the distribution of estimates was reported as the chromosome specific recombination rate, and separate values were reported for alignments against the three different divergent reference genomes. Regressions were carried out using the R-package lm and recombination rate estimates were visualized using ggplot2 v3.2.1 (Wickham and Chang 2008). This process was repeated for male and female maps in order to obtain sex specific chromosomal recombination rates.

RESULTS

BIOINFORMATICS AND GENOTYPING

We obtained a mean of 2,685,178 demultiplexed paired end (PE) reads for offspring from diploid crosses (range = 660,474 – 4,317,086, SD = 598,973.3) and a mean of 4,701,286 reads for parents (range = 3,483,973 – 5,868,449, SD = 787,251.1). On average, 21.97% of reads were removed by clone_filter for these individuals. De novo assembly of RAD loci with gstacks produced 146,525 RAD loci ranging in size from 140 to 754 bp in length. Between 92.86% and 95.20% of reads were mapped to de novo assembled RAD loci (mean = 93.52%, SD = 0.33%) using bwa mem. The Lepmap3 genotyping pipeline reported genotype probabilities for 212,158 SNPs, 147,920 of which were informative for linkage mapping. Of those, 72,549 SNPs passed missingness, segregation distortion, and minor allele frequency filters.

For gynogenetic diploid offspring, we obtained an average of 3,873,649 PE reads (range = 1,517,646 – 5,789,490, SD = 1,004, 905). We generated 3,536,915 reads for the mother of this family. On average, 32.63% of reads for these samples were removed by clone_filter. Between 89.0% and 89.6% of those reads were mapped to the de novo assembly using bwa mem (mean = 89.3%, SD= 0.12%). After genotyping with freebayes and filtering data to remove non-informative markers, we identified 893 SNPs that were informative for half-tetrad analysis.

LINKAGE AND CENTROMERE MAPPING

We were able to assign 15,740 RAD loci to LGs with between 878 and 113 loci mapped to each LG (Figure 1.2, Table 1.1, Supplementary Table 1.1). The total male map length was 2043.41cM and the female map was 2842.22 cM (overall female:male map ratio

= 1.391). Male LG map lengths ranged from 1.97 cM -92.87 cM, while female LG map lengths ranged from 43.07 cM – 134.64 cM (Table 1.2). SNPs were mapped to between 60 and 244 unique positions on linkage groups. As expected, we identified 42 LGs, 8 of which were metacentric and 34 that were acrocentric or telocentric (Supplementary Tables 1.1 and 1.2). These linkage groups likely correspond to the 42 chromosomes identified by previous karyotyping studies (Phillips and Zajicek 1982; Reed and Phillips 1995). Half-tetrad analysis yielded centromere intervals for 7 of 8 metacentric chromosomes and 22 of 34 LGs identified as acrocentric or telocentric (Figure 1.2, Supplementary Figure 1.8, Supplementary Table 1.2).

RFm analysis identified centromeres for 8 of 8 metacentric chromosomes and 32 of 34 acrocentric or telocentric chromosomes (Figure 1.2, Supplementary Table 1.2). We were ultimately able to determine the location of centromeres for 41 out of 42 chromosomes using the two methods. We were not able to map the centromere for Sna42; however, this chromosome is likely acrocentric or telocentric based on the size of the linkage group relative to others (Table 1.2) and karyotyping work suggesting the existence of 34 acrocentric and telocentric chromosomes (Phillips and Zajicek 1982; Reed and Phillips 1995).

HOMOLOGY

Alignment of the linkage map to divergent salmonid reference genomes revealed that the resulting map was highly congruent with existing assembled salmonid genomes (Supplementary Figures 1.1-1.6). Large synteny blocks were detected between Lake Trout linkage groups and the Arctic char genome for linkage groups Sna1-Sna41 (Table 1.3). Alignments suggested that Sna42 is syntenic with sal34; however, fewer than 50 loci with

MQ60 mapped to this chromosome from Sna42. Syntenies were detected between Lake Trout and all rainbow trout and Atlantic salmon chromosomes. MapComp identified homologies with all brook trout linkage groups identified by Sutherland et al. (2016) (Supplementary Figure 1.7).

The Lake Trout karyotype is differentiated from Arctic char by multiple Robertsonian translocations including one possible chromosomal fusion (Sal6.1 and Sal6.2) and four chromosomal fissions (Sal8, Sal14, Sal20, Sal4q.1.29). Sal6.1 and Sal6.2 are fused and Sal4q.1.29 is split into two LGs in Lake Trout, similar to the Arctic char linkage map presented by Nugent et al. (2017). The two Salvelinus species are also differentiated by at least 6 putative chromosomal inversions (Table 1.4), primarily on acrocentric or telocentric chromosomes. Arctic char chromosome Sal14 in particular appears to be the result of a fusion between Sna24 and Sna33. Sna24 also contains multiple chromosomal inversions that differentiate the two karyotypes (Figure 1.3). With the exception of inversions detected on Sna24, all putative inversions differentiating the two species were found to be near, or overlapping, the centromere (n=5, Table 1.4). MapComp results suggest inversions on Sna10, Sna11, Sna24, and Sna34 are shared with brook trout; however, large inversions differentiating brook trout and Lake Trout were identified on Sna28 (brook trout BC35), Sna12 (brook trout BC9), and Sna23 (brook trout BC25).

TRAIT MAPPING

Multiple quantitative trait loci were detected for the evaluated traits (Table 1.5, Figure 1.4). A highly significant QTL for presence of spots and vermiculations mapped to a sex-averaged position of 3 cM on Sna3 (VPA1, 95% CI = 0-4.485 cM, LOD = 6.563, p=0.001). We identified 16 candidate genes associated with this peak, including melanoregulin-like

(MREG-L, Arctic char scaffold NW_019942894.1: 64734-79619; Sna3, 1.575 cM). A second QTL for this trait mapped 21.095 cM on the same linkage group (VPA2, 95% CI = 19.685 – 30.175, LOD = 4.850, p=0.014). A total of 176 candidate genes were identified within this QTL credible interval. The four genes closest the highest LOD value were transcription factor 20-like (TCF20-L), retinoic acid induced 1-like (RAI1-L), sterol regulatory element binding factor 1-like (SREBF1-L), and calcium channel voltage-dependent T type alpha 1I subunit-like (CACNA1I-L; Supplementary Table 1.3). These QTL explained 11.5 and 10.8 percent of phenotypic variance, respectively (Table 1.5).

Significant QTL for fork length mapped to two locations on Sna1 (FL1, 39.00 cM, 95% CI = 36.94 – 44.6cM, LOD = 4.401, p = 0.03, and FL2, 60.265 cM, 95% CI = 51.475 – 66.07 cM, LOD = 4.224, p = 0.043) and one location on Sna12 (FL3, 57.630 cM, 95% CI = 51.835-63.03 cM, LOD = 4.226, p = 0.043). A significant QTL for condition also mapped to 60.265 cM on Sna1 (CF1, 95% CI = 52.6 – 73.105, LOD = 3.796, p = 0.045) and a QTL for weight mapped to 57.665 cM on Sna12 (W1, 95% CI = 50.55 - 64.15 cM, LOD = 4.13, p = 0.045). Suggestive QTL (LOD > 3, p > 0.05) were detected on Sna1 (60.265 cM, 95% CI = 37.365 – 72.4, LOD = 4.052, p = 0.062) and Sna12 (60.095 cM, 95% CI = 47.72 – 64.15 cM, LOD = 0.009, p = 0.278) for weight and condition factor, respectively. We identified 39 candidate genes associated with peak FL1, 137 genes associated with FL2, and 77 genes associated with FL3. We did not search for candidate genes for other growth and body condition related QTL (weight and condition factor) because the locations and credible intervals for these QTL overlapped almost perfectly with those detected for fork length (Table 1.5).

Suggestive QTL were detected for the PCA1 body shape variable on Sna5 (11.830 cM, 95% CI = 10.8 – 16.145, LOD = 3.651, p = 0.156), Sna24 (PCA1_1, 35.990 cM, 95% CI = 27.3 – 44.5 cM, LOD = 4.259, p = 0.049), and Sna33 (4.550 cM, 95% CI = 0 – 6.39 cM, LOD = 3.554, p = 0.184); however only the peak on Sna24 was statistically significant. Suggestive QTL were detected for PCA2 on Sna2 (64.464 cM, 95% CI = 45.94 – 80.32, LOD = 3.594, p = 0.188), Sna32 (45.745 cM, 95% CI = 27.925 – 50.59 cM, LOD = 3.041, p = 0.451), and Sna34 (22.820 cM, 95% CI = 0 – 39.405 cM, LOD = 3.087, p = 0.423); however, none of these QTL were found to be statistically significant. 111 candidate genes were identified for the significant QTL interval for PCA1 on Sna24.

Our sdY presence-absence assay produced a male specific peak in melt curve derivative plots at approximately 84ºC. A melt curve derivative peak existed for 18S at approximately 85ºC. Of the 16 known males used to validate the assay, 15 produced sdY peaks and 16 produced 18S peaks. The 16 known females tested all yielded 18S peaks; however, sdY peaks were absent in all females as expected. We were ultimately able to determine genotypic sex for 323 offspring produced from diploid crosses. Mapping sdY presence-absence as a binary trait using qtl2 identified strong peaks of association at 78.54 cM (95% CI = 75.66 – 82.14 cM, LOD = 8.538, p <0.001) and 84.43 cM (95% CI = 82.14 – 86.12 cM, LOD = 8.04, p <0.001) on Sna4.

RECOMBINATION RATES

Sex averaged recombination rates estimated by alignment to the Arctic char genome ranged from 0.138 to 2.935 cM/MB with a mean of 0.817 cM/MB (Supplementary Table 1.4; SD = 0.537) across LGs. In general, recombination rate estimates generated by mapping to the Arctic char genome were lower than those obtained from mapping to rainbow trout

or Atlantic salmon genomes (Supplementary Table 1.4, Supplementary Figure 1.9). Consistent with previous studies on salmonids, male recombination rates were considerably lower than those observed for females (Supplementary Figure 1.9). For example, the mean male recombination rate base on alignment to the Rainbow Trout genome was 0.203 cM/MB (SD = 0.133), while the mean female recombination rate was 1.31 cM/MB (SD = 0.602). Alignment of male and female linkage maps to divergent reference genomes demonstrated that male recombination is highly suppressed except for near telomeres (Supplementary Figures 1.1-1.3).

DISCUSSION

MAP EVALUATION

Multiple lines of evidence suggest this linkage map provides an accurate representation of the Lake Trout genome. First, we identified a single centromere for each chromosome (except Sna42), suggesting that linkage groups were appropriately split. In all cases, centromere mapping locations derived from half-tetrad and RFm analysis either overlapped or were in close proximity (Figure 1.2, Supplementary Table 1.2). For acrocentric and telocentric chromosomes, centromeres always mapped to the end of the chromosome with the highest female recombination rate and lowest male recombination rate, which matches results from previous centromere mapping efforts in salmonids suggesting that male recombination occurs almost exclusively near telomeres (Moen et al 2008; Miller et al. 2012; McKinney et al. 2016). Second, our homology analysis demonstrated a high degree of contiguity between existing genome assemblies and the map presented here (Supplementary Figures 1.1-1.7). Finally, our sex determination locus mapping results are concordant with cytogenetic studies. Previous cytogenetic analysis of

male and female Lake Trout identified sex-specific quinacrine and C-banding patterns on a large submetacentric chromosome (Phillips and Ihssen 1985). Mapping sdY presence-absence using the linkage map demonstrated with high certainty that the sex locus exists near one of the telomeres of Sna4 (Figure 1.4), which is metacentric or submetacentric based on RFm and half tetrad analysis (Figure 1.2). Although two significant peaks were detected on this LG, they were in close proximity and credible intervals were adjacent, suggesting that they likely represent a single peak of association. This study and others suggest that Salvelinus species, and salmonids in general, have highly variable sex chromosome configurations. Specifically, the brook trout sex determination locus maps to a region that is homologous to the Arctic char sex chromosome; however, it localizes to a different arm (Sutherland et al. 2017), while the Lake Trout sex chromosome identified here lacks homology with those of all species examined (Sal4p.1 – Nugent et al. 2017; BC35 – Sutherland et al. 2017; Ssa02, Ssa03, and Ssa06 – Kijas et al. 2018; Omy29 – Pearse et al. 2019). Many previous studies have identified variation in sex locus mapping position both within and between salmonid species (Woram et al. 2003; Lubieniecki et al. 2015; Sutherland et al. 2017; Kijas et al. 2018), even though the same gene ultimately underlies sex determination in most cases (Yano et al. 2013). Our results add to a growing body of literature suggesting that sdY is a conserved, yet highly mobile, sex determination gene in salmonids.

Furthermore, the Lake Trout linkage map presented here is of similar density to those used to scaffold genome assemblies for other salmonids (Christensen et al. 2018a, 2018b) and provides valuable information on the order of loci along chromosomes and recombination rates between loci. In general, male:female map length ratios and estimated

sex averaged map lengths were highly similar to those observed for other salmonids. For instance, Leitwein et al. (2018) found that chromosome specific recombination rates varied from 0.21 – 4.1 cm/MB (mean =0.88) for brown trout, compared with 0.138 to 2.935 cM/MB (mean = 0.817) for Lake Trout based on mapping linkage mapped RAD loci to the Arctic char reference genome. Similar to other salmonids, we observed pronounced heterochiasmy, with male recombination being almost entirely suppressed except for near telomeres (Moen et al. 2004; Moen et al 2008; McKinney et al. 2016; Leitwein et al. 2018)

STRUCTURAL VARIATION

Combining centromere mapping locations with synteny analysis revealed that Lake Trout are differentiated from Arctic char, rainbow trout, and Atlantic salmon by multiple pericentric inversions, suggesting that centromeric instability (specifically on acrocentric and telocentic chromosomes) is potentially an important component of the salmonid evolutionary legacy (Table 1.4). Future work should evaluate whether these detected inversions are truly species diagnostic or if they are polymorphic within species. Large structural variants have previously been found to be associated with adaptive differentiation and life history variation within rainbow trout (Miller et al. 2012; Pearse et al. 2019) and inversions can contribute to pre or post-zygotic isolation between species or ecotypes (Kirkpatrick 2010). Future studies should evaluate the extent to which the structural variation detected here contribute to reproductive isolation and adaptive divergence within and between salmonid species. Sna24 presents one of the most striking examples of extensive structural variation in the genus Salvelinus, with multiple paracentric and pericentric inversions differentiating the lineages containing Arctic char and Lake Trout (Figure 1.3). With the exception of a putative inversion located between 0

and 12 cM, all other inversions on this LG were not observed in other salmonid species examined, suggesting that the other inversions on this chromosome (Sna24, Ssa14; Figure 1.3, column 2) are fixed or segregating within the Arctic char lineage or within the Salvelinus clade containing Arctic char, bull trout (S. confluentus), dolly varden trout (S. malma), and white char (S. albus). This hypothesis is supported by results from MapComp which suggested that brook trout, the most closely related extant species to Lake Trout (Crête-Lafrenière et al. 2012), and Lake Trout are not differentiated by any inversions on this linkage group. A large inversion spanning the centromere of Sna11 shows clear evidence of being differentially fixed between Lake Trout and all other salmonids except brook trout (Figure1.3, Supplementary Figures 1.1-1.3, Supplementary Figure 1.7). Inversions located on Sna10, Sna24, and Sna34 also appear to be differentially fixed between the Lake Trout – brook trout lineage and all other salmonids; however, interpretation is complicated by subsequent translocations and inversions that occurred in other taxa (Table 1.4). A large pericentric inversion on Sna12 appears to differentiate Lake Trout from all other salmonids, including brook trout (Supplementary Figure 1.7). MapComp results also suggest that two large inversions on Sna28 (homologous to the brook trout sex chromosome - BC35; Sutherland et al. 2017) and Sna23 (BC25) differentiate Lake Trout from closely related brook trout (Supplementary Figure 1.7). It is unclear if these structural variants are truly fixed between species, or if they might be polymorphic within Lake Trout or brook trout. The inversion polymorphisms identified above could be associated with chromosomal speciation within the genus Salvelinus or adaptive divergence within salmonid species and warrant further examination.

The majority of detected inversions differentiating Lake Trout from other salmonids are pericentric, which is not entirely unexpected. Repeat-rich eukaryotic centromeres often demonstrate exceptionally high rates of evolution (Henikoff et al. 2001) and are prone to chromosomal breakage and the accumulation of structural variation (Kalitsis and Choo 2012; Barra and Fachinetti 2018). Sutherland et al. (2016) also identified evidence for multiple inversions differentiating salmonid species, including one pericentric inversion differentiating pink, chum, and sockeye salmon from other salmonids.

Evidence for F2 inviability and reduced reproductive success between hybrids are widespread (Stebbins 1958), including for pairs of closely related species within the salmonid lineage (Renaut et al. 2011). Bull trout and brook trout, for instance, readily produce F1 offspring but F2 offspring are rarely observed (Leary 1993). Hybrids between westslope cutthroat (Oncorhynchus clarkii lewisi) and rainbow trout are viable but have dramatic reductions in reproductive success (Muhlfeld et al. 2009). Future work should evaluate if instances of reduced fitness in inter or intraspecific salmonid hybrids might be linked to combined deleterious effects of recombination at multiple pericentrically inverted loci. It would also be interesting to ascertain whether centromeric regions tend to harbor signals of adaptive divergence between salmonid species and morphotypes. For example, Ellegren et al. (2012) found elevated levels of divergence between Fidecula flycatcher species near centromeres. Given the prevalence of pericentric inversions on acrocentric and telocentric chromosomes, we also might expect these loci to be associated with adaptive ecophenotypic radiations that have occurred within Salvelinus (Eschenroder 2008) and Coregonus (Lu and Bernatchez 1999).

GENOMIC BASIS FOR ADAPTIVE TRAITS

Suggestive QTL for traits that differentiate Lake Trout morphotypes were detected on multiple linkage groups. This supports the hypothesis proposed by Perreault-Payette et al. (2017) that ecophenotypic divergence in Lake Trout has a polygenic basis. Our results suggest that the presence or absence of spots and vermiculations is controlled by either one or two loci on the same arm of linkage group Sna3. A search for candidate genes within the QTL mapping intervals identified melanoregulin-like (MREG-L) as a potential causal locus. The homolog of this gene is involved in the transfer of melanosomes from melanocytes to keratinocytes (Wu et al. 2012), and appears to control the distribution of pigments within mice hair (O'Sullivan et al. 2004). Pigmentation polymorphisms are common in Lake Trout (Wilson and Mandrak 2004; Zimmerman et al. 2007) and other trout and char (Gomez-Uchida 2008), although it is unclear if the genes identified here explain skin pigmentation variation in other species and populations. Skin pigmentation variation has been shown to be associated with depth of capture in multiple Lake Trout populations and is hypothesized to be adaptive in some environments (Protas and Patel 2008); however, it is also possible that the trait is simply linked with some other adaptive traits. Pigmentation patterns are often linked to variation in behavior, immune response, and energy homeostasis in vertebrates, likely owing to pleiotropic effects of melanocortins (Ducrest et al 2008). Pigmentation traits have also been linked to stress response in rainbow trout, Atlantic salmon, and Arctic char (Hoglund et al. 2000; Kittilsen et al 2009).

Suggestive QTL for the composite body shape variable with the highest explanatory power (PCA1) were detected on Sna5, Sna24, and Sna33. Interestingly, each of these chromosomes appear to have undergone structural reorganization in relatively recent

evolutionary history, based on alignment to the Arctic char genome (Supplementary Figure 1.1). Specifically, Sna24 and Sna33 are fused in Arctic char and Sna5 is split into two chromosomes. Sna24 in particular appears to have accumulated multiple large inversions that differentiate this linkage group from the homologous region of the syntenic Arctic char chromosome. A QTL for condition factor, which is closely related to body shape (Froese 2006), has been previously detected on the brook trout linkage group homologous to Sna33 (linkage group BC16, Sutherland et al. 2017). Additional mapping crosses, ideally generated using ancestral populations with highly differentiated body shapes (leans vs. siscowet or divergent hatchery strains for example) would be valuable for further validating the existence of QTL detected here and improving our understanding of the genetic basis for adaptive divergence within Lake Trout.

Growth and body condition related traits all have suggestive QTL on linkage groups Sna1 and Sna12, indicating that genes on these chromosomes likely harbor variation that underlies differences in growth between populations. Linkage group Sna12 also appears to harbor an inversion that differentiates Lake Trout from other salmonid species examined, including brook trout. A previous study identified a putative growth rate QTL on the brook trout linkage group homologous to Sna12 (Supplementary Figure 1.7; Sutherland et al 2017; BC9). The same study identified a stress response QTL, measured as change in blood cortisol levels following handling stress, in brook trout on the chromosome homologous to Sna1 (Sutherland et al. 2017, BC6). Increased cortisol levels have been found to be negatively corelated with growth and condition factor in other salmonids (Barton et al. 1987; Reinecke 2010), suggesting that variation observed in our families could actually be due to variation in stress response. There is evidence for variation in fitness among Lake

Trout hatchery strains used to supplement and restore Lake Trout populations in the Great Lakes, with the strain from Seneca Lake, New York appearing to have a fitness advantage (Scribner et al. 2018). Great Lakes Lake Trout populations are heavily impacted by predation by invasive sea lamprey and previous work has shown that larger individuals have a greater probability of surviving lamprey attack (Swink 1990). Similarly, size-selective fisheries have also been shown to impose strong natural selection on growth in multiple species (Enberg et al 2012). Future work could examine whether the chromosomal regions identified here are associated with size-at-age or are under selection in populations experiencing lamprey predation or size-selective fisheries.

Associations between environmental conditions and phenotype have been observed across the Lake Trout range and across salmonid species for the afore mentioned traits, suggesting adaptive significance in some contexts. For example, patterns and intensity of skin pigmentation, along with divergence in other traits, is commonly associated with depth-of-capture in both Lake Trout (Zimmerman et al. 2007; Marin et al. 2016) and Arctic char populations (Gomez-Uchida et al. 2008). Variation in skin pigmentation is potentially involved with predator avoidance and camouflage, feeding behavior, mate choice (Protas and Patel 2008), or protection from ultraviolet radiation (Yan et al. 2013). Differences in age specific growth rates are also frequently observed between humper, lean, and siscowet-like Lake Trout morphotypes (Burnham-Curtis and Bronte 1996; Hansen et al. 2012) — as well as between Arctic char morphotypes (Jonsson et al. 1988; Snorrason et al. 1994; Adams et al. 1998). These differences in growth rate likely reflect variation in allocation of resources toward growth and reproduction, adaptation to nutrient stress (Arendt 1997), or plastic responses to environmental variation (Hindar and Jonsson 1993).

Morphotypes can also often be differentiated based on body shape differences, which are hypothesized to be optimized for different feeding behaviors and modes of locomotion (Bond 1996; Muir et al. 2014; Perreault-Payette et al. 2017). For example, the streamlined body shape of leans has been hypothesized to be adaptive for swimming and predation in shallower nearshore environments (Bond 1996; Muir et al. 2014), while the more deep-bodied shape of siscowet Lake Trout is believed to reflect adaptation for vertical migration and foraging in deep-water habitats (Webb 1984; Muir et al 2014). Morphotypes with traits reflecting those observed in the species native range have the potential to emerge rapidly in some introduced invasive populations (Stafford et al. 2013), suggesting a high degree of phenotypic plasticity or exceptionally strong selection favoring divergence.

Unfortunately, many Lake Trout metapopulations of conservation concern have experienced reductions in abundance and decreases in ecophenotypic diversity as a result of overexploitation and introduction of invasive species (Krueger and Ihssen 1995; Hansen 1999). For example, in the Great Lakes the siscowet morphotype has been extirpated from all lakes except Lake Superior (Krueger and Ihssen 1995). The results presented here enhance understanding of the genetic architecture of traits that underlie trophic specialization in Lake Trout and could aid in restoring genetic and phenotypic diversity in lakes where it has been lost.

CONCLUSIONS

We identified multiple structural variants potentially involved in speciation and adaptation within the genus Salvelinus, mapped the Lake Trout sex determination locus, and identified QTL for traits believed to be adaptively significant in Lake Trout populations.

Future work should use additional QTL mapping crosses and association studies in wild populations to evaluate if the QTL identified here are consistently associated with the phenotypic variation examined, as well as other phenotypes that differentiate Lake Trout morphotypes. Trophically specialized Lake Trout morphotypes and adaptively diverged populations are differentiated by multiple other traits (i.e. tissue lipid content, fin size, diet; Thurston 1962; Eschmeyer and Phillips 1965; Zimmerman et al. 2006; Zimmerman et al. 2007). QTL mapping studies using later generation crosses or genome-wide association studies in wild populations would be particularly useful for fine-scale localization of genotype-phenotype associations within QTL credible intervals identified here. Additionally, QTL mapping efforts can yield different results for different families and the genetic basis for some traits often varies across populations (Santure et al. 2015). The Lake Trout linkage map will allow further examination of the genetic basis for ecophenotypic variation in Lake Trout and will enable additional exploration of chromosomal evolution within the genus Salvelinus. Perhaps most important, this resource will allow for the assembly of a chromosome-anchored reference genome for Lake Trout, which will greatly facilitate future genomic research on this important species.

APPENDIX

Figure 1.1: Map displaying the locations of hatchery facilities (dots) and locations of wild progenitor populations (diamonds) used for mapping. Locations of hatchery facilities used for conducting crosses are marked with black circles. The locations of the progenitor populations are identified with black diamonds. Longitude is displayed on the Y-axis and latitude is displayed on the X-axis.

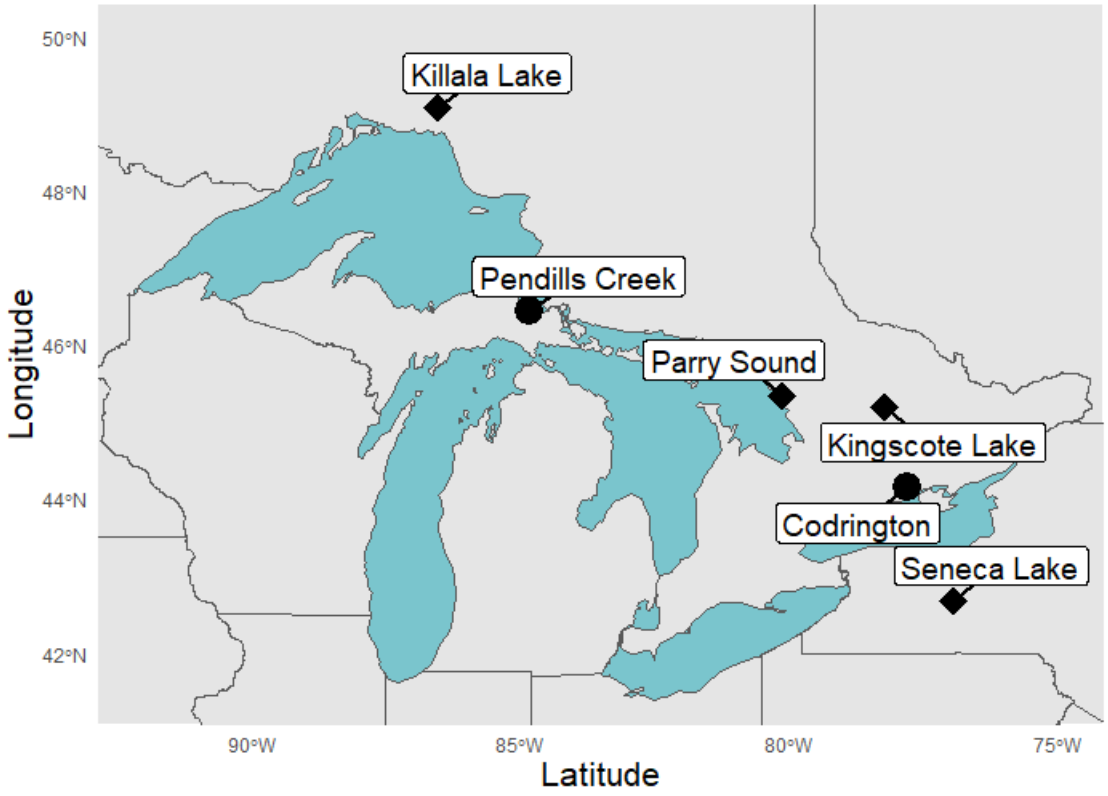

Figure 1.2: Locations of 15,740 RAD loci along 42 Lake Trout linkage groups. Orange boxes highlight centromeres identified using half tetrad analysis with a y-threshold of 0.1. Blue boxes span the intervals of centromeres identified using the RFm method (Limborg et al. 2016) combined with half-tetrad analysis. Locations are in centimorgans on the female linkage map.

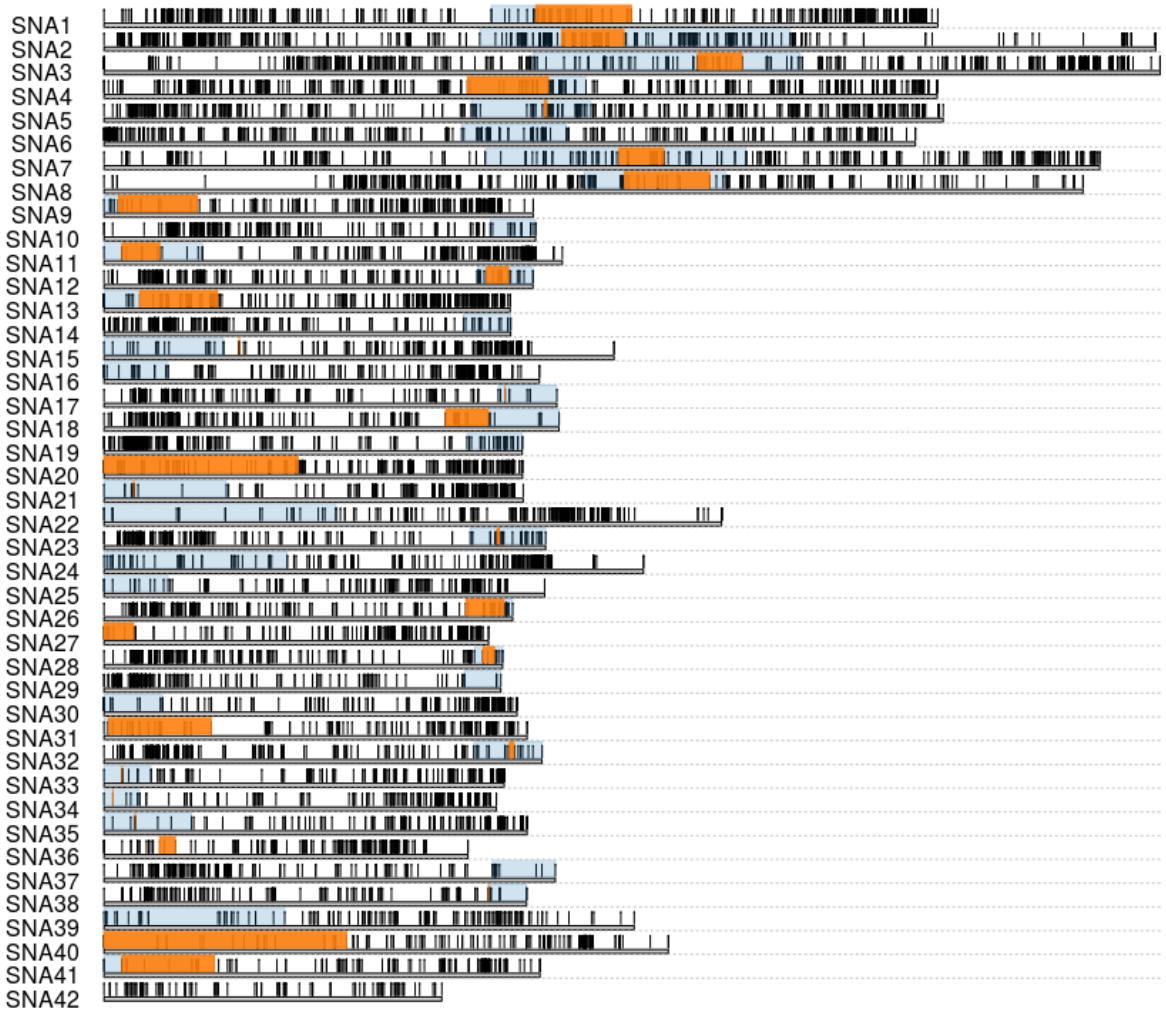

Figure 1.3: Examples of two linkage groups (Sna11 and Sna24) with evidence of inversions differentiating Lake Trout from other salmonids. Female Lake Trout linkage groups are colored blue (top curves). Male Lake Trout linkage groups are colored red (bottom curves). Sna11(first column) is differentiated from all homologs by a single large pericentric inversion spanning 0-30cM on the female linkage map (left side of each panel). Sna24 is differentiated from Omy04 and Ssa06 by an inversion spanning 0-10cM on the female map. It is unclear if the same inversion exists in Arctic char due to extensive structural differentiation relative to Lake Trout and other salmonids (Sna24 vs. Sal14).

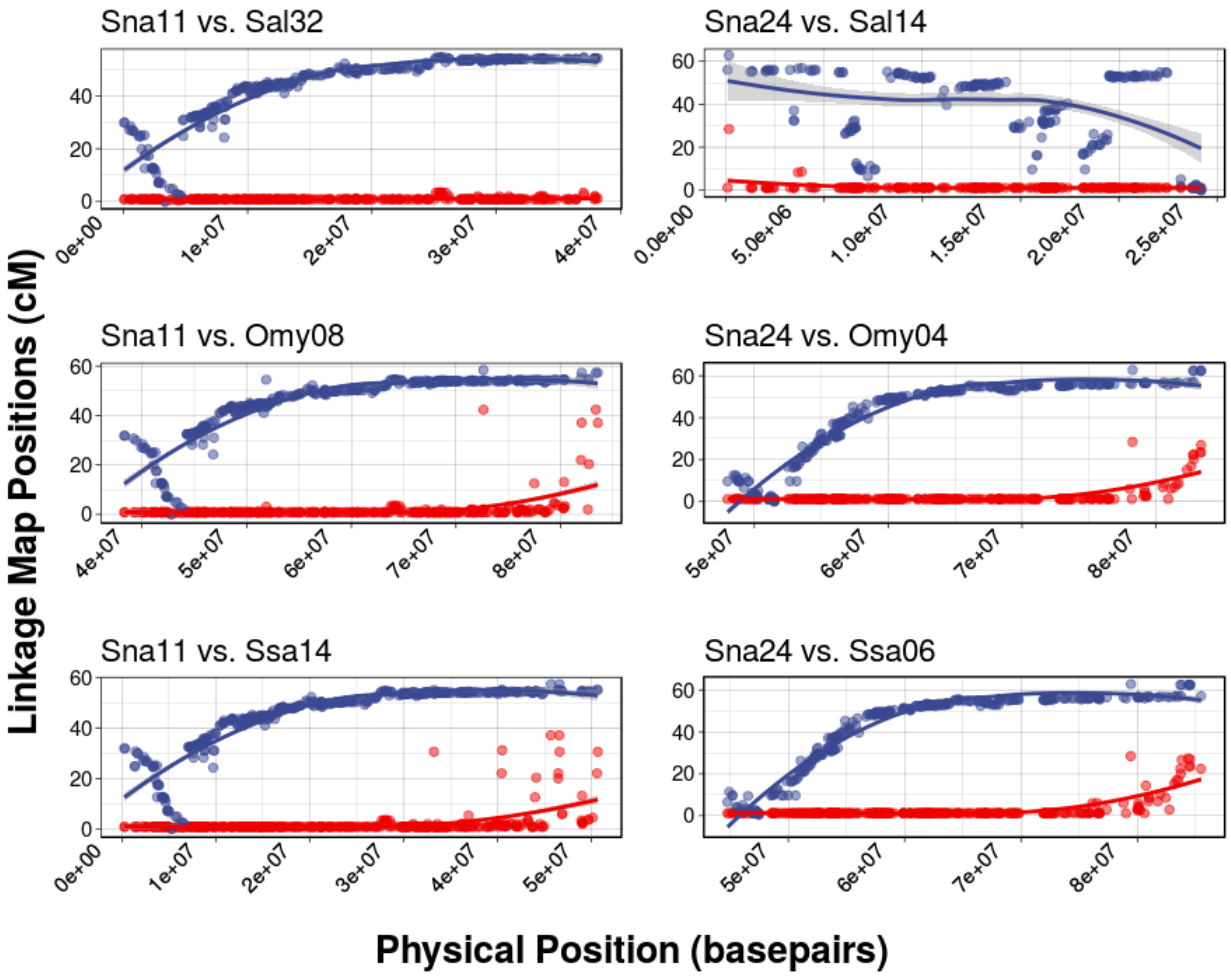

Figure 1.4: Panels display LOD values on the Y-axis versus sex averaged map position (cM) for QTL scans for (A) the sex determination locus, (B) presence of spots and vermiculations, (C) fork length, (D) weight, (E) condition factor, (F) PCA1, and (G) PCA2. The dashed red line corresponds to the p< 0.05 significance threshold for LOD scores. The solid green line corresponds to the LOD threshold of 3 used to identify peaks putatively associated with each trait.

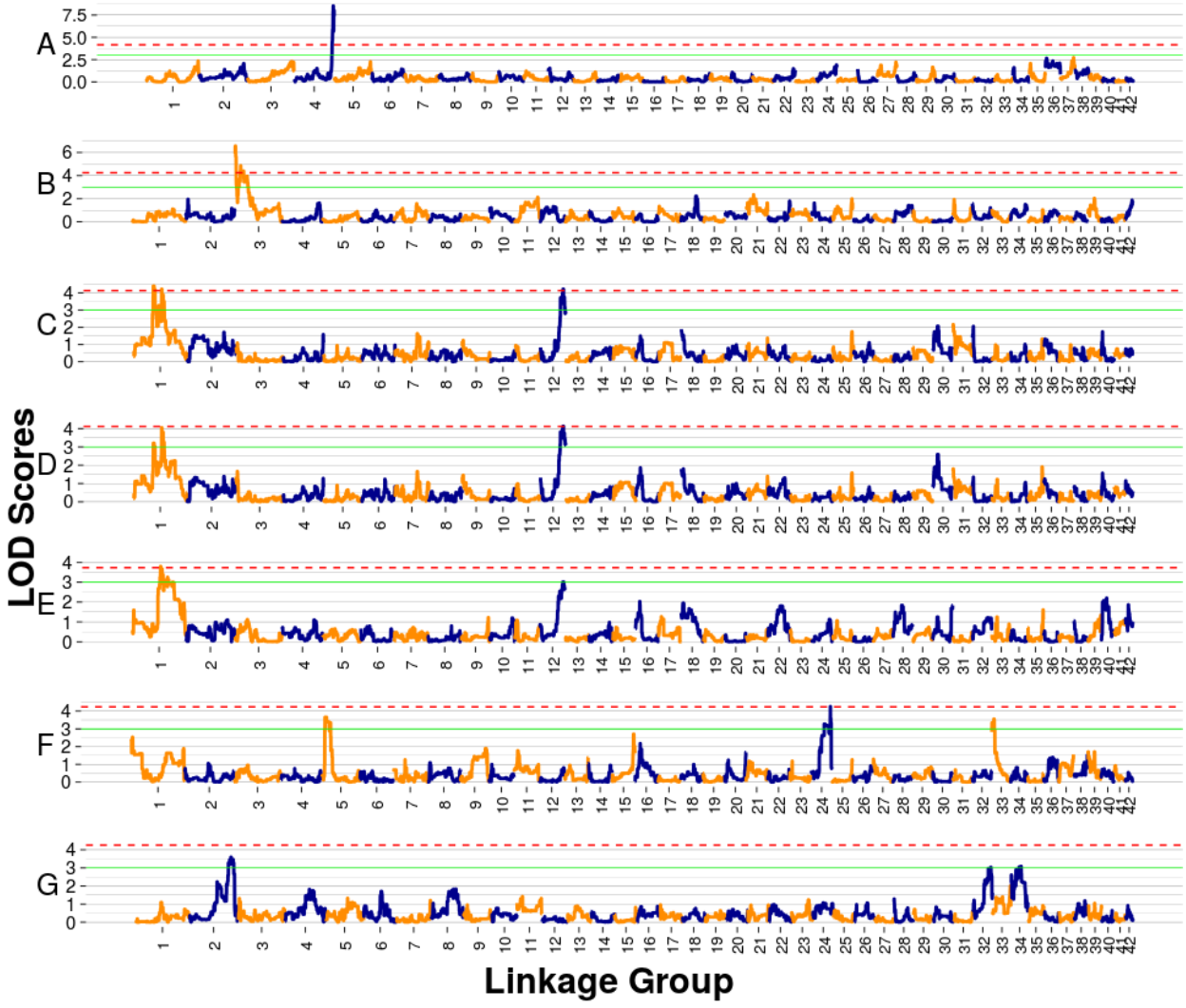

Table 1.1: Family IDs, cross type (diploid or gynogenetic diploid), number of genotyped offspring per family, and maternal and paternal origins for the five families used for linkage and QTL mapping.

Family	Type	No. Offspring	Mother Origin	Father Origin
S1	Diploid	88	Parry Sound	Seneca Lake
S2	Diploid	91	Parry Sound	Seneca Lake
P1	Diploid	91	Killala X Kingscote F1	Killala X Kingscote F1
P3	Diploid	88	Killala X Kingscote F1	Killala X Kingscote F1
G1	Gyn. Diploid	45	Killala X Kingscote F1	None

Table 1.2: Summary statistics for each of the 42 linkage groups. No. Mapped Loci corresponds to the number of unique RAD contigs mapped to each linkage group. Male and Female map lengths are in centimorgans (cM). No. Unique Positions corresponds to the number of unique linkage map positions to which RAD loci were assigned. Female:Male Ratio is the ratio of Female Length and Male Length in centimorgans.

Name	No. Mapped Loci	Male Length (cM)	Female Length (cM)	Female:Male Ratio	No. Unique Positions
Sna1	878	85.28	106.3	1.246	217
Sna2	789	71.98	134.08	1.863	207
Sna3	761	59.12	134.64	2.277	244
Sna4	648	66.01	106.22	1.609	207
Sna5	618	50.95	106.98	2.1	221
Sna6	515	77.88	103.44	1.328	223
Sna7	514	73.49	126.94	1.727	195
Sna8	497	92.87	124.81	1.344	157
Sna9	460	48.39	54.72	1.131	112
Sna10	406	54.29	55.02	1.013	112
Sna11	404	42.38	58.43	1.379	100
Sna12	395	73.61	54.69	0.743	121
Sna13	389	34.02	51.78	1.522	116
Sna14	377	59.65	51.85	0.869	110
Sna15	360	48.12	65.02	1.351	101
Sna16	358	53.9	55.51	1.03	102
Sna17	357	48.72	57.73	1.185	95
Sna18	356	44.67	58.01	1.299	101
Sna19	348	41.15	53.32	1.296	109
Sna20	344	36.93	53.35	1.445	102
Sna21	340	41.96	53.44	1.274	92
Sna22	333	70.57	78.73	1.116	109
Sna23	332	61.14	56.31	0.921	106
Sna24	325	28.44	68.78	2.418	105
Sna25	322	63.3	56.18	0.888	98
Sna26	319	36.71	52.1	1.419	95
Sna27	317	33.52	49.03	1.463	94
Sna28	313	37.01	50.81	1.373	102
Sna29	312	48.84	50.59	1.036	86
Sna30	310	66.35	52.65	0.794	83
Sna31	307	36.9	53.94	1.462	93
Sna32	302	56.79	55.84	0.983	102
Sna33	286	50.85	51.02	1.003	89
Sna34	255	52.77	50.02	0.948	85
Sna35	244	42.17	53.94	1.279	94
Sna36	242	30.19	46.4	1.537	80
Sna37	225	26.84	57.54	2.144	82
Sna38	218	35.06	53.87	1.537	83
Sna39	194	22.56	67.59	2.996	91
Sna40	185	33.84	71.95	2.126	90
Sna41	172	2.12	55.58	26.217	60
Sna42	113	1.97	43.07	21.863	60

Table 1.3: Synteny between Lake Trout linkage groups and Arctic Char, Rainbow Trout, Atlantic Salmon, and Brook Trout genomes.

Lake Trout	Arctic Char	Rainbow Trout	Atlantic Salmon	Brook Trout
Sna1	Sal15	Omy6	Ssa24, Ssa26	BC6
Sna2	Sal1	Omy17	Ssa12	BC3
Sna3	Sal20	Omy12	Ssa03, Ssa13	BC8, BC14
Sna4	Sal18	Omy16, Omy23	Ssa01, Ssa19	BC1
Sna5	Sal6.1, Sal6.2	Omy2, Omy14	Ssa05	BC7
Sna6	Sal3	Omy21	Ssa07	BC2
Sna7	Sal27	Omy15, Omy18	Ssa16, Ssa29	BC5
Sna8	Sal13	Omy4, Omy10	Ssa04, Ssa23	BC4
Sna9	Sal26	Omy1	Ssa16	BC20
Sna10	Sal16	Omy5	Ssa10	BC17
Sna11	Sal32	Omy8	Ssa14	BC22
Sna12	Sal23	Omy10	Ssa04	BC9
Sna13	Sal2	Omy3	Ssa25	BC24
Sna14	Sal7	Omy9	Ssa15	BC30
Sna15	Sal9	Omy19	Ssa01	BC12
Sna16	Sal17	Omy16, Omy20	Ssa13, Ssa19	BC18
Sna17	Sal8	Omy25	Ssa09	BC33
Sna18	Sal33	Omy11	Ssa20	BC40
Sna19	Sal36	Omy22	Ssa21	BC26
Sna20	Sal11	Omy7	Ssa22	BC21
Sna21	Sal4q.1:29	Omy2	Ssa10	BC15
Sna22	Sal25	Omy1	Ssa18	BC36
Sna23	Sal22	Omy27	Ssa20	BC25
Sna24	Sal14	Omy4	Ssa06	BC31
Sna25	Sal19	Omy28	Ssa03	BC11
Sna26	Sal5	Omy29	Ssa11	BC10
Sna27	Sal31	Omy18	Ssa27	BC23
Sna28	Sal4q.2	Omy25	Ssa09	BC35
Sna29	Sal28	Omy8	Ssa15	BC19
Sna30	Sal10	Omy26	Ssa11	BC28
Sna31	Sal4q.1:29	Omy5	Ssa01	BC13
Sna32	Sal30	Omy14	Ssa14	BC34
Sna33	Sal14	Omy11	Ssa19	BC16
Sna34	Sal4p	Omy24	Ssa09	BC38
Sna35	Sal8	Omy20	Ssa28	BC27
Sna36	Sal37	Omy9	Ssa18	BC32
Sna37	Sal35	Omy3	Ssa02	BC29
Sna38	Sal24	Omy15	Ssa17	BC37
Sna39	Sal21	Omy13	Ssa02	BC42
Sna40	Sal12	Omy7	Ssa17	BC39
Sna41	Sal20	Omy13	Ssa06	BC14
Sna42	Sal34*	Omy19	Ssa08	BC41

Table 1.4: Inversions differentiating salmonid species. The first column is the Lake Trout linkage group in question and columns 2-4 list the approximate location of any detected inversions that differentiate species. The type of inversion is stated in parenthesis. Locations are listed in centimorgans on the female map. Whenever multiple inversions were detected on a chromosome, at least one was pericentric. ** = Inverted region appears to be translocated to a separate chromosome. * = suggestive evidence of structure variation but unable to determine if an inversion occurred. Centromeres were not localized for Sna42, so centricity of inversions could not be determined.

Linkage Group	Arctic Char	Rainbow Trout	Atlantic Salmon
Sna6	-	-	30-43 (Paracentric)
Sna10	45-55 (Pericentric)	**	*
Sna11	0-30 (Pericentric)	0-30 (Pericentric)	0-30 (Pericentric)
Sna12	48-54 (Pericentric)	**	48-54 (Pericentric)
Sna16	-	12-40 (Multiple Inversions)**	25-40 (Multiple Inversions)
Sna19	-	5-30 (Paracentric), 25 - 58 (Pericentric)	30-58 (Pericentric)
Sna20	-	0-10 (Pericentric)	0-10 (Pericentric)
Sna24	0-57 (Multiple inversions)	0-12 (Pericentric)	0-12 (Pericentric)
Sna25	-	-	0-30 (Pericentric)
Sna28	-	-	43-52 (Pericentric)
Sna30	*	0-10 (Pericentric)	-
Sna31	-	-	0-7 (Pericentric)
Sna34	0-30 (Pericentric)	*	*
Sna35	-	0-47 (Pericentric)	-
Sna41	*	-	-
Sna42	-	35-43 (Unknown)	20-43 (Unknown)

Table 1.5: Map locations of detected QTL. Linkage map positions (in centimorgans; cM) of QTL peaks detected for the sex determination locus, presence-absence of vermiculations and spots, fork length, shape variable PCA1, shape variable PCA2, weight, and condition factor (Trait column). CI_Low and CI_High are the upper and lower bounds of the 95% credible interval for map positions for each QTL peak. LG is the linkage group on which the QTL was detected. Model lists the model used for QTL mapping in r/qtl2. Positions are sex averaged map positions. LOD scores are the differences in log10 likelihoods for models assuming presence or absence of a QTL at the locus in question (reported by r/qtl2). The estimates proportion of phenotypic variance explained by each QTL peak is listed in the PVE column. Estimated additive and dominant effects for the peak in question are also listed. P-values are those obtained via the permutation test described.

Trait	LG	Position	CI (Low)	CI (High)	Model	LOD	Additive Effect	Dominance Effect	PVE	p-value
Sex	Sna4	78.54	75.66	82.14	Binary	8.538	1.049	-0.045	0.115	<0.001***
Sex	Sna4	84.43	82.14	86.12	Binary	8.04	1.055	-0.044	0.108	<0.001***
Vermiculation	Sna3	3	0	4.49	Binary	6.563	1.595	0.278	0.191	0.001***
Vermiculation	Sna3	21.1	19.69	30.18	Binary	4.855	0.103	1.048	0.145	0.014*
Fork Length	Sna1	39	36.94	44.6	Normal	4.401	18.905	8.058	0.107	0.030*
Fork Length	Sna1	60.27	51.48	66.07	Normal	4.224	15.91	9.172	0.103	0.043*
Fork Length	Sna12	57.63	51.84	62.03	Normal	4.226	-11.693	10.91	0.103	0.043*
PCA1	Sna5	11.83	10.8	16.15	Normal	3.651	-0.005	0.011	0.111	0.156
PCA1	Sna24	35.99	27.3	44.5	Normal	4.259	-0.003	-0.011	0.128	0.049*
PCA1	Sna33	4.55	0	6.39	Normal	3.554	0.008	0.007	0.108	0.184
PCA2	Sna2	64.47	45.94	80.32	Normal	3.594	0.006	0	0.109	0.188
PCA2	Sna32	45.75	27.93	50.59	Normal	3.041	0.004	-0.001	0.093	0.451
PCA2	Sna34	22.82	0	39.41	Normal	3.087	0.005	0	0.095	0.423
Weight	Sna1	60.27	37.37	72.4	Normal	4.052	48.657	29.021	0.099	0.062
Weight	Sna12	57.67	50.55	64.15	Normal	4.13	-40.95	29.692	0.101	0.049*
Condition	Sna1	60.27	52.6	73.11	Normal	3.796	0.05	0.033	0.093	0.045*
Condition	Sna12	60.1	47.72	64.15	Normal	3.009	-0.053	0.001	0.074	0.278

Supplemental Methods 1.1

This document provides a more detailed description of library preparation methods, targeted sequence capture methods, and a description of how baits used for targeted sequence capture were designed.

I. BestRAD Library Preparation for Mapping Families

DNA was digested for 1 hour at 37□ C using the restriction enzyme PstI-HF (R3140S, New England Biolabs, Ipswich, Massachusetts). Following digestion, the enzyme was heat killed by incubating at 80□ C for 20 minutes. Biotinylated bestRAD adapters (Ali et al. 2016) were ligated to PstI overhangs using T4 Ligase (M0202M, New England Biolabs, Ipswich, Massachusetts) by incubating at 25□ C for 12 hours before heat killing the enzyme at 80□ C for 20 minutes. Digested and adapter ligated DNA from all individuals within a plate was then pooled and concentrated into 55ul of low EDTA Tris-EDTA buffer (T0230, Teknova, Hollister, California) before shearing on a Covaris E220 Ultrasonicator (Covaris Inc., Woburn, Massachusetts) using the recommended settings for a 300bp mean fragment length. The fragment size distributions and concentrations for sheared and pooled libraries were assessed using a D5000 Tapestation assay (5067-5365, Agilent Technologies, Santa Clara, California) before proceeding.

RAD loci were isolated using M280 streptavidin beads (11205D, Thermo Fisher Scientific, Waltham, Massachusetts) following the exact protocol from Ali et al. (2016) and sequencing adapters were added using the NEB Next Ultra Library Prep Kit for Illumina (E7370S, New England Biolabs, Ipswich, Massachusetts). Adapters were diluted 1:10 prior to ligation and a unique 6 base pair i7 indexing primer was used for each library so they could be pooled. Libraries were amplified for 10 cycles and library quantity and insert size

were assessed using Quantit Picogreen (P11496, Thermo Fisher Scientific, Waltham, Massachusetts) and Tapestation D5000 assays (5067-5365, Agilent Technologies, Santa Clara, California), respectively. Finished libraries were then pooled in sets of two and bead cleaned twice using a 0.9:1 bead-to-DNA ratio Ampure XP cleanup (A63881, Beckman Coulter, Brea, California) to remove any residual indexing primers.

II. Hybridization Capture for Mapping Families

After preparing RAD libraries for mapping families, the following procedure was used to selectively enrich for 58,889 RAD loci that were previously found to be polymorphic within Lake Trout (see next section). Target enrichment reactions were carried out using a MyBaits Custom Target Enrichment kit using manufacturer recommendations (MycroArray, Ann Arbor, Michigan; Protocol Version 3). 400 nanograms of DNA were used as input for each reaction. Hybridization reactions were carried out at 65□ C for 24 hours and wash reactions were done at 65-67□ C in 1.5mL tubes. Following enrichment, pools were PCR amplified for 9 cycles using the KAPA Library Amplification Kit for Illumina (KK2620, KAPA Biosystems, Wilmington, Massachusetts) with universal primers according to the manufacturer recommended protocol. Final enriched pools were quantified using Quantit Picogreen assays (P11496, Thermo Fisher Scientific, Waltham, Massachusetts) run in triplicate and insert sizes were determined using D5000 Tapestation assays (5067-5365, Agilent Technologies, Santa Clara, California). All libraries were pooled in equal amounts before sequencing.

III. RAD-Capture Bait Selection and Design

The following procedures were used to discover polymorphic RAD loci within Lake Trout, select loci for targeted genotyping, and select the final bait panel used for genotyping

Lake Trout families. Variable RAD loci were discovered using PstI RAD sequencing carried out on 48 individuals collected from across the Lake Trout range using the bestRAD protocol (Ali, et al. 2016). These individuals included 3 individuals from the Lewis Lake hatchery strain, 3 individuals from the Seneca Lake hatchery strain, 3 individuals from the Apostle Island hatchery strain, 3 individuals from the Isle Royale hatchery strain, 9 individuals from the Marquette hatchery strain, 3 individuals from the Green Lake hatchery strain, 12 wild born individuals collected from Lake Huron, 2 humpers from Lake Superior, 2 leans from Lake Superior, 2 siscowet from Lake Superior, 2 individuals from Flathead Lake (Montana, USA), one individual from Lake Opeongo (Ontario, Canada), one individual from Lake of the Woods (Ontario, Canada), one individual from Schrader Lake (Alaska, USA), and one individual from Ugashik Lakes (Alaska, USA). Fifty nanograms of double stranded DNA from each sample was used as input for library preparation. Libraries were prepared exactly as described above; however, libraries were sheared using the recommended protocol for 250bp fragments. Prior to sequencing, the concentration and size of the library measured using a combination of Qubit dsDNA high sensitivity (Thermo Fisher, Waltham, Massachusetts), KAPA Illumina Library Quantification qPCR (KAPA Biosystems, Wilmington, Massachusetts), and Caliper LabChipGX HS DNA assays (Caliper Life Sciences, Waltham, Massachusetts). The library was sequenced in two HiSeq 4000 lanes in the 2X150 PE read format, using HiSeq 4000 SBS reagents (Illumina, San Diego, California).

Reads from 3 males and 3 females from the Marquette hatchery strain were used for de novo assembly of RAD loci using Stacks v1.44 (Catchen et al. 2013). These individuals were chosen because they had exceptionally high read depth. We chose individuals from a

single hatchery strain in order to minimize diversity and promote the assembly of long contigs. Fastq files were purged of clonal reads using clone_filter. and reads were trimmed whenever the average base quality score across a sliding window of 4 base pairs dropped below q20 using Trimmomatic v0.36 (Bolger et al. 2014). Reads were re-oriented such that the bestRAD barcodes were always found at the beginning of read 1 using a custom perl script before demultiplexing with process_radtags. Reads less than 140 bp were discarded and reads greater than 140 bp were cropped to 140bp before being used for read clustering and RAD locus assembly.

Loci were identified using ustacks (-m 2 -M 4 --model_type bounded --alpha 0.05 --bound_low 0 --bound_high 0.05 -p 32 -d -r –gapped), cstacks (-n 2 –gapped), and sstacks (--gapped). The paired end reads for all identified loci were deposited in separate fasta files using sort_read_pairs.pl and were assembled into contigs using the exec_velvet.pl wrapper script. Resulting contigs were concatenated onto 10 pseudo-scaffolds with 500 Ns between each contig. Contig sequences were renamed based on their coordinates within the pseudo-scaffolds and extracted to a new fasta file, which was normalized using Picard NormalizeFasta v2.8 (http://broadinstitute.github.io/picard/). The process resulted in the discovery and assembly of 1,292,171 RAD loci. Contig lengths ranged from 170 to 669 base pairs, with an average length of 316.71 bp.

In order to discover variable SNP loci, fastq files from two sequencing lanes (including all 48 individuals) were concatenated, and read quality was assessed for read 1 and read 2 files using Fastqc v0.11.5 (Andrews 2014). In order to avoid genotyping biases associated with clonal reads, duplicates were removed using the clone_filter program. Sequencing adapter contamination was removed from reads using Trimmomatic v0.36 and

reads were truncated whenever the mean Phred score across a window of 4 nucleotides dropped below q15. We further required reads to be greater than 60 bp after applying the trimming steps above. Reads were then re-oriented using a custom script such that the inline index for each individual was always found at the beginning of read 1. Properly oriented fastq files were demultiplexed by individual barcode using process_radtags. Reads were mapped to the de novo assembled reference using bwa-mem and resulting SAM files were sorted, converted to bam format, and indexed using samtools v1.4 (Li et al. 2009). BAM files were genotyped using HaplotypeCaller and the Genome Analysis Toolkit (GATK) incremental joint genotyping workflow (v3.7; McKenna et al. 2010) and the resulting VCF file was filtered using bcftools v1.4.1 (Li 2011).

Data were filtered using a variety of criteria meant to minimize the prevalence of false-positive variants. Before filtering on any variable, we evaluated the distribution of the variable relative to the frequency of the first alternate allele in order to ensure that we were not systematically truncating the distribution of allele frequencies. We also checked that the distributions of z-scores resulting from Wilcoxon Sign Rank Tests output by GATK genotypeGVCFs (Mapping Quality Rank Sum, Base Quality Rank Sum, Read Position Rank Sum distributions) were centered on zero and approximately normal. Deviation from normality or a mean different than 0 would indicate systematic biases associated with sequencing, de novo assembly, or genotyping.

For SNP loci, we required that QD (Quality standardized by depth) be greater than 2, SOR (strand odds ratio) be less than 3, MQ (Mapping Quality reported by GATK) be greater than 40, FS (Fisher Strand) be less than 60, Base Quality Rank Sum be between -2 and 2, Mapping Quality Rank Sum be between -2 and 2, Read Position Rank Sum be between -2

and 2, and depth across all samples be less than 5000x. We excluded RAD loci containing greater than 10 variants. Additionally, we required that contigs map to a single location in the Atlantic salmon genome and that mapping locations not overlap by more than the length of the cut-site overhang on the Atlantic Salmon genome (Lien et al. 2016).

All remaining SNPs were masked in the de novo reference using Picard FastaAlternateReferenceMaker v2.8 and RAD locus consensus sequences were extracted using bedtools getfasta v2.26 (Quinlan and Hall 2010) for RAD loci containing variable SNP loci. RAD loci were re-oriented such that the remainder of the PstI cut-site was always found at the beginning of the contig. Loci were removed from the dataset if the remainder of a PstI cut-site existed on both ends of a contig. We then used RepeatMasker v4.0.7 (Smit et al. 1996; http://www.repeatmasker.org/) to mask low complexity sequence and repeats using the Atlantic Salmon repeat library, while allowing for up to 10% divergence. Consensus sequences for RAD loci were trimmed whenever a repeat masked or low complexity region was encountered (with the end containing the cut-site being maintained) and loci less than 200bp in length after trimming were excluded. All remaining consensus sequences were cropped to 200bp and aligned to the complete set of 1,292,171 de novo assembled RAD loci using blat v0.35 (Kent 2002). Loci were removed if they aligned to an off-target locus with greater than 40% similarity (calculated as (matches-mismatches-gaps)/200). The remaining 64,242 loci were submitted to MycroArray for bait design on July 6th 2017 (Ann Arbor, Michigan). Two baits were designed for each locus; one adjacent to the PstI cut-site and another offset 80 bp from the cut-site. These baits were input into MycroArray's complementary bait quality control software. We retained baits only if the following criteria (reported by MycroArray) were met; the top blast hit to

the de novo assembly was less than 25% soft masked, delta G > -9, zero blast hits to Atlantic salmon or rainbow trout mtDNA, 0 heterodimers with other baits, at most 10 off target blast hits with Tm between 62.5 and 65 C, and fewer than 2 off target blast hits with Tm >= C. For loci where both baits passed all filtering criteria, preference was given to the bait closest to the PstI cut-site. This resulted in the retention of baits for 58,889 polymorphic RAD loci, which were subsequently used for targeted enrichment of RAD libraries prepared for mapping families (see above).

The following supplementary materials referenced herein are too large to be usefully displayed here and are available upon written request to the author. Descriptions of these documents are provided below.

Supplementary Data 1.1: Map information, phenotypes, and genotypes used for QTL and sex locus mapping.

Supplementary Table 1.1: The Lake Trout linkage map and sequences of mapped RAD loci.

Supplementary Table 1.2: Centromere mapping intervals for each linkage group

Supplementary Table 1.3: Candidate genes identified for significant QTL peaks.

Supplementary Table 1.4: Recombination rate estimates

Supplementary Figure 1.1: Alignments of Lake Trout linkage groups with Arctic Char chromosomes.

Supplementary Figure 1.2: Alignments of Lake Trout linkage groups with Rainbow Trout chromosomes.

Supplementary Figure 1.3: Alignments of Lake Trout linkage groups with Atlantic Salmon chromosomes.

Supplementary Figure 1.4: Dot plot grid comparing the Lake Trout linkage map with the Arctic Char genome assembly.

Supplementary Figure 1.5: Dot plot grid comparing the Lake Trout linkage map with the Rainbow Trout genome assembly.

Supplementary Figure 1.6: Dot plot grid comparing the Lake Trout linkage map with the Atlantic Salmon genome assembly.

Supplementary Figure 1.7: MapComp dot plot grid comparing the Lake Trout linkage map with the Brook Trout linkage map from Sutherland et al. (2016).

Supplementary Figure 1.8: Centromere mapping using half tetrad analysis.

Supplementary Figure 1.9: Recombination rate estimates for males and females.

CHAPTER 2: A CHROMOSOME-ANCHORED GENOME ASSEMBLY FOR LAKE TROUT (SALVELINUS NAMAYCUSH)

ABSTRACT

Here we present an annotated, chromosome-anchored, genome assembly for Lake Trout (Salvelinus namaycush) – a highly diverse salmonid species of notable conservation concern and an excellent model for research on adaptation and speciation. We leveraged Pacific Biosciences long-read sequencing, paired-end Illumina sequencing, proximity ligation (Hi-C) sequencing, and a previously published linkage map to produce a highly contiguous assembly composed of 7,378 contigs (contig N50 = 1.8 Mb) assigned to 4,120 scaffolds (scaffold N50 = 44.975 Mb). Long read sequencing data were generated using DNA from a female double haploid individual. 84.7% of the genome was assigned to 42 chromosome-sized scaffolds and 93.2% of Benchmarking Universal Single Copy Orthologs were recovered, putting this assembly on par with the best currently available Salmonid genomes. Estimates of genome size based on k-mer frequency analysis were highly similar to the total size of the finished genome, suggesting that the entirety of the genome was recovered. A mitochondrial genome assembly was also produced. Self-vs-self synteny analysis allowed us to identify homeologs resulting from the Salmonid specific autotetraploid event (Ss4R) as well as regions exhibiting delayed re-diploidization. Alignment with three other salmonid genomes and the Northern Pike (Esox lucius) genome also allowed us to identify homologous chromosomes in related taxa. We also generated multiple resources useful for future genomic research on Lake Trout, including a repeat library and a sex-averaged recombination map. A novel RNA sequencing dataset for liver tissue was also generated in order to produce a publicly available set of annotations for

49,668 genes and pseudogenes. Potential applications of these resources to population genetics and the conservation of native populations are discussed.

INTRODUCTION

Key questions in evolution and conservation biology can only be addressed using genomic approaches and appropriate study species. Lake Trout (Salvelinus namaycush; Figure 2.1) are a top predator in many lentic ecosystems across northern North America and express exceptional levels of ecotypic variation (Muir et al., 2014; Muir et al., 2016), making them an ideal study species for exploring the processes of ecological speciation and adaptive diversification. The post-Pleistocene parallel evolution of diverse Lake Trout ecotypes has been likened to the adaptive radiation of cichlid species in the Great Lakes of east Africa (Muir et al., 2016); however, the radiation of Lake Trout ecotypes appears to have occurred over a relatively short evolutionary timescale (Harris et al., 2015, ~8000 years). At least three distinct Lake Trout ecotypes (lean, siscowet, and humper) once existed throughout the Laurentian Great Lakes (Hansen, 1999) and anecdotal evidence suggests that as many as 10 easily differentiable forms once existed in Lake Superior (Goodier, 1981). High levels of ecotypic variation have also been documented in contemporary populations across the species range (Blackie et al., 2003; Zimmerman et al., 2006; Hansen et al., 2016; Chavarie et al., 2015), with as many as five trophic ecotypes being found in a single lake (Marin et al., 2016).

Lake Trout are also ancestrally autotetraploid, with the common ancestor of all salmonids having undergone a whole genome duplication event (WGD) roughly 80-100 million years ago (Macqueen & Johnston 2014; Lien et al., 2016, Berthelot et al., 2014). For this reason, Salmonids have long been considered ideal study species for understanding the

evolutionary consequences of WGD (Ohno, 1970; Allendorf & Thorgaard, 1984). Previous studies have demonstrated that salmonid genomes exhibit a mixture of disomic and tetrasomic inheritance (Allendorf & Danzmann, 1997) and have suggested that salmonid homeologs can be partitioned into two broad categories – ancestral ohnolog resolution regions (AORe) and lineage specific ohnolog resolution regions (LORe; Roberston et al., 2017). AORe regions exhibit elevated differentiation between homeologs because these regions returned to a state of disomic inheritance prior to the radiation of Salmonid species. Conversely, LORe regions are characterized by extremely low levels of sequence differentiation between homeologs due to delayed rediploidization. Given the high levels of ecotypic diversity observed in Lake Trout, and the potential for WGD to facilitate the evolution of novel phenotypes (Ohno, 1970; Macqueen & Johnston 2014; Van De Peer et al., 2017) and reproductive isolation (Lynch & Force, 2000), research exploring the genetic basis for ecotypic differentiation and incipient speciation in Lake Trout could provide important insights about the role of LORe and AORe regions in more recent adaptive radiations.

Furthermore, many Lake Trout populations, particularly those in the Laurentian Great Lakes, have been severely reduced in abundance or distribution, or extirpated, due to invasive species introductions and overfishing (Smith, 1968). Following the basin-wide collapse of the lake whitefish (*Coregonus clupeaformis*) commercial fishery in the Great Lakes during the early 20th century, fishing pressure was transferred to Lake Trout populations, which partially contributed to population declines starting in the 1930s (Hansen, 1999). A novel predator, the sea lamprey (Petromyzon marinus), also invaded the Great Lakes during this time, leading to further increases in adult Lake Trout mortality and

functional extirpation from all lakes except Lake Superior and a small, isolated, population in Lake Huron (Hansen, 1999). The restoration program that commenced largely focused on reducing sea lamprey predation, reducing fishing pressure, creating aquatic refuges, and stocking juvenile Lake Trout from a diverse collection of domesticated strains originating from multiple source populations (Krueger et al., 1983; Hansen, 1999). Lake Trout populations in Lake Superior rebounded relatively quickly; however, the re-emergence of natural reproduction in other lakes was hindered by high levels of lamprey predation on adult Lake Trout (Pycha et al., 1980), predation on juveniles by invasive alewife (Madenjian et al., 2008), reduced juvenile survival caused by thiamine deficiency (Fitzsimmons et al., 2009), and potentially reduced hatching success associated with PCB contamination (Mac & Edsall 1991). Today, Lake Superior populations remain relatively stable and recruitment has been observed in lakes Huron (Riley et al., 2007), Michigan (Hanson et al., 2013), and Ontario (Lantry, 2015). Recent research suggests that domesticated strains used for reintroduction have variable fitness in contemporary Great Lakes environments (Scribner et al., 2018; Larson et al., 2021) and may be differentially contributing to recent recruitment. However, the biological mechanisms that underly these differences in fitness and recruitment remain unclear.

Genomic and transcriptomic approaches have been widely used to identify loci associated with adaptive diversity and ecotypic divergence in salmonids (Prince et al., 2017; Veale & Russelo, 2017; Willoughby et al., 2018; Rougeux et al., 2019). This work has been partially driven by the publication of high-quality genome assemblies and linkage maps for numerous salmonid species (Gagnaire et al., 2013; Lien et al., 2016; Christensen et al., 2018a, Christensen et al., 2018b; Pearse et al., 2019; De-Kayne et al., 2020); however,

genomic resources are notably lacking for Lake Trout. An annotated, chromosome-anchored, genome assembly is arguably the most valuable resource for advancing genomic research on any species. A publicly available reference genome for Lake Trout would eliminate many challenges associated with conducting conservation-oriented genetic research aimed at restoring ecotypic diversity and viable wild populations. Until recently, the assembly of non-model eukaryotic genomes was prohibitively expensive, computationally challenging, and required the collaborative efforts of large genome consortia; however, the development of long-read ('third generation') sequencing technologies has to some extent eliminated these hurdles (Hotaling & Kelley, 2020; Whibley et al., 2020).

Long-read sequencing data can be useful for scaffolding and filling gaps in existing, fragmented, short-read assemblies (English et al., 2012). A number of assembly algorithms also seek to assemble contigs directly from long-read sequencing data (Falcon, Chin et al., 2016; Canu, Koren et al., 2017; wtdbg2, Ruan & Li 2020) and recent work suggests that this approach can be highly effective for assembling chromosome-anchored salmonid genomes when combined with additional scaffolding information (De-Kayne et al., 2020; also see RefSeq: GCF_002021735.2).

Salmonid genomes are highly complex and relatively difficult to assemble owing to the existence of large LORe regions (Robertson et al., 2017) and high repeat content (Lien et al., 2016; De-Kayne et al., 2020; Kajitani et al., 2014). Sequencing low-diversity individuals from inbred lines or homozygous individuals produced via chromosome set manipulations provides one route for simplifying the assembly process and correctly assembling regions with low levels of differentiation between homeologs. Previous

salmonid genome assemblies have made use of doubled haploid individuals (Lien et al., 2016; Christensen et al., 2018b; Pearse et al., 2019) because these individuals are theoretically homozygous at all loci (but see Hansen et al., 2020). Additionally, long read sequencing data has been shown to be highly effective for assembling polyploid genomes (Du et al., 2020), and these data would likely improve our ability to resolve LORe regions in salmonids. For instance, De-Kayne et al. (2020) recently published a highly contiguous assembly for European Whitefish (Coregonus sp. balchen); however, this assembly was produced using data from an outbred, wild-caught, individual rather than a double haploid.

Here we present a chromosome-anchored reference genome for a female Lake Trout that was assembled using Pacific Bioscience long-read sequencing data and scaffolded using a high-density linkage map (Smith et al., 2020) and genome-wide chromatin conformation capture followed by massively parallel sequencing (Hi-C). We also produced a number of complementary resources including a custom repeat library and an interpolated recombination map in order to facilitate additional research on this important species. A publicly available set of gene annotations was also produced using the NCBI Eukaryotic Genome Annotation Pipeline. Additionally, we identify Lake Trout homeologs resulting from the Salmonid specific autotetraploid event (Ss4R) and establish homologous relationships with chromosomes from other salmonid species.

MATERIALS AND METHODS

CROSSING AND SAMPLE COLLECTION

Gynogenetic double haploids were produced by fertilizing eggs with UV irradiated sperm, then pressure shocking embryos immediately following the first mitotic division (as described in Thorgaard et al., 1983; Limborg et al., 2016). Double haploid (DH) offspring

were created at Iron River National Fish Hatchery using eggs and sperm collected from captive adult Lake Trout from the U.S. Seneca Lake brood stock. The U.S. Seneca Lake hatchery strain was entirely founded by early Autumn spawning Lake Trout initially collected from Seneca Lake, New York (see Page et al., 2003 and Krueger et al., 1995). Due to low survivorship of DH offspring (Komen & Thorgaard, 2007), we tested multiple UV and pressure shock treatments on eggs from five different females. Batches of 900 eggs from each female were fertilized with sperm that was irradiated for 140, 280, or 1,260 seconds. Each batch was then split and sub-batches were pressure shocked at 11,000 PSI for five minutes at either 6.5, 7, 7.5, 8, 8.5, 9, 9.5, or 10 hours post-fertilization. A total of 13,500 eggs were exposed to various UV and pressure shock treatments. One batch of 900 eggs from each female was also exposed to a control treatment which involved no sperm irradiation or pressure shock. Embryos were incubated in heath trays at ambient temperature until eye-up stage (E36 per Balon, 1980), with dead embryos being removed from trays on a daily basis. A single individual that survived past post-embryo stage (sensu Marsden et al., 2021) was grown to a size of approximately 5 centimeters before being sampled post-mortality and stored at -20ºC. The post-embryo stage in Lake Trout is characterized by a fully absorbed yolk sac, parr marks, and an inflated gas bladder (Marsden et al., 2021).

LABORATORY METHODS

High molecular weight (HMW) DNA was extracted from white muscle sampled from the DH individual using a MagAttract HMW DNA Extraction kit (Qiagen, Hilden, Germany). The manufacturer recommended protocol was used except tissue digestion was done at room temperature for 140 minutes rather than 12-16 hours at 55ºC. Fragment size and

yield were determined using pulse field gel electrophoresis and a AccuClear Ultra High Sensitivity DNA Quantification assay (Biotium, Fremont, California). Prior to sequencing and assembly, we verified that the DH individual was completely homozygous at 15 microsatellite loci that are typically highly heterozygous in Lake Trout populations (Valiquette et al., 2014). A long-read sequencing library was then prepared using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences, Menlo Park, California), with the optional DNA Damage Repair step after size selection. Size selection was made for fragments >10 kb using a Blue Pippin instrument (Sage Science, Beverly, Massachusetts) according to the manufacturer recommended protocol for 20kb template preparation. 5ug of concentrated DNA was used as input for the library preparation reaction. Library quality and quantity were assessed using a genomic DNA Tape Station assay (Agilent, Santa Clara, California), as well as Broad Range and High Sensitivity Qubit fluorometric assays (Thermo Fisher, Waltham, Massachusetts). Single-Molecule Real Time sequencing was performed on the Pacific Biosciences Sequel instrument at the McGill Genome Centre (McGill University, Montreal, Canada, https://www.mcgillgenomecentre.ca/) using an on-plate concentration ranging from 1.5-7.5pM and the Sequel Sequencing Kit 2.0 with diffusion loading. 38 SMRTCells were run with 600-minute movies and two SMRTCells were run with 1200-minute movies. All HMW DNA for the DH individual was expended over the course of PacBio sequencing runs. This necessitated the use of DNA from diploid individuals for generating additional libraries needed for scaffolding and polishing.

Hi-C proximity ligation libraries were generated using tissue from a 7-year-old diploid female Lake Trout originating from the Killala Provincial hatchery strain. Four Hi-C libraries were prepared using spleen and white muscle tissue using the Arima Hi-C kit

according to the manufacturer's protocol (A510008, Arima-HiC_AnimalTissue_A160132_v00, Arima Genomics, San Diego, CA) and library preparation kits from Kapa Biosystems (Wilmington, Massachusetts) and Lucigen (Middleton, Wisconsin). Each Hi-C library was spiked into a portion of an Illumina HiSeqX lane in order to assess how effectively reads could be mapped against the draft contig assembly. HiCUP version 0.7.2 (Wingett et al., 2015) within Genpipes version 3.1.5 (Bourgey et al., 2019) was used to map Hi-C sequencing reads against draft contigs. The Hi-C library prepared using muscle tissue and prepared using the Arima-Hi-C and Lucigen Kits, was selected for further sequencing given that this library produced the highest proportion of reads mapped to draft contigs. This kit employs a restriction enzyme cocktail that digests chromatin at N^GATC and G^ANTC sequence motifs. The selected library was sequenced to high coverage in a single HiSeqX lane using the 2X150 bp paired end read format. Sequencing produced 182,781,953 paired end reads.

DNA was also extracted from fin tissue collected from an adult (diploid) female Lake Trout from the Seneca Lake broodstock using a MagAttract HMW DNA extraction kit (Qiagen, Hilden, Germany) and the manufacturer recommended protocol. Sequencing reads from this Seneca strain female were later used for contig polishing and correction (described below in Assembly and Scaffolding). The library was prepared using 100ng of input DNA and the NEBNext Ultra Library Preparation Kit for Illumina (New England Biolabs, Ipswich, Massachusetts). The library was sheared to approximately 400 bp using a Covaris M220 Ultrasonicator, amplified for eight cycles, and quantified using Quant-It Picogreen dsDNA assays (Thermo Fisher, Waltham, Massachusetts) run in triplicate. Fragment size was assessed using a genomic DNA Tape Station assay (Agilent, Santa Clara,

California). The library was sequenced in multiplex with three other Lake Trout in two HiSeqX lanes using the paired end 2x150 read format. Sequencing produced 316,557,707 read pairs for this individual.

ASSEMBLY AND SCAFFOLDING

Contig assembly using PacBio reads was carried out using the polished_falcon_fat assembly workflow run using the SMRT Analysis v3.0 pbsmrtpipe workflow engine provided with an installation of SMRT Link v5.0 (smrtlink-release_6.0.0.47841; https://github.com/PacificBiosciences/pbsmrtpipe). Read metadata were extracted using the SMRT Analysis v3.0 dataset tool with the merge option. Sequencing read metadata, pipeline settings, and an output directory were specified for the polished_falcon_fat pipeline option. Default assembly settings were used except genome size (HGAP_GenomeLength_str) was set to 3 gigabases (Gb), seed coverage (HGAP_SeedCoverage_str) was set to 40X, and the minimum read length to use a read as a seed (HGAP_SeedLengthCutoff_str) was set to 1000. Multiple settings were also changed. The resulting assembly settings file, read metadata file, and commands used to run the pipeline are available at https://github.com/smithsr90/LakeTroutGenome.

The polished_falcon_fat workflow uses FALCON assembly algorithm (Chin et al., 2013) and the Quiver/Arrow consensus tool (https://github.com/PacificBiosciences/GenomicConsensus) to generate a polished contig assembly. The Falcon method operates in two phases: First, overlapping sequence reads were compared to generate accurate consensus sequences with read N50 greater than 10.9Kb. Next, overlaps between the corrected longer reads were used to generate a string graph. The graph was reduced so that multiple edges formed by heterozygous structural

variation were replaced to represent a single haplotype. Contigs were formed by using the sequences of nonbranching paths. Two supplemental graph cleanup operations were applied to improve assembly quality by removing spurious edges from the string graph: tip removal and chimeric duplication edge removal. Tip removal discards sequences with errors that prevent 5′ or 3′ overlaps. Chimeric duplication edges may be produced due to the production of chimeric molecules during library preparation or during the first sequence cleanup step and these errors artificially increase the copy number of a duplication. In a second and final workflow stage, the polished_falcon_fat workflow used the Arrow consensus tool to perform error correction on the assembly using PacBio reads in order to generate an initial polished assembly. The resulting contigs were passed through a second round of error correction using Pilon in order to resolve SNP, indel, and local assembly errors before proceeding with scaffolding (https://github.com/broadinstitute/pilon). The Illumina paired-end sequencing dataset from a Seneca strain female (described above) was mapped to draft contigs using BWA mem with default settings (Li 2013). Reads with mapping qualities less than 20 were removed from the dataset in order to exclude low quality alignments and reads mapping to multiple locations. Improperly paired reads were also excluded using samtools view (Li & Handsaker, et al., 2009). The resulting filtered bam file was used as input for Pilon with the --fix all --mindepth 5, and --diploid options. Pilon was run prior to scaffolding in order to identify and correct local assembly errors that could potentially cause downstream scaffolding errors.

We adopted a multifaceted scaffolding approach leveraging information from Hi-C sequencing and a high-density linkage map for Lake Trout (Smith et al., 2020). Hi-C reads

were mapped to Pilon corrected contigs with default setting using the Arima Genomics Mapping pipeline (Arima Genomics, https://github.com/ArimaGenomics/mapping_pipeline), which included four primary steps. First, forward and reverse reads were mapped to the reference genome using bwa version 0.7.17 (Li, 2013) separately. Next, the 5' end of the mapped reads were trimmed. Samtools version 1.9 (Li & Handsaker, et al., 2009) was then used to filter reads with mapping qualities less than 10 in order to remove low quality alignments and reads mapping to multiple locations. Finally, Picard version 2.17.3 (https://broadinstitute.github.io/picard/) was used to add read group information and mark duplicate reads. The resulting BAM file was used as input for SALSA v2.2 (Ghurye et al., 2017) run with default settings (three iterations). We also tested Salsa2 using five iterations and compared results with those produced using default settings by calculating Spearman's rank order correlation coefficients between the order of loci on the Lake Trout linkage map (Smith et al., 2020) and the order of loci on the 50 largest scaffolds. Linkage mapped RAD contigs were aligned to the reference assembly using Minimap2 (Li 2018) using the -asm5 option. RAD contigs with mapping qualities less than 60 were removed before calculating correlation coefficients using the R function cor from the stats package (R Core Team, 2017) and the method argument set to "spearman."

Additional scaffolding was carried out using Chromonomer v1.13 (Catchen et al., 2020). The assembly was initially scaffolded using default settings, which yielded chromosome length scaffolds with a high degree of concordance with the linkage map; however, structural differences between the linkage map and scaffolds were apparent on six chromosomes. In order to resolve these inconsistences, we aligned the full set of PacBio

subreads to the assembly using Minimap2 (Li, 2018) using the preset option for PacBio data. The resulting bam file was sorted, indexed, and per-base coverage was calculated for all positions using samtools depth with the --aa option. We then ran a second round of Chromonomer using the --rescaffold, --depth, and depth_stdevs = 2 options, which allowed for gaps to be opened in contigs if the site-specific depth within a sliding window of 1000 base pairs was greater than 2 standard deviations from the mean, suggesting an assembly error. This resulted in an assembly with improved concordance with the linkage map; however, linkage group 41 still exhibited a large inversion relative to the scaffolds. We determined the approximate location of this assembly error by identifying the pair of linkage mapped loci for which the level of discordance between the linkage map and assembly was maximized. The scaffold was manually broken and reoriented using an existing gap that existed between these two loci.

Gaps were filled using PBJelly from PBSuite v15.8.24 (English et al., 2012). All PacBio reads were aligned to the draft assembly using Minimap2 using the -pb preset option and reads mapping within 5000 base pairs of a gap were retained for gap filling using bedtools intersect (Quinlan & Hall, 2010). Retained reads were re-mapped with Blasr v5.3.2 (Chaisson & Tesler 2012) using the options --minMatch 11, --minPctIdentity 75, --bestn 1, --nCandidates 10, --maxScore -500, and --fastSDP. The "maxWiggle" argument was set to 100 kilobases (Kb) for the PBJelly assembly stage in order to account for gaps of unknown length. After filling gaps, we corrected single nucleotide and short indel errors by running 3 iterations of Polca (distributed with MaSuRCA v. 3.4.2; Zimin & Salzberg, 2020) using Illumina data from a Seneca strain female as input. Polca was chosen because this error correction approach has been shown to be more effective for correcting single

nucleotide and indel errors than comparable tools (Zimin & Salzberg, 2020). Default settings were used except low quality alignments (MQ<10) and alignments overlapping gaps were removed from bam files using bedtools intersect (Quinlan and Hall 2010) prior to running the Polca variant calling step.

Illumina paired end data from the same individual used for genome polishing and PacBio data from one SMRTcell were aligned to the Arctic Char (Salvelinus alpinus) mitochondrial genome (RefSeq: NC_000861.1) in order to obtain reads useful for assembling the Lake Trout mitochondrial genome. Reads were aligned using Minimap2 using the sr and map-pb present options for short-reads and long-reads, respectively. Reads aligning to the Arctic Char mitochondrial genome were extracted from original fastq files using seqtk subseq (https://github.com/lh3/seqtk) and hybrid assembly was conducted using Unicycler v0.4.8 (Wick et al., 2017) using the settings --min_fasta_length 15000 and --keep 0. Unicycler implements a hybrid-assembly approach using Spades (Bankevich et al., 2012), SeqAn (Döring et al., 2008), and Pilon. First, Spades (v3.13.1) was used to assemble Illumina short-reads and contigs with graph coverage less than half the median coverage were removed due to potential contamination from the nuclear genome. Contigs were then scaffolded using long-reads and SeqAn (Döring et al., 2008) was used to generate gap consensus sequences. Finally, Pilon was used to resolve assembly errors using short-read alignments as input. The resulting mitochondrial genome assembly was aligned to all Salmonid sequences in the NCBI Nucleotide Collection using blastn and standard settings in order to verify that it was consistent with previous Lake Trout mitochondrial assemblies. A neighbor-joining tree constructed from blast pairwise alignments was exported from the NCBI website and is available in Figure S.2.6.

ASSEMBLY QUALITY CONTROL

We used multiple approaches to assess the accuracy, contiguity, and completeness of the genome assembly. First, we determined the proportion of the genome that was recovered in our assembly by comparing total assembly size with an estimate of genome size based on the distribution of k-mer frequencies from Illumina paired-end 2x150 data generated using DNA from a Seneca strain female. The frequency of all 19mers in the read data was calculated using the count function in Jellyfish v2.2.6 (Marçais & Kingsford, 2011) with the options -m 19 and -C. K-mer counts were then exported to the histogram format using the histo function. This file was used as input for GenomeScope v1.0 (http://qb.cshl.edu/genomescope/; Vurture et al., 2017) with read length set to 150 bp and k-mer length set to 19.

Basic assembly statistics were calculated using the program summarizeAssembly.py from PBSuite v15.8.24 (English et al., 2012). Statistics included total assembly size, contig and scaffold N50s, and minimum and maximum contig and scaffold lengths. Assembly statistics were calculated with and without gaps. Contig and scaffold N50s and counts were obtained for 14 additional salmonid assemblies from NCBI for comparison. Single base consensus accuracy was estimated during each iteration of polishing with Polca as the proportion of bases in input sequences overlapping detected errors.

Next, we calculated percentages of complete singleton, complete duplicated, fragmented, and missing Benchmarking Single Copy Orthologs (BUSCOs) for seven chromosome-level salmonid assemblies and compared these with scores for the Lake Trout assembly discussed here. These included genomes for Brown Trout (Salmo trutta; GCA_901001165.1), European Whitefish (Coregonus sp. balchen; GCA_902810595.1; De-

Kayne et al., 2020), Atlantic Salmon (Salmo salar; GCA_000233375.4; Lien et al., 2016), Coho Salmon (Oncorhynchus kisutch; GCA_002021735.1), Rainbow Trout (Oncorhynchus mykiss; GCA_002163505.1; Pearse et al., 2019), Chinook Salmon (Oncorhynchus tshawytscha; GCA_002872995.1; Christensen et al., 2018b), and Dolly Varden (Salvelinus malma; GCA_002910315.1; Christensen et al., 2018a). It should be noted that the assembly originally produced for Arctic Char (GCA_002910315.1; Christensen et al., 2018a, referred to as the Dolly Varden assembly here) was later found to be from a Dolly Varden or potentially a Dolly Varden – Arctic Char hybrid (see Shedko 2019 and Christensen et al., 2021). BUSCO scores were also calculated for the Northern Pike genome (Esox lucius; GCA_000721915.3; Rondeau et al., 2014), a member of the order Esociformes that is commonly used as a pre-Ss4R outgroup species. BUSCO scores were calculated using BUSCO v4.0.6, the actinopterygii_odb10 database (created November 20th, 2019), and the -genome option.

Finally, we aligned the linkage mapped contigs from Smith et al., (2020) to the final assembly and calculated Spearman's rank order correlation coefficients between physical mapping locations and the order of loci along linkage groups. Linkage mapped contigs were aligned to the reference assembly using Minimap2 using the -asm5 preset parameters and the resulting sam file was filtered to exclude contigs with mapping qualities less than 60. Correlation coefficients were calculated using the cor function in R (R Core Team, 2017) with the method argument set to "spearman." Correlation coefficients were then converted to absolute values using the abs function in order to compare chromosomes and linkage groups with reversed orientations.

REPETITIVE DNA

A custom repeat library was created using RepeatModeler v2.0.1 (Flynn et al., 2020) and repeats were subsequently classified using RepeatClassifier (Smit et al., 2015). Repeats were then masked using RepeatMasker (Smit et al., 2015) and the output of RepeatMasker was used to determine the genome-wide abundance of different repeat families and the relative density of repeat types across chromosomes. The density of the most abundant repeat type (Tcl-mariner) was visualized across chromosomes using the R-package circlize (Gu et al., 2014; Figure 2.2).

HOMEOLOG IDENTIFICATION AND SYNTENY

We performed a self-vs-self synteny analysis using SyMap v5 (Soderlund et al., 2006; Soderlund et al., 2011) to identify Lake Trout homeologs resulting from the Salmonid specific autotetraploid event (Macqueen & Johnston 2014; Lien et al., 2016). Prior to running SyMap, we hard-masked the genome using RepeatMasker v4.1.0 (Smit et al., 2015) using our custom repeat library as input and RMblast as the search engine (-e ncbi). Nucmer (Marçais et al., 2018) was used for SyMap alignments and options were set to min-dots = 30, top_n = 2, and merge_blocks = 1. We then used Symap to identify blocks of synteny between Lake Trout and Dolly Varden, Rainbow Trout, and Atlantic Salmon. These alignments were conducted using Promer (Marçais et al., 2018), and we used the options min_dots = 30, top_n = 1, merge_blocks = 1, and no_overlapping_blocks = 1. Results from self-vs-self synteny analysis were visualized using the R package circlize (Gu et al., 2014). Additionally, we identified syntenic relationships with Northern Pike using SynMap2 (Haug-Baltzell et al., 2017). We used the last algorithm to align genomes, DAGChainer to identify syntenic blocks (-D20, -A5), Quota Align Merge to merge syntenic blocks (-Dm 0),

and Quota Align (Overlap Distance = 40) to enforce a 1-to-2 ploidy relationship between Northern Pike and Lake Trout (Haas et al., 2004; Tang et al., 2011).

We also repeated our self-vs-self synteny analysis using SynMap2 while enforcing 2-to-2 synteny relationships. Sequence identity between homeologs was extracted from the output of SynMap2 for all merged regions composed of more than 1000 blocks. We then computed the moving average of local homeolog identity across chromosomes using sliding windows containing 200 blocks. We then fit a Gaussian mixture model to the distribution of homeolog identities using mixtools (Benaglia et al., 2009) and the function normalmixEM (k=2) after observing that the values were bimodally distributed. Posterior probabilities of assignment to clusters with high and low homeolog divergence were determined for each window in addition to cluster means and mixing proportions (lambdas) for the dataset. Results from the Lake Trout-vs- other salmonids synteny analysis were visualized using the Chromosome Explorer option in Symap v5. Syntenic relationships between Lake Trout and Northern Pike were visualized as a dotplot generated in R (Figure S.2.5).

RNA SEQUENCING AND GENE ANNOTATION

RNA samples derived from liver were obtained from the offspring of Seneca Lake hatchery strain fish held within the Ontario Ministry of Natural Resources and Forestry (OMNRF) hatchery system. Offspring were produced using four males and four females in a full factorial mating cross, by dry-spawning anesthetized fish (anesthetic: 0.1 g L-1 MS-222; Aqua Life, Syndel Laboratories Ltd., B.C., Canada). Eggs (140 mL) were stripped from each female, divided evenly among four jars, and fertilized by pipetting milt directly onto them. After fertilization, embryos were transported to the Codrington Fish Research Facility (Codrington, Ontario, Canada) where they were transferred from the jars into perforated

steel boxes with one family per box. These boxes were contained in flow-through tanks receiving freshwater at ambient temperature (5-6°C) and natural photoperiod under dim light. When the embryos fully absorbed their yolk sacs and were ready to feed exogenously (i.e., free embryos; approximately March 2016), 14 individuals from each family were randomly selected and split into two groups of seven, then transferred into one of four larger (200 L) tanks.

Tissue sample collection occurred between June 28 to August 9, 2016. The mean body mass and fork length of fish at sampling was 3.60 grams (SD=1.30) and 7.23 centimeters (SD=0.78), respectively. Each fish was euthanized in a bath of 0.3 g L-1 of MS-222 and dissected to remove the whole liver. The liver was gently blotted on a lab wipe and stored in RNAlater (Invitrogen, Thermo Fisher Scientific) for 24-48 hours at room temperature. RNALater was pipetted from the liver tissue and the samples were stored at -80°C until RNA isolation. Liver tissues were homogenized individually in 2 mL Lysing Matrix D tubes (MP Biomedicals) with 1 mL of Trizol reagent (Invitrogen, Thermo Fisher Scientific). RNA was extracted from the homogenate using phenol-chloroform extraction (Chomczynski & Sacchi, 2006). RNA was precipitated with RNA precipitation solution (Sambrook & Russel, 2001) and isopropanol, and washed with 75% ethanol. RNA samples were resuspended in nuclease-free water (Thermo Fisher Scientific). The purity and concentration of the RNA were initially determined using a NanoDrop-8000 spectrophotometer. RNA quality was also assessed using a Bioanalyzer (Agilent) and resulting RNA integrity numbers (RIN). All RNA samples met our minimum RIN threshold of 7.5.

RNA sequencing was performed over two years. Twenty-four samples were sent to The Centre for Applied Genomics (Sick Kids Hospital, Toronto, Ontario, Canada) in 2018, and another 30 samples were sent to the Centre d'expertise et de services Génome Québec (Montreal, Quebec, Canada; https://cesgq.com/) in 2020. cDNA libraries were produced by enriching the poly(A) tails of mRNA with oligo dT-beads using the NEBNext Ultra II Directional polyA mRNA Library Prep kit (New England Biolabs; Ipswich, Massachusetts). The group of 24 individuals was sequenced in 2.5 Illumina HiSeq 2500 lanes using 2X126 bp paired end reads. The additional thirty individuals were sequenced in three Illumina HiSeq 4000 lanes using 2X126 bp paired end reads. Data were deposited in sequence read archives associated with BioProject PRJNA682236. These sequencing reads, along with those from two previous RNAseq experiments for liver and muscle tissue (Goetz et al., 2010; Goetz et al., 2016: SRA Accessions SRS005644 and SRS387865), were used as input for NCBI's Eukaryotic Genome Annotation Pipeline (Thibaud-Nissen et al., 2016). A collection of 3547 Atlantic Salmon transcripts were also used as input. The annotation was produced by NCBI during January, 2021. For a complete record of data sources used for transcript alignments and gene prediction see: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Salvelinus_namaycush/100/).

RECOMBINATION RATES AND CENTROMERES

Sex averaged recombination rates were estimated across chromosomes using the sliding window interpolation approach implemented in MareyMap (Rezvoy et al., 2007). Restriction site associated DNA (RAD) contigs from the Lake Trout linkage map (Smith et al., 2020) were mapped to chromosomes using minimap2 using the -asm5 preset option and reads with mapping qualities less than 60 were removed. At this point, RAD loci

overlapping centromere mapping intervals for each linkage group were extracted and the centromere center was considered to be the mean mapping position for centromere associated RAD tags. Centromere positions were visualized using the R-package circlize (Gu et al., 2014).

In order to remove contigs with anomalous mapping positions that could bias recombination rate estimates, we fit a loess model describing linkage map position as a function of physical position for each chromosome, extracted model residuals, and removed markers with residuals that were greater than one standard deviation from the mean. Loess models were fit using the loess function in R with the span argument set to 0.2 and the degree argument set to 2. The remaining markers were output to MareyMap format and were manually curated using MareyMap Online (Siberchicot et al., 2017). A sex averaged recombination map was calculated using sliding window interpolation and exported from the program (Supplemental Material 2.1).

RESULTS

SEQUENCING, ASSEMBLY, AND SCAFFOLDING

Of 13,500 embryos exposed to UV irradiation and pressure shock treatments, only two individuals survived beyond post-embryo stage and one individual survived to a size of approximately 5 cm. This individual was found to be homozygous at all 15 genotyped microsatellite loci, suggesting that chromosome set manipulations were successful at inducing double haploidy. HMW DNA extraction yielded a DNA concentration of 70ng/ul based on nanodrop readings. We proceeded with PacBio sequencing, and produced a dataset with an estimated genome coverage of 89X, with 53X coverage provided by reads longer than 12 Kb in length.

The Falcon-based assembly pipeline and polishing with Arrow and Pilon yielded an initial assembly with 8,321 contigs, a total length of 2.3 Gb, and a contig N50 of 1.3 megabases (Mb) with a maximum contig length of 19.6 Mb (Table 2.1). Our analysis comparing the correlation between the Lake Trout linkage map and Hi-C scaffolds indicated that three iterations of Salsa (the default setting) produced moderately large scaffolds suitable for downstream use. We opted to use these settings for scaffolding. Salsa v2.2 split multiple contigs, resulting in 8,367 contigs with an N50 of 1.25 Mb and 5,171 scaffolds with an N50 of 5.15 Mb. Additional scaffolding with Chromonomer v1.13 increased scaffold N50 to 44 Mb and reduced the total number of scaffolds to 4,122. Chromonomer v1.13 also reduced contig N50 to a small degree due to the insertion of additional gaps at likely misassembles. Scaffolding with Hi-C and the Lake Trout linkage map ultimately allowed us to assign 84.7% of the genome to chromosomes. Gap filling with PBJelly increased scaffold N50 to 44.97 Mb, increased the total assembly size to 2.345 Gb, and increased contig N50 to 1.8 Mb (Table 2.1). Gap filling increased the maximum contig length to 34.78 Mb and the maximum scaffold length to 98.19 Mb. The consensus accuracy reported during the third round of error correction with Polca was 99.9959 %. The polished assembly was submitted to GenBank for public use (accession GCA_016432855.1).

ASSEMBLY QUALITY CONTROL

We estimated the total haploid genome size for Lake Trout to be between 2.119 and 2.122 Gb using k-mer analysis and GenomeScope v1.0, with 38% of the genome composed of unique sequence and 62% composed of repetitive sequence (Table S.2.1 and Figure S.2.1). Heterozygosity for the sample used for polishing was estimated to be between 2.78 and 2.9 heterozygous sites per 1000 base pairs. It should be noted that the individual used

for polishing was a diploid and not a gynogenetic double haploid. The estimated coverage for the sample used for genome-size estimation was 16X, which should be sufficient for k-mer based methods (Williams et al., 2013).

We recovered 93.2% of BUSCO genes with 60.3% and 32.9% being present as singletons and duplicates, respectively (Figure 2.3; Table S.2.2). The salmonid genomes evaluated recovered between 88.1% and 95.3% complete BUSCOs with between 25.3% and 34.9% being duplicated and between 58.3% and 65% being singletons. The proportion of duplicated BUSCOs in the Lake Trout genome was the second highest among the compared salmonid genomes (32.9%) and appears to be comparable to the Brown Trout genome (GCA_901001165.1; River Trout), which was also assembled using Falcon (Falcon-unzip) and polished using a method based on the Freebayes variant caller (Garrison and Marth 2012).

We found that the mitochondrial genome assembly produced here falls within a monophyletic group entirely composed of mitochondrial sequences previously generated for Lake Trout (Figure S.2.6; Schroeter et al., 2020). The assembly was most similar to one produced for a Lake Trout sampled from Lake Ontario, Pennsylvania (99.96 % Identity; Accession:MF621746.1). The Seneca Lake hatchery strain is heavily stocked in Lake Ontario and appears to have elevated fitness in this environment (Perkins et al., 1995).

The mean linkage map versus Hi-C scaffold Spearman's correlation was 0.89 across the 50 largest Hi-C scaffolds. These were calculated prior to integrating linkage information from the map. Thirty-three of the 50 largest Hi-C scaffolds had correlations greater than 0.95 and 42 had correlations greater than 0.8. Spearman's rank order correlations between finished chromosomes and the linkage map ranged from 0.89 to 1.0 for the 42 Lake Trout

chromosomes. High correlation coefficients are expected in this case because the linkage map was used to scaffold chromosomes. The mean correlation coefficient was 0.98 and 39 of 42 finished chromosomes had correlations greater than or equal to 0.96, suggesting that the linkage map and genome assembly provide a concordant representation of the order of loci along chromosomes (Figure 2.2E).

REPETITIVE DNA

RepeatModeler 2 identified 2,810 interspersed repeats and 462 of these were classified by RepeatClassifier. RepeatMasker reported that 53.8% of the Lake Trout genome is composed of sequences from this repeat library. A total of 13.04% of the genome was composed of retroelements, with 10.47% being LINEs and 2.57% being LTR elements, and 9.97% of the genome was composed of DNA transposons. As has been observed in other salmonids, TcMar-Tc1 was the most abundant superfamily and these repeats were most abundant near centromeres (Figure 2.2; Lien et al., 2016; Pearse et al., 2019). A total of 30.79% of the genome was composed of interspersed repeats that were not classified by RepeatClassifier (Table 2.2).

HOMEOLOG IDENTIFICATION AND SYNTENY

Self-vs-self synteny analysis conducted using Symap v5 identified 126 syntenic blocks shared between putative Lake Trout homeologs (Figure 2.2). Blocks ranged in size from 477,153 bp to 57,126,662 bp. Fifty-two blocks were longer than 10 Mb and 70 were longer than 5 Mb (Figure 2.2, inner links). The distribution of local homeolog identity was bimodal and our Gaussian mixture model estimated that the means of these two distributions were 82.72% and 90.64%. The lambda estimates for the model were 0.5848 and 0.4152 suggesting that approximately 41.52% of the Lake Trout genome exhibits a

signal of delayed re-diploidization (Figure 2.2D). Loci with elevated homeolog sequence identity were primarily located near the telomeres of metacentric chromosomes (Chr1-Chr8); however, one pair of acrocentric homeologs (Chr23 and Chr34) also exhibited elevated sequence identity.

We identified 50 syntenic blocks shared between Rainbow Trout and Lake Trout and identified homologous Rainbow Trout chromosomes for all Lake Trout chromosomes. Syntenic blocks shared between these two species ranged in size from 1.9 Mb to 97.2 Mb. Symap identified homologous chromosomes in Atlantic Salmon for all chromosomes except Lake Trout chromosomes 32 and 39. However, we expect that Lake Trout chromosome 39 is homologous to a region of Atlantic Salmon chromosome 2 and Lake Trout chromosome 32 is homologous with a region of Atlantic Salmon chromosome 14 based on the size of missing synteny blocks. Specifically, Lake Trout chromosomes 32 and 39 are 37.24 Mb and 23.59 Mb in length, respectively. The two regions with missing homology in Atlantic Salmon on chromosomes 2 and 14 are approximately 27.9 Mb and 42.9 Mb, respectively. Fifty-four syntenic blocks were detected between these species that ranged in size from 208,516 bp to 88 Mb. We identified 42 syntenic blocks shared between Dolly Varden and Lake Trout and identified homologs for all chromosomes except Chr 41. Syntenic blocks ranged in size from 6.8 Mb to 79.9 Mb. Pre-Ss4R ancestral chromosomes were also detected in Northern Pike (Figures S.2.2-S.2.5).

GENOME ANNOTATION

We generated a total of 3.45 billion RNA-seq reads from liver tissue that were subsequently used as input for the NCBI Eukaryotic Genome Annotation Pipeline v8.5 (July 9, 2020 release date). An additional 528,760 reads were used from previous Lake Trout

gene expression studies. A total of 86% of reads were aligned to the genome assembly, and 12 Lake Trout transcripts from GenBank and 3,547 known Atlantic Salmon transcripts from RefSeq were ultimately used as input for the pipeline.

The pipeline produced annotations for 49,668 genes and pseudogenes. A total of 3,307 non-transcribed pseudogenes and two transcribed pseudogenes were identified. Gene length ranged from 53 to 1,198,409 bp, with a median length of 8,676 bp. Gene densities for chromosomes ranged from 15.45 to 31.39 genes/Mb with an average genome-wide density of 21.07 genes/Mb (Figure 2.2C). A total of 422,014 exons were identified, with between 1 and 224 exons per transcript (mean=10.31, median=8).

RECOMBINATION RATES AND CENTROMERES

We were able to map between 1 and 238 centromere-associated RAD contigs per chromosome and determine approximate centromere locations for all chromosomes except chromosome 42 (Table S.2.4; Figure 2.2A). Smith et al. (2020) did not determine the location of the centromere for chromosome 42, which prohibited us from identifying its location here. Across all chromosomes, we mapped 35 centromere-associated RAD loci per chromosome on average. Between 39 and 238 centromeric loci were mapped to metacentric chromosomes (mean = 93), while between 1 and 59 loci were mapped for acrocentric or telocentric chromosomes (mean = 21).

In all, 14,438 linkage-mapped contigs were mapped to the genome with mapping qualities greater than 60 (Figure 2.2E). A total of 11,232 loci were retained for recombination rate estimation after manual curation and filtering using loess model residuals. We determined the mean sex averaged recombination rate to be 1.09 centimorgans/Mb, with recombination rates varying between 0 and 6.58 centimorgans/Mb

across the genome. The interpolated recombination map produced by MareyMap is available in Supplemental Material 2.1 – Recombination Map.

DISCUSSION

The adoption of multiple complementary scaffolding approaches resulted in an assembly of similar quality to the best available salmonid genomes. Multiple lines of evidence suggest that the genome presented here represents a nearly complete and accurate model of the female Lake Trout genome. First, the total size of the finished genome was slightly greater than the genome size estimate obtained from GenomeScope (2.3 Gb vs. 2.1 Gb). Pflug et al. (2020) found that k-mer based methods for genome size estimation tend to underestimate genome size by 4.5% on average, so this result is not entirely unexpected. Additionally, BUSCO scores were similar to those obtained for the highest quality salmonid genomes available at the time of analysis. Among the genomes examined, Brown Trout, Lake Trout, Atlantic Salmon, and European Whitefish had the highest proportion of complete BUSCOs (95.3, 93.2, 92.2, and 91.7 percent, respectively). Overall, Lake Trout BUSCO scores were most similar to those obtained for Brown Trout; however, the proportion of missing BUSCOs was 1.9% higher for Lake Trout and the proportion of complete duplicated BUSCOs was 2% lower suggesting that some duplicated regions might be missing from the Lake Trout genome. Nonetheless, these two assemblies had the highest percentage of complete BUSCOs and the highest percentage of complete duplicated BUSCOs out of the genome assemblies examined, suggesting that these two assemblies more effectively resolve LORe regions with high sequence similarity. Furthermore, the order of loci on the Lake Trout linkage map and the order of loci on Lake Trout chromosomes was shown to be highly concordant; however, it should be noted that

the linkage map cannot be considered an independent source of validation. The genome presented here is also highly contiguous, with a contig N50 higher than any published salmonid genome at the time of analysis (but see the recently released assemblies for Arlee Strain Rainbow Trout - GCF_013265735.2 and Atlantic Salmon - GCA_905237065.2).

Interestingly, the PacBio data used for assembly were of similar coverage to the data used for assembling the European Whitefish genome (De-Kayne et al., 2020); however, the Lake Trout genome contig N50 is >3X higher (1.8Mb vs. 0.53 Mb; Table S.2.3). It is worth noting that this assembly was produced using a different assembler (wtdbg2; Ruan & Li 2020); however, an assembly with a contig N50 of 211 Kb was also generated from these data using Falcon (De-Kayne et al., 2020). There are at least two reasonable explanations for the pronounced difference in contig N50 between the Lake Trout genome and European Whitefish assemblies produced using Falcon and wtdbg2. First, the European Whitefish genome was assembled using DNA from a wild-caught, outbred individual rather than a double haploid. Second, the European Whitefish genome was not gap filled after scaffolding. Gap filling the Lake Trout genome with PBJelly increased contig N50 by 561,496 bp, which partially explains the difference.

Additionally, our analysis of homeolog sequence identity across the Lake Trout genome indicates that regions exhibiting delayed rediploidization (ie. LORe regions; Robertson et al., 2017) are primarily associated with metacentric chromosomes and their acrocentric homeologs in Lake Trout. We identified one pair of acrocentric homeologs with elevated sequence identity (Chr23 and Chr34). Similar to Lien et al. (2016), these results suggest that homeologous pairing might not necessarily require one chromosome to be metacentric as suggested by Kodama et al. (2014). Interestingly, the region with elevated

homeolog sequence identity on chromosome 4 does not appear to be associated with one of the telomeres. A previous quantitative trait locus mapping study suggested that this chromosome harbors the sex determining gene (SdY) in Lake Trout (Smith et al., 2020).

It is important to note that the assembly presented here does not represent a true haploid assembly even though contigs were assembled using DNA from a double haploid individual. The assembly was error corrected using sequencing reads from a diploid female, the linkage map was generated using families from multiple hatchery strains (Smith et al., 2020), and Hi-C data were generated from a diploid individual from a separate strain. Therefore, the chromosome sequences presented here represent consensus sequences for female Lake Trout (from the Seneca L. strain) rather than haplotypes existing within the DH individual we sequenced. Additionally, PacBio assemblies are known to have an elevated prevalence of short indel errors relative to short read assemblies. These errors can interfere with annotation and necessitate error correction using short-read data (Watson & Warr, 2019). For this assembly, we excluded multimapping reads with low mapping qualities in order to avoid homogenizing variation between homeologs during error correction. This could have resulted in an elevated prevalence of short indel errors within duplicated regions with high homeolog sequence identity, which could make it more difficult to annotate genes in these regions.

Nonetheless, the genome presented here represents a significant improvement compared to existing genomic resources for the genus Salvelinus (Figure 2.3, Table S.2.2-S.2.3). Improvements could likely be made to the assembly using supplementary scaffolding resources such as a higher density linkage map or optical map (Pan et al., 2020). The annotation could also be improved by generating additional RNA-seq data.

Nevertheless, the number of annotated genes and pseudogenes (n=49,668) is similar to what has been obtained for other recent salmonid assemblies (e.g., Chum salmon, Oncorhynchus keta, GCF_012931545.1, n = 45,643; Sockeye salmon, Oncorhynchus nerka, GCF_006149115.1, n= 46,184; Dolly Varden, Salvelinus sp., GCF_002910315.2, n=46,775) using the same annotation pipeline. However, it is important to note that annotation completeness is markedly reduced relative to other assemblies with similar BUSCO scores such as Atlantic Salmon (57,783; GCF_000233375.1; Annotation Release 100), Coho Salmon (63,465; GCF_002021735.2; Annotation Release 101), Brown Trout (61,583; GCF_901001165.1; Annotation Release 100), Rainbow Trout (55,630, GCF_002163495.1, Annotation Release 100), and Chinook Salmon (53,685, GCF_002872995.1, Annotation Release 100). These annotations were produced using RNA-seq evidence from a greater diversity of tissue types, which likely explains this discrepancy. The Lake Trout annotation, as well as annotations for other salmonids, could also be further improved by directly sequencing full length transcripts using long-read sequencing technologies (Workman et al., 2019). We predict that the completeness of the Lake Trout genome annotation will be improved as more gene expression data from a greater diversity of tissue types becomes available for the species (Salzberg, 2019). Nonetheless, the current genome annotation will undoubtably aid in the interpretation of future findings by allowing researchers to link signals of selection and loci associated with phenotypes with putatively causal genes and biological processes. Publicly available gene expression and functional annotation resources, like those being developed by the Functional Annotation of All Salmonid Genomes (FAASG) initiative, will also aid in this effort (Macqueen et al., 2017).

The availability of a second high-quality genome assembly for a Salvelinus species will likely benefit comparative genomic research aimed at understanding the evolutionary consequences of genome duplication. Salmonids have long been appreciated as a model system for understanding evolution following whole genome duplication (Ohno, 1970) and a variety of recent studies have utilized the wealth of genomic resources for salmonids to shed light on the evolutionary processes at play following autotetraploid genome duplication events (see Gundappa et al., 2021 and Gillard et al., 2021). Additionally, multiple recent studies have highlighted the importance of structural genetic variation for promoting adaptive diversification within salmonid species (Pearse et al., 2019; Bertolotti et al., 2020), and chromosome-anchored genome assemblies are typically needed for detecting and genotyping structural variants (Mérot et al., 2020).

Genomic methods have dramatically increased the precision of population genetic analyses and have enabled researchers to address qualitatively unique questions that require some knowledge of genome structure and function (Waples et al., 2020). The genome assembly presented here will enable researchers to identify loci and candidate genes associated with phenotypic differentiation and reproductive isolation among Lake Trout ecotypes. Additionally, this resource will allow for the identification of loci associated with variation in fitness between Lake Trout hatchery strains in contemporary Great Lakes environments (Scribner et al., 2018; Larson et al. 2021) and loci that are adaptively diverged between hatchery strains. This information could help fisheries managers to maximize adaptive genetic diversity in re-emerging wild populations and prioritize hatchery populations for continued propagation. Overall, the availability of a high-quality

reference genome for Lake Trout will likely have important implications for ongoing conservation projects in the Great Lakes region and elsewhere.

APPENDIX

Figure 2.1 – The study species. Photograph of an adult Lake Trout (*Salvelinus namaycush*) from Great Bear Lake, Northwest Territories, Canada. Photo credit: Andrew Muir.

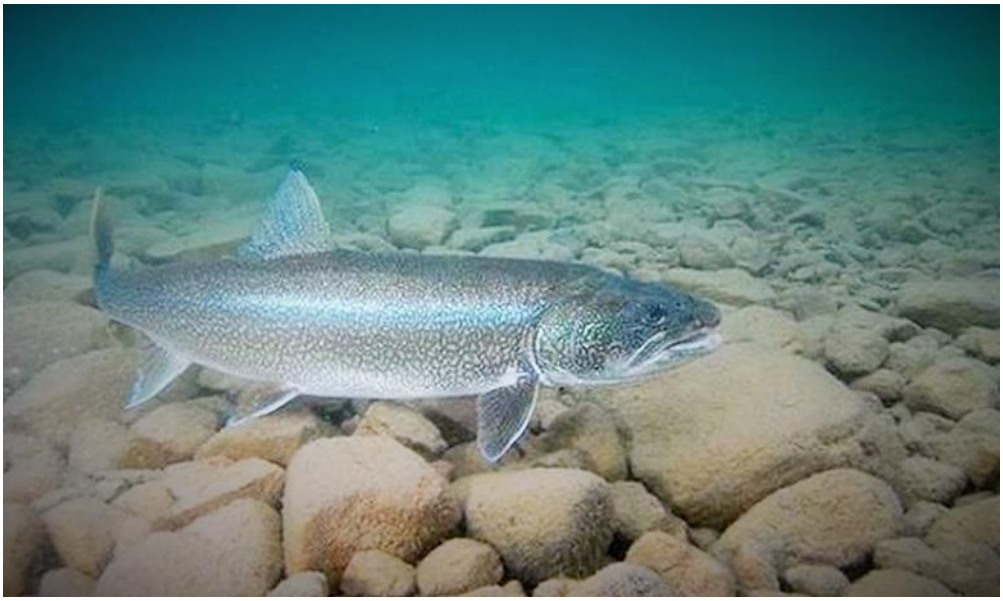

Figure 2.2: Circos plot displaying centromere positions, Tcl-Mariner abundance, density of annotated protein coding genes, local homeolog sequence identity, male and female Lake Trout (*Salvelinus namaycush*) linkage maps, and homeolog pairs resulting from Ss4R. (A) Black boxes in the outside ring display the mean mapping positions (+/- 5 Mb) for centromere associated RAD loci from Smith et al., (2020). (B) The second ring displays Z-transformed Tcl-Mariner repeat abundance in 5 Mb sliding windows with an offset of 100 kilobases. (C) The third ring displays the density of annotated genes in 5 Mb sliding windows with an offset of 100 kilobases. (D) The fourth ring displays local homeolog identity between syntenic blocks detected by SynMap2. Red points correspond to windows with elevated sequence identity putatively resulting from delayed re-diploidization (posterior probability > 0.5). Blue points correspond to windows with elevated sequence divergence between homeologs. (E) The fifth ring displays map distance (centimorgans) for male (red) and female (blue) linkage maps (y-axis) versus physical distance (x-axis) for each of the 42 chromosomes. Connections are drawn between syntenic blocks identified by SyMap v5 putatively resulting from Ss4R.

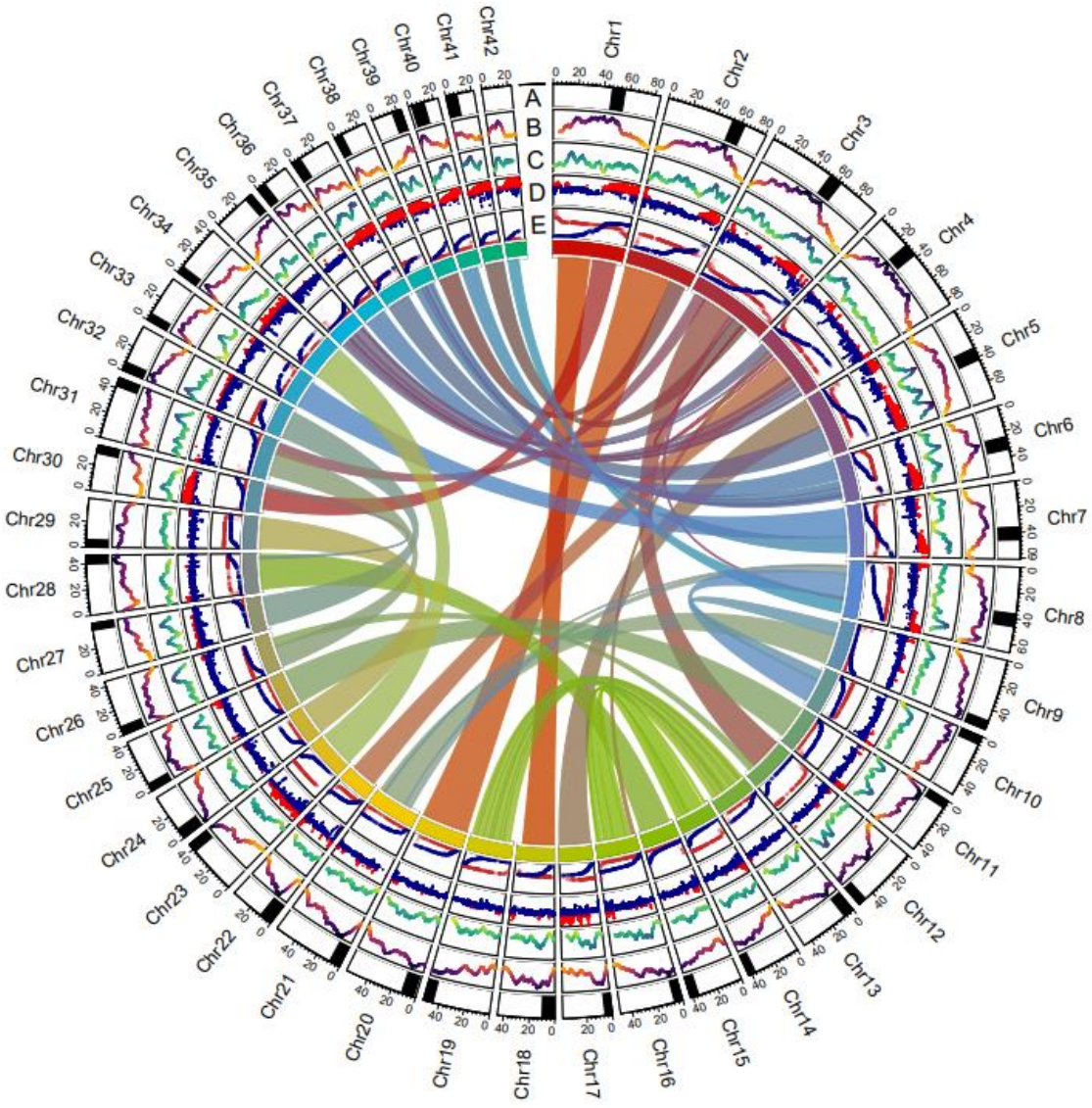

Figure 2.3: Comparison of BUSCO scores across multiple chromosome-level salmonid assemblies. Scores for the pre-duplication outgroup species (Northern Pike; Esox lucius) are also included for comparison. Assemblies are listed top-to-bottom according to the total percentage of complete BUSCOs. Complete single-copy, complete duplicated, fragmented, and missing BUSCO percentages are delineated with green, blue, yellow, and red bars, respectively.

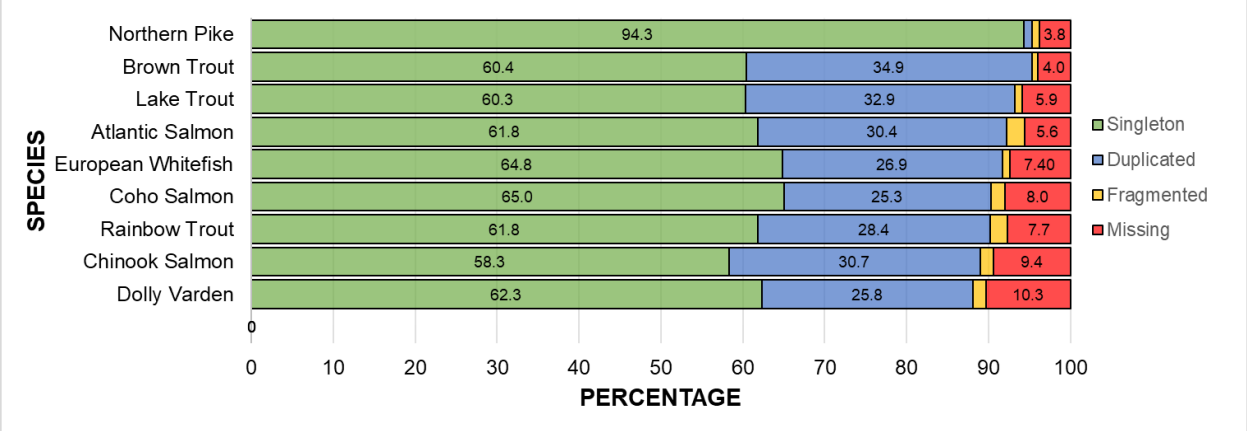

Table 2.1: General summary statistics for the Lake Trout (*Salvelinus namaycush*) genome assembly. The total number of chromosomes, scaffolds (including chromosomes), and contigs are listed in the top row. Metrics reported for chromosomes and scaffolds include gaps of unknown length. Consensus accuracy was obtained from the output of POLCA after running three iterations of the program.

	Chromosomes	**Scaffolds**	**Contigs**	**Gaps**
Count	42	4,120	7,378	3,258
Minimum Length (bp)	22,041,605	9,606	84	100
Mean Length (bp)	47,175,710	569,295	317,859	100
Max Length (bp)	98,200,354	98,200,354	34,788,501	100
Total Length (bp)	1,981,379,816	2,345,496,355	2,345,170,555	325,800
N50 (bp)	48,336,861	44,976,251	1,804,090	100
N90 (bp)	34,530,387	249,999	114,532	100
N95 (bp)	26,015,404	84,453	61,568	100
Consensus Accuracy (%)	-	-	99.9959	-

Table 2.2: Number of elements, total sequence length, and percent of the Lake Trout (*Salvelinus namaycush*) genome occupied by retroelements, DNA transposons, and other repeat types.

		No. Elements	Length	Percent
Retroelements:		**551376**	**305755720**	**13.04**
	SINEs:	**0**	**0**	**0.00**
	Penelope:	**11724**	**3138292**	**0.13**
	LINEs:	**483866**	**245479169**	**10.47**
	CRE/SLACS	0	0	0.00
	L2/CR1/Rex	337340	178461635	7.61
	R1/LOA/Jockey	9131	2778587	0.12
	R2/R4/NeSL	705	573357	0.02
	RTE/Bov-B	28238	14293769	0.61
	L1/CIN4	12257	6142123	0.26
	LTR Elements:	**67510**	**60276551**	**2.57**
	BEL/Pao	1533	1173630	0.05
	Ty1/Copia	1427	1007823	0.04
	Gypsy/DIRS1	55237	49788865	2.12
	Retroviral	9313	8306233	0.35
DNA Transposons:		**533707**	**233872078**	**9.97**
	hobo-Activator	34814	15807935	0.67
	Tc1-IS630-Pogo	473487	209441783	8.93
	En-Spm	0	0	0.00
	MuDR-IS905	0	0	0.00
	PiggyBac	9091	3370797	0.14
	Tourist/Harbinger	3105	834759	0.04
Other (Mirage, P-elements, Transib):		**1104**	**292535**	**0.01**
Rolling-Circles		**348**	**227654**	**0.01**
Unclassified:		**2885512**	**722299456**	**30.79**
All Interspersed Repeats:			**1261927254**	**53.80**

Table S.2.1: GenomeScope Output

GenomeScope version 1.0		
k = 19, Read Length = 150, Max Coverage = -1		
Sample SLW_52_F		
Property	**Minimum**	**Maximum**
Heterozygosity	0.278%	0.290%
Genome Haploid Length	2,119,589,342	2,122,166,134
Genome Repeat Length	1,316,156,520	1,317,756,576
Genome Unique Length	803,432,822	804,409,558
Model Fit	92.373%	99.196%
Read Error Rate	0.0288%	0.0288%

Figure S.2.1: GenomeScope Output

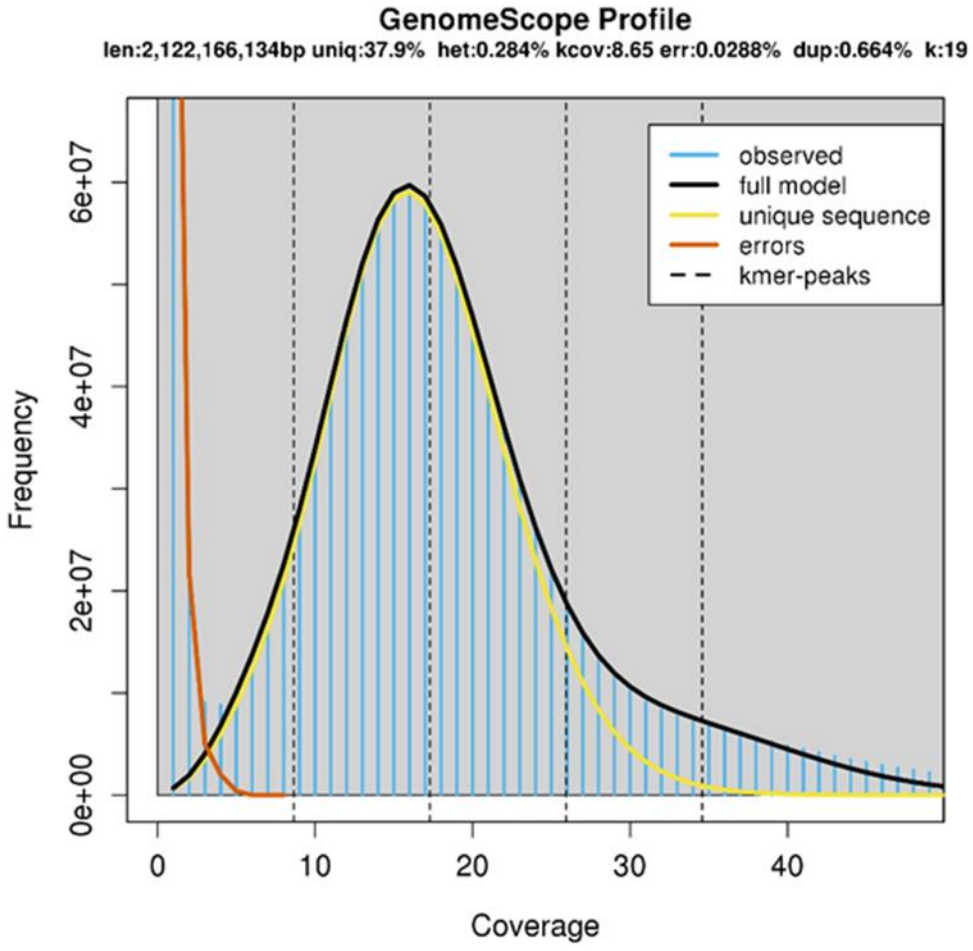

Figure S.2.2: Syntenic relationships between Lake Trout and Dolly Varden (previously Arctic Char) assemblies. The circos plot below identifies syntenic blocks shared between the Lake Trout and Dolly Varden genomes. Links are drawn between homologous regions in the two assemblies. Syntenic blocks were identified using SyMap version 5. Genomes were aligned using Promer and we used the Symap options min_dots = 30, top_n = 1, merge_blocks = 1, and no_overlapping_blocks = 1. The plot was generated using the Chromosome Explorer option in SyMap. A complete record of syntenic blocks between these two genomes is available in tab delimited format upon request.

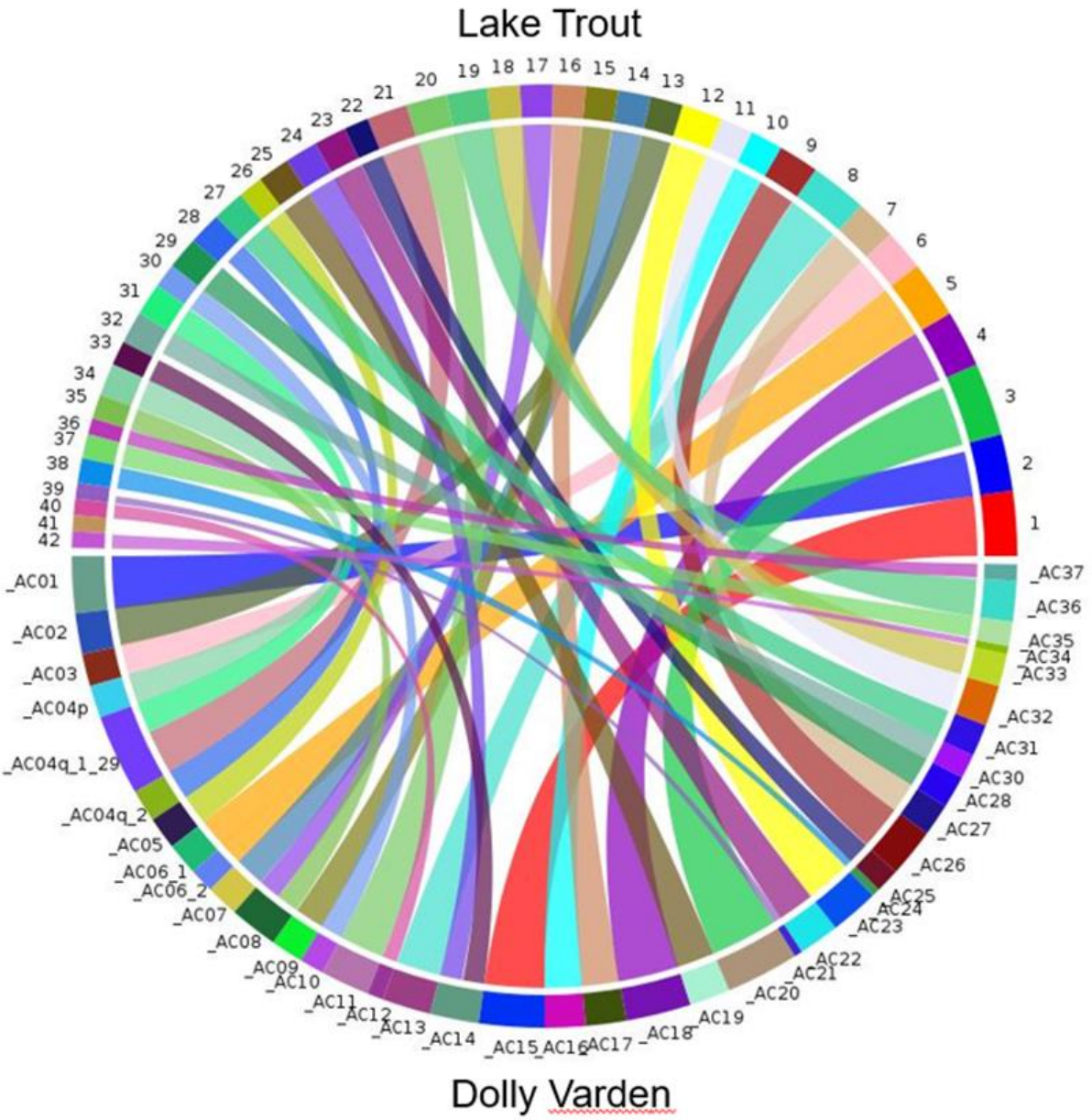

Figure S.2.3: Syntenic relationships between Lake Trout and Atlantic Salmon genome assemblies. The circos plot below identifies syntenic blocks shared between the Lake Trout and Atlantic Salmon genomes. Links are drawn between homologous regions in the two assemblies. Syntenic blocks were identified using SyMap version 5. Genomes were aligned using Promer and we used the Symap options min_dots = 30, top_n = 1, merge_blocks = 1, and no_overlapping_blocks = 1. The plot was generated using the Chromosome Explorer option in SyMap. A complete record of syntenic blocks between these two genomes is available in tab delimited format upon request.

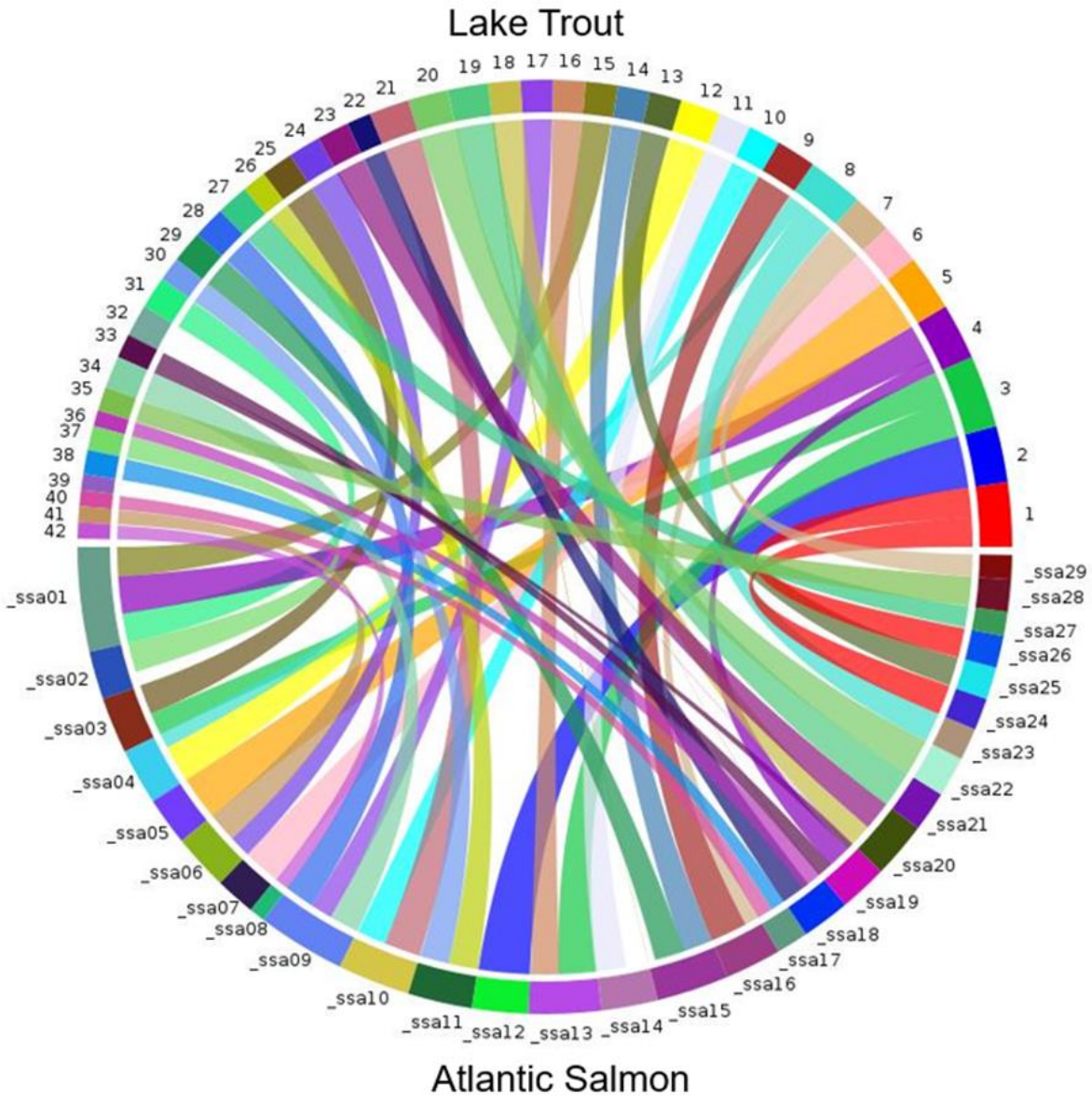

Figure S.2.4: Syntenic relationships between Lake Trout and Rainbow Trout genome assemblies. The circos plot below identifies syntenic blocks shared between the Lake Trout and Rainbow Trout genomes. Links are drawn between homologous regions in the two species. Syntenic blocks were identified using SyMap version 5. Genomes were aligned using Promer and we used the Symap options min_dots = 30, top_n = 1, merge_blocks = 1, and no_overlapping_blocks = 1. The plot was generated using the Chromosome Explorer option in SyMap. A complete record of syntenic blocks between these two genomes is available in tab delimited format upon request.

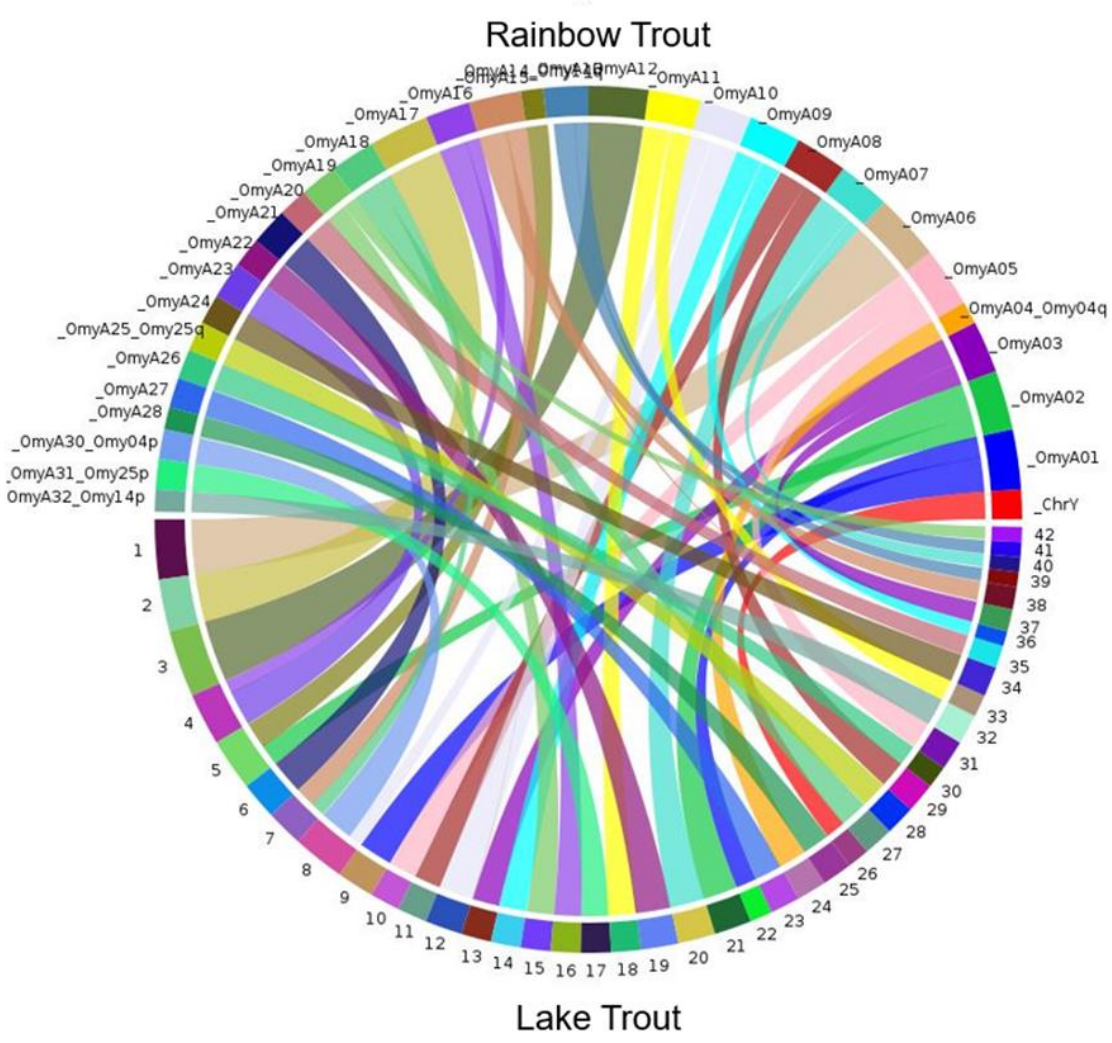

Figure S.2.5: Syntenic relationships between Lake Trout and Northern Pike genome assemblies. The dot plot below identifies syntenic blocks shared between the Lake Trout and Northern Pike genomes. Within SynMap2, we used the last algorithm to align genomes, DAGChainer to identify syntenic blocks (-D20, -A5), Quota Align Merge to merge syntenic blocks (-Dm 0), and Quota Align (Overlap Distance = 40) to enforce a 1-to-2 ploidy relationship between Northern Pike and Lake Trout.

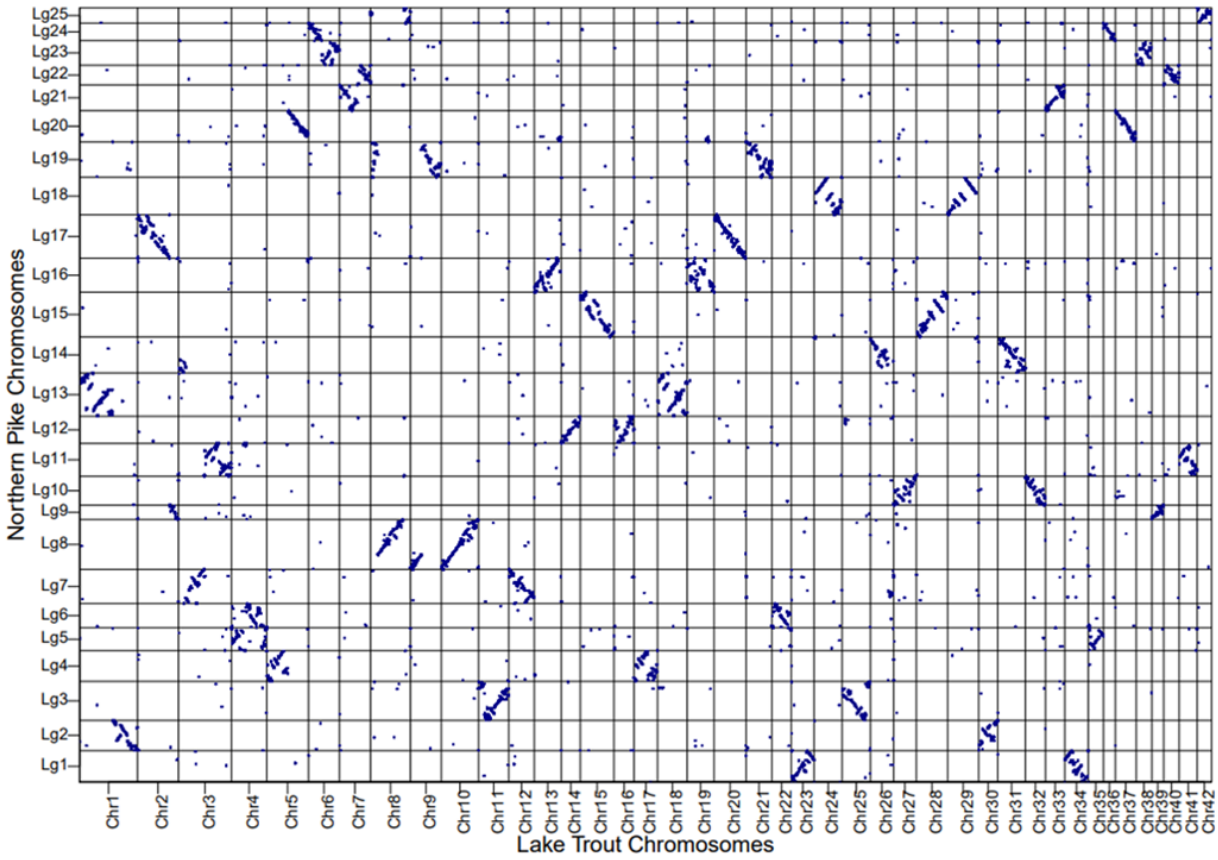

Figure S.2.6: Neighbor joining tree comparing the Lake Trout mitochondrial genome assembly with blast hits in the NCBI nucleotide collection. The sequence assembled here is highlighted in yellow.

Table S.2.2: Comparison of BUSCO scores among salmonid genomes

Species	Accession	Percent BUSCOs				
		Total Compete	Single Copy	Duplicated	Fragmented	Missing
Char	GCA_002910315.1	88.1	62.3	25.8	1.6	10.3
Chinook Salmon	GCA_002872995.1	89.0	58.3	30.7	1.6	9.4
Rainbow Trout	GCA_002163505.1	90.2	61.8	28.4	2.1	7.7
Coho Salmon	GCA_002021735.1	90.3	65.0	25.3	1.7	8.0
Whitefish	GCA_902810595.1	91.7	64.8	26.9	0.9	7.4
Atlantic Salmon	GCA_000233375.4	92.2	61.8	30.4	2.2	5.6
Lake Trout	-	93.2	60.3	32.9	0.9	5.9
Brown Trout	GCA_901001165.1	95.3	60.4	34.9	0.7	4.0
Northern Pike	GCA_000721915.3	95.3	94.3	1.0	0.9	3.8

Table S.2.3: N50 Comparison between salmonid genomes

Species	Resources	Type	Accession	Contig N50 (bp)	Scaffold N50 (bp)	No. Contigs	No. Scaffolds
Oncorhynchus mykiss	USDA_OmykA_1.1	Chromosome	GCF_013265735.2	15,579,713	39,165,350	1,229	939
Salvelinus namaycush	SaNama_1.0	Chromosome		1,804,090	44,974,654	7,378	4,120
Salmo trutta	fSalTru1.1	Chromosome	GCA_901001165.1	1,703,178	52,209,666	5,378	1,441
Oncorhynchus kisutch	Okis_V2	Chromosome	GCA_002021735.2	1,159,298	68,265,700	8,771	4,087
Coregonus balchen	AWG_v2	Chromosome	GCA_902810595.1	533,126	52,020,451	8,707	40
Oncorhynchus nerka	Oner_1.0	Chromosome	GCA_006149115.1	329,583	1,058,586	57,813	38,027
Oncorhynchus tshawytscha	Otsh_v1.0	Chromosome	GCA_002872995.1	133,169	1,728,323	69,485	15,946
Salmo salar	ICSASG_v2	Chromosome	GCA_000233375.4	57,618	1,366,254	368,060	241,573
Salvelinus sp.	ASM291031v2	Chromosome	GCA_002910315.2	55,619	1,018,695	97,014	16,702
Oncorhynchus keta	Oket_V1	Chromosome	GCA_012931545.1	50,313	28,109	117,794	28,109
Hucho hucho	ASM331708v1	Scaffold	GCA_003317085.1	37,639	287,338	221,746	71,639
Thymallus thymallus	ASM434828v1	Chromosome	GCA_004348285.1	31,774	32,985,317	204,386	3,831
Oncorhynchus tshawytscha	CHI06	Chromosome	GCA_002831465.1	19,113	153,278	234,121	115,115
Oncorhynchus mykiss	Omyk_1.0	Chromosome	GCA_002163495.1	13,827	1,670,138	559,854	139,799
Oncorhynchus mykiss	AUL_PRJEB4421_v1	Scaffold	GCA_900005705.1	9,390	383,627	221,128	79,942

Table S.2.4: Mean, median, minimum and maximum mapping positions for centromere associated RAD loci from the Smith et al. (2020) linkage map.

Chromosome	Mean	Median	Minimum	Maximum	SD	No. Markers	Centricity
Chr1	51,189,585	48,795,202	37,397,989	82,535,863	13,239,344	74	Metacentric
Chr2	57,816,187	58,747,044	26,935,986	79,541,631	6,773,213	238	Metacentric
Chr3	57,655,059	56,167,146	49,804,167	93,429,921	6,070,385	96	Metacentric
Chr4	32,737,543	32,297,665	22,548,690	82,690,766	7,117,895	66	Metacentric
Chr5	37,260,346	36,840,130	33,503,123	41,310,615	2,686,454	52	Metacentric
Chr6	29,444,214	30,120,981	23,975,449	32,146,173	2,495,169	60	Metacentric
Chr7	40,759,178	39,476,371	31,083,727	58,525,437	6,640,410	116	Metacentric
Chr8	41,354,659	40,619,039	38,829,067	44,485,090	1,751,458	39	Metacentric
Chr9	53,943,232	53,888,689	52,786,588	55,592,224	799,756	25	Acrocentric or Telocentric
Chr10	1,502,018	1,527,575	875,997	2,259,439	400,484	12	Acrocentric or Telocentric
Chr11	1,284,931	1,461,087	236,806	2,132,682	697,789	14	Acrocentric or Telocentric
Chr12	52,130,995	52,297,247	50,434,181	53,485,846	926,967	22	Acrocentric or Telocentric
Chr13	3,285,216	3,751,149	100,123	5,703,361	1,439,883	41	Acrocentric or Telocentric
Chr14	41,840,915	41,980,094	41,075,031	42,528,650	510,134	16	Acrocentric or Telocentric
Chr15	40,275,515	40,377,639	38,069,590	41,415,590	839,081	18	Acrocentric or Telocentric
Chr16	1,393,990	1,410,042	434,505	1,705,546	341,379	13	Acrocentric or Telocentric
Chr17	755,314	810,441	85,582	1,285,115	408,257	12	Acrocentric or Telocentric
Chr18	4,937,158	5,080,550	3,861,515	6,173,195	682,645	34	Acrocentric or Telocentric
Chr19	49,810,077	49,915,374	48,442,196	51,647,025	854,842	39	Acrocentric or Telocentric
Chr20	6,582,644	6,192,348	3,132,187	9,456,364	1,942,820	59	Acrocentric or Telocentric
Chr21	3,727,782	3,715,482	2,557,011	7,890,542	1,295,706	15	Acrocentric or Telocentric
Chr22	4,946,703	3,997,840	915,422	33,256,077	6,603,243	21	Acrocentric or Telocentric
Chr23	42,821,763	42,599,182	40,901,960	44,899,486	926,385	43	Acrocentric or Telocentric
Chr24	3,245,082	3,457,044	27,140	5,986,432	1,738,099	49	Acrocentric or Telocentric
Chr25	1,096,375	1,190,442	305,887	1,598,429	404,584	10	Acrocentric or Telocentric
Chr26	1,687,022	1,258,180	109,925	4,529,231	1,170,118	41	Acrocentric or Telocentric
Chr27	39,531,659	39,627,398	39,313,814	39,677,610	164,815	5	Acrocentric or Telocentric
Chr28	44,945,793	44,939,363	44,173,866	45,819,910	563,424	18	Acrocentric or Telocentric
Chr29	936,853	862,386	672,151	1,350,487	296,389	4	Acrocentric or Telocentric
Chr30	34,228,173	34,423,767	33,440,433	34,520,145	424,016	6	Acrocentric or Telocentric
Chr31	46,573,530	46,873,899	44,658,724	48,275,262	1,185,624	26	Acrocentric or Telocentric
Chr32	1,923,186	2,043,603	64,284	3,760,530	1,160,232	21	Acrocentric or Telocentric
Chr33	886,467	940,845	386,894	1,403,126	294,088	15	Acrocentric or Telocentric
Chr34	943,912	606,933	260,163	1,964,641	900,821	3	Acrocentric or Telocentric
Chr35	33,823,887	33,882,683	32,886,783	34,439,463	397,528	13	Acrocentric or Telocentric
Chr36	1,930,526	1,930,526	-	-	-	1	Acrocentric or Telocentric
Chr37	2,245,735	1,764,446	1,507,395	5,377,712	1,169,553	10	Acrocentric or Telocentric
Chr38	1,028,942	704,679	486,418	2,548,049	768,439	6	Acrocentric or Telocentric
Chr39	21,362,859	21,755,949	18,202,186	23,519,178	1,801,297	27	Acrocentric or Telocentric
Chr40	7,321,542	2,186,440	57,467	18,848,102	8,031,557	31	Acrocentric or Telocentric
Chr41	3,651,909	1,919,310	236,245	17,829,128	4,773,770	19	Acrocentric or Telocentric
Chr42	-	-	-	-	-	0	Acrocentric or Telocentric

Supplementary Material 2.1: This file contains an interpolated, sex averaged recombination map for Lake Trout in raw text format. This file is too large to usefully display in this document and is available upon written request to the author.

CHAPTER 3: THE GENOMIC BASIS FOR ECOMORPHOLOGICAL VARIATION IN LAKE SUPERIOR LAKE TROUT (SALVELINUS NAMAYCUSH)

ABSTRACT

We use a combination of conventional and low-coverage genotyping-by-sequencing methodologies to localize genomic regions associated with ecomorphological variation in native Lake Superior Lake Trout (*Salvelinus namaycush*). We identified 601 SNPs (out of 225,700 SNPs tested) that were significantly associated with ecomorphological differentiation based on concordant results from several selection-detection methods. Multiple islands of divergence spanning between 1 and 6.4 Mb were also detected based on results from our most conservative scan for selection. In some cases, we determined that islands of divergence were associated with putative chromosomal inversions based on patterns of linkage disequilibrium. These include a putative inversion on chromosome Sna1 spanning between 6.8 and 14.93 megabases. Interestingly, some of the strongest signals of adaptive divergence between ecomorphotypes were in close proximity to proteins involved in canonical and non-canonical Wnt signaling including Wnt5a, a Frizzled-1-like protein, and a Dishevelled 2-like protein, suggesting that ecomorphological divergence in Lake Trout might be partially associated with selection on proteins involved with this highly conserved signaling pathway. Additionally, we identified a group of SNPs associated with the lean ecomorphotype in close proximity to the vgll3 locus, which was found previously to be associated with sea-age at maturity in Atlantic Salmon. Multiple other candidate genes related to lipid metabolism, circadian rhythm, immune function, eye development, habituation, and carbohydrate metabolism were also identified. Additionally, we find that the Lake Superior metapopulation was primarily stratified by ecomorphotype rather than

sampling location in the 1990s and earlier. Interestingly, individual ancestry coefficients suggest that hybridization primarily occurred between leans and humpers and humpers and siscowets during the time period examined. Results from a limited dataset of historical samples from the Apostle Islands suggest that hybridization between ecomorphotypes increased substantially between the 1960s and 1990s.

INTRODUCTION

Adaptive radiation is a primary process by which diversity is generated in nature (Schluter, 2000). In many cases, radiations were driven by selective pressures on traits favoring niche segregation and the ability to exploit novel resources (Schluter,1996). Multiple classic examples have been reported of adaptive radiations resulting in the evolution of biological species (Butler, 2007; Schluter, 1996; Grant & Grant, 2020); however, the evolution of phenotypically distinct ecotypes that occupy separate niches is often considered to be an intermediate step towards the evolution of novel species (Lowrey, 2012). Ecotypes often exhibit divergent physiology, morphology, and behavior (Wood et al., 2008; Brawand et al., 2014) and this metapopulation diversity can promote population, community, and ecosystem resilience (Schindler et al. 2010). The maintenance of variation that facilitates adaptive diversification is critical to the goal of sustaining population viability and evolutionary potential (Teixeira & Huber, 2020).

Lake Trout (*Salvelinus namaycush*) is a highly diverse salmonid species (Muir et al., 2016) that has been severely impacted by anthropogenic activities over the last century–particularly in the Laurentian Great Lakes (Hansen, 1999). Prior to European settlement, Lake Trout were an abundant top predator in the Great Lakes and multiple distinct ecotypes existed in sympatry across the region, often with marked differences in trophic

niche and allocation of resources to growth, reproduction, and survival (Moore & Bronte, 2001; Zimmerman et al., 2009; Hansen et al. 2012; Chaverie et al., 2013; Goetz et al., 2014; Hansen et al., 2016a;). Four forms, commonly referred to as "ecomorphotypes", exist in present day Lake Superior (Muir et al., 2014), however, anecdotal evidence suggest that levels of diversity were higher historically (Goodier, 1981; Rakestraw, 1968). These extant ecomorphotypes are referred to as leans, siscowets, humpers, and redfins in Lake Superior. Multiple lines of evidence suggest that a diversity of forms also existed in Lakes Michigan, Huron, and Erie (Jordan & Evermann, 1923; Brown et al. 1981; Goodier 1981; Eshenroder et al., 1995; Hansen 1999); however, only a single form is currently present in these locations.

Declines in diversity and abundance were largely driven by increased fishing pressure following the collapse of the lake whitefish commercial fishery and the invasion of the Great Lakes by Sea Lamprey (*Petromyzon marinus*) in the first half of the 20th century. Dramatic increases in adult mortality ultimately led to the functional extirpation of the species from all lakes except Lake Superior and a small isolated population in Lake Huron (Hansen, 1999). Only the lean ecomorphotype was initially reintroduced throughout the Great Lakes (Hansen 1999).

Lean Lake Trout (Figure 3.1B) dominate nearshore, shallow water, habitats and were reintroduced across the Great Lakes following the collapse of native populations (Krueger & Ihssen, 1995; Eschenroder et al., 1995). These individuals are primarily piscivorous as adults, have pointed snouts, are typically found at depths less than 50 meters, have a streamlined body shape, and spawn in late autumn (Moore & Bronte, 2001; Khan & Qadri, 1970; Muir et al., 2014; Hansen et al. 2016). Siscowet Lake Trout (Figure

3.1C) typically occupy habitats greater than 80 meters in depth; are deep bodied; and have rounded snouts, large eyes, and high tissue lipid content (Muir et al. 2014; Sitar et al., 2008). Siscowets are known for feeding on Mysis shrimp (Mysis diluviana) and Coregonus species as they undergo diel movements through the water column (Ahrenstorff et al., 2011). Humper Lake Trout (Figure 3.1A) occupy offshore shoals, and have large eyes, reduced growth rates, mature when relatively young, and spawn in late Summer (Burnham-Curtis & Bronte, 1996; Rahrer, 1965; Muir et al., 2016). Overall, these forms are differentiated by spawning time, feeding behavior, physiology, morphology, multiple life history characteristics, and habitat occupancy (Hansen et al., 2016a).

Similar patterns of ecomorphological diversity have been documented in other lakes across Northern North America (Blackie et al., 2003; Zimmerman et al., 2006; Hansen et al., 2012; Marin et al., 2016; Chavarie et al., 2015), with a much larger diversity of forms existing in some lakes (Marin, et al., 2016). Divergent Lake Trout forms; often similar to leans, siscowets, and humpers; have been identified in Great Slave Lake, Northwest Territories (Hansen et al., 2016b); Lake Mistassini, Quebec (Hansen et al., 2012); Flathead Lake, Montana (Stafford et al. 2014); Great Bear Lake, Northwest Territories (Chavarie et al. 2013); and Rush Lake, Michigan (Muir et al. 2016). The similarity of forms across northern North America suggests that ecomorphological differentiation might be driven and maintained by a common set of selective pressures in multiple lakes. Interestingly, the rapid evolution of dwarf and siscowet-like Lake Trout has been observed in Flathead Lake, Montana over the last 100 years, following introduction of only one form (Stafford et al. 2014; Craig Stafford – Personal Communication). It is unclear if this rapid diversification is

due to epigenetic effects, phenotypic plasticity, or rapid changes in allele frequency resulting from selective pressures favoring niche segregation.

Patterns of population genetic structure in Lake Superior appear to have changed substantially between the 1990s and early 2000s. The results of Perrault-Payette et al. (2017) suggest that genetic variation in Lake Superior Lake Trout is primarily partitioned among sampling locations rather than between ecomorphotypes for contemporary time periods (2013-2014). Conversely, Guinand et al. (2012) evaluated patterns of population genetic structure using samples collected in 1995 and 1999 and found that ecomorphotypes represented genetically differentiated subpopulations during this time period. Additionally, Ballie et al. (2016) documented a 60.7% reduction in genetic differentiation among ecomorphotypes between recovery (1995-1999) and contemporary (2004-2013) sampling periods. They predicted that this was caused by increased hybridization associated with overlap between foraging and breeding habitat; however, hybridization dynamics between Lake Superior ecomorphotypes have not been thoroughly evaluated. The restoration of the deep-water community in the Great Lakes necessitates the quantification of genetic differentiation between ecomorphotypes at neutral and adaptive loci and an improved understanding of the causes and consequences of genetic homogenization (Zimmerman & Krueger, 2009).

Multiple previous studies have evaluated the genetic basis for ecomorphological variation in Lake Trout. Goetz et al. (2010) explored differences in gene expression between lean and siscowet Lake Trout and found that genes associated with immunity and lipid metabolism, transport, and synthesis were differentially expressed in the two forms. Importantly, they found that morphological and physiological characteristics were

maintained when ecomorphotypes were raised in a common garden setting, suggesting that phenotypic differences between these forms are due to genetic differences and not phenotypic plasticity.

Smith et al. (2020) constructed a linkage map for Lake Trout and identified quantitative trait loci (QTL) that underly differences in growth and condition related traits, skin pigmentation, and body shape; however, only lean Lake Trout were used to produce mapping families in this study. It is unclear if the QTL detected in this study are also associated with phenotypic differences between Lake Superior ecomorphotypes. Perrault-Payette et al. (2017) preformed a scan for loci under selection between lean, siscowet, humper, and redfin ecomorphotypes. They identified loci near genes associated with vision and lipid metabolism; however, they did not identify loci that were consistently associated with morphotype across sampling locations. This work was carried out using relatively low marker density (n=6822) SNPs, which could have hindered their ability to identify loci with consistent associations across populations. For instance, Barria et al. (2019) recently suggested that a minimum of 74,000 SNPs should be used for mapping genotype-phenotype associations in an aquaculture population of Coho Salmon, although this value will vary depending on the extent of linkage disequilibrium in populations. Evaluating adaptive differences between ecomorphotypes using higher marker density would increase the probability of detecting loci associated with adaptive differences and would increase our ability to localize potential causal genes within peaks of association. The recently completed Lake Trout genome and associated annotation would also greatly aid in this effort (Smith et al, 2021, in-press).

The genetic basis for morphological, physiological, and life history variation has been extensively studied in Pacific Salmon species (Oncorhynchus sp.) and Atlantic Salmon (*Salmo salar*). Two of the most well-known genes associated with life history variation in salmonids are greb1-l and vgll2. The greb1-l locus explains a significant proportion of variation in the pre-mature migration phenotype in Chinook Salmon and anadromous Rainbow Trout (Steelhead; Thompson et al., 2020; Prince et al., 2017). Additionally, the vgll3 locus explains a substantial proportion of variance in sea-age at maturity in Atlantic Salmon (Barson et al., 2015); however, this association does not appear to exist in Pacific Salmon species (Waters et al., 2021). The six6 gene has also been identified as a significant quantitative trait locus for the age-at-maturity phenotype in Atlantic Salmon (Sinclair-Waters et al., 2020), Sockeye Salmon, and anadromous Rainbow Trout (Waters et al., 2021). Lake Trout ecomorphotypes differ with respect to spawning time (Burnham-Curtis & Smith 1994) and maturation age (Rahrer 1965; Madenjian et al., 1998; Hansen et al., 2012); however, it is unclear if the genetic basis for these traits is shared with other salmonid species.

Multiple studies have identified large islands of divergence associated with ecotypic and life history variation within salmonid species (Larson et al., 2017; Kess et al., 2021). Probably the most extreme example is a 55 Mb double inversion on chromosome Omy5 in Rainbow Trout that is strongly associated with sex-specific migratory behaviors (Miller et al., 2012; Pearse et al., 2019). Large chromosomal inversions have also been found to be associated with adaptive divergence between Atlantic Cod ecotypes (Berg et al., 2017). Inversions inhibit recombination between inverted and non-inverted haplotypes and can act to maintain adaptive alleles on the same chromosome in strong linkage (Merot, 2020;

Wellenreuther & Bernatchez, 2018). To date, no study has sought to identify polymorphic structural variants within Lake Trout or test whether or not adaptively diverged regions are associated with structural variants in this species.

Here we provide a thorough characterization of the adaptive genetic variation that facilitated the exploitation of multiple niches by Lake Superior Lake Trout. We utilize the recently developed Lake Trout genome (Smith et al. 2021) and linkage map (Smith et al. 2020) to identify genomic regions, genes, and biological pathways associated with adaptive differences between leans, humpers, and siscowets because similar forms commonly occur across the species range. Additionally, we sought to determine whether or not any detected structural variation or islands of divergence were associated with chromosomal inversions. Our secondary goal was to characterize the biological pathways associated with ecomorphological divergence using gene-set enrichment analysis (GSEA; Alexa & Rahnenführer, 2009). We specifically hypothesized that adaptively diverged regions would disproportionately contain genes related to phenotypic and physiological traits widely recognized as definitive features of each ecomorphotype including lipid metabolism, locomotion, eye development, circadian rhythm, and somatic growth. We were also interested in evaluating population structure in Lake Superior using samples from the 1990s and earlier. We were specifically interested in characterizing levels of hybridization between ecomorphotypes in the 1990s and determining if levels of hybridization have increased through time.

MATERIALS AND METHODS

LABORATORY METHODS AND SAMPLES

Samples were obtained from the collection described in Page et al. (2004) including lean, siscowet, and humper Lake Trout collected from the Apostle Islands, Isle Royale, Stannard Rock, Whitefish Point, and Caribou Reef during the summer and fall of 1995 (Figure 3.2). We supplemented this dataset with historical scale samples collected from 38 lean and siscowet Lake Trout collected near the Apostle Islands between 1969 and 1986. Samples from 1994 and 1995 were originally extracted from liver tissue stored in urea buffer using the Puregene extraction protocol (Gentra, Inc). Scale samples were extracted using the bead-based protocol described in Ali et al. (2001) with Serapure beads (Rohland and Reich, 2012) substituted for Ampure beads. Additionally, Serapure beads were prepared with a 3X higher concentration of magnetic particles compared with the original protocol of Rohland and Reich (2012). Sample quality and quantity were initially assessed using a Nanodrop 2500 instrument. Double stranded DNA concentrations were determined using Quant-it Picogreen assays (Life Technologies). Samples were purified and concentrated 2-fold using Serapure beads if dsDNA concentrations were less than 5ng/ul or if Nanodrop 260/280 absorbance ratios were less than 1.8.

BestRAD libraries were prepared using the protocol from Ali et al. (2016) with modifications described in Smith et al. (2020). Libraries were quantified using Quant-it Picogeen assays run in triplicate before pooling equal amounts of DNA from each library. Pooled libraries were enriched for 58,889 variable Pst1 RAD loci (Smith et al. 2020) using a MyBaits V3 Custom Target Enrichment Kit. The target enriched pool was amplified for 10 cycles using the KAPA Library Amplification Kit for Illumina using manufacturer

recommended PCR conditions. Amplified DNA was purified twice using 0.9:1 Ampure XP clean-ups and eluted in low-EDTA TE buffer. The library was sequenced in a single HiSeq 4000 lane at the Michigan State University Research Technology Support Facility using 2X150 base pair paired-end reads.

BIOINFORMATICS

Data quality and quantity was initially assessed using FastQC (Andrews, 2010). Clonal reads were then removed from the dataset using the clone_filter program distributed with Stacks version 2 (Rochette et al., 2019). Paired-end reads were then demultiplexed using process_radtags after re-orienting reads such that sample-specific indices were located at the beginning of the first read (see Smith et al., 2020 for script). Low quality bases were trimmed from the ends of reads using Trimmomatic v0.32 (Bolger et al., 2014). Specifically, reads were trimmed whenever the mean base quality across a sliding window of 4 bases dropped below Q15. Trimmomatic was also used to remove sequencing adapter contamination from reads. Next, reads were mapped to the Lake Trout genome (Smith et al., 2021 - in press; GCF_016432855.1) using bwa mem (Li, 2013) using standard settings, and resulting BAM files were sorted and indexed using Samtools (Li et al., 2013). Secondary and supplementary alignments and reads with mapping qualitied less than 10 were removed from the dataset using samtools view.

Genotype likelihoods were calculated with ANGSD (Korneliussen et al., 2014) using the GATK model (-GL 2, -doMajorMinor 1, -doMaf 2, minMaf 0.05). We required a minimum mapping and base quality of 10 to use a read or base for calculating genotype likelihoods and only retained high confidence variable sites (SNP_pval 1e-6) for subsequent analysis. We also required that greater than 50% of individuals have at least 1 read at a site in order

to report genotype likelihoods. Genotype likelihoods were exported to Beagle format (Browning & Browning 2016) for further analyses. We also produced a high confidence set of genotype calls using gStacks2 (Rochette et al., 2019), which was converted to VCF format (Danecek et al., 2011) using the populations module in Stacks. Gstacks genotypes were called using default settings and loci were removed from the initial call set if the minor allele was observed in fewer than 2 individuals or if genotypes were called with a genotype quality (GQ) less than 20 in more than 50% of individuals. Additional filtering applied to gStacks genotypes is described in the following sections and all filtering was conducted using vcftools (Danecek et al., 2011).

POPULATION STRUCTURE AND DIVERSITY

We initially assessed population structure using the individual based clustering algorithm implemented in NGSadmix (Skotte et al., 2013) using genotype likelihoods from ANGSD. NGSAdmix was run 10 times for all K values between 1 and 8 with random seeds for each iteration. We required a minimum minor allele frequency of 0.05 and only used sites for which data were available for 100 samples. Next, output files from NGSadmix were converted to log probability tables, which were used to identify the most likely K using the Delta K method (Evanno et al., 2005). Individual ancestry coefficients for each individual were averaged across runs for the most likely K. A principal components analysis (PCA) was also performed using PCAngsd (Meisner & Albrechtsen, 2018) with default settings using genotype likelihoods from ANGSD.

Additionally, using genotypes called with gStacks, we preformed PCA using all loci (complete set), loci within detected islands of divergence (adaptive set; see below for a description of how islands of divergence were delineated), and all loci outside regions with

signals of divergence (neutral set; similar to Berg et al., 2017). These datasets were filtered to exclude genotypes called with fewer than 5 reads in addition to genotypes with GQ less than 20. We then excluded individuals with greater than 50% missing data, loci with minor allele frequencies less than 0.05, and loci with genotypes called for fewer than 70 percent of individuals. Loci located on unplaced scaffolds and the mitochondrial genome were also excluded from this set. Missing genotypes were imputed using Beagle (v5.2; Browning & Browning, 2016) and the dataset was thinned on an increment of 1000 base pairs in order to exclude loci originating from the same restriction cut-site. Discriminant analysis of principle components analysis (DAPC; Jombart et al., 2010) was also conducted using the R package adegenet (Jombart, 2008; R Core Team, 2017) for each of these datasets using 10 principal components and 2 linear discriminant functions. These analyses were conducted for the purpose of verifying results from PCAngsd and determining whether observed patterns of population structure were biased by the inclusion of adaptive loci.

F_{ST} was also calculated between all sample group pairs using the R package SeqVarTools (Gogarten et al. 2014) after removing sample collections with fewer than 8 samples. This was repeated for the complete locus set, the adaptive set, and the neutral set (Supplemental Material 3.6). Additionally, we calculated mean observed heterozygosity (HO, mean expected heterozygosity (HE), and mean FIS using the R package heirfstat (Goudet 2005). The total number of loci with minor allele frequencies (MAF) > 0 and MAF > 0.05 were also reported. We also conducted the exact test for deviations from Hardy-Weinberg Proportions as implemented in the --hardy function from vcftools (Wigginton et al. 2005). Tests for heterozygote excess and deficit were considered to be significant at a

Bonferroni corrected p-threshold of 0.05. The total count of significant tests is reported for each sample group.

Finally, we tested for evidence for deviation from panmixia using analysis of molecular variance (AMOVA) with samples stratified by sampling location and ecomorphotype. AMOVA was conducted using the R package poppr (Kamvar et al. 2014). For this analysis, we thinned the neutral gStacks dataset described above on an increment of 1 Mb prior to analysis. We tested for significant deviations from panmixia associated with these two sampling strata (ecomorphotype and sampling location) using the randomization test described in Excoffier et al. (1992). Significance was assessed at an alpha threshold of 0.05.

OUTLIER ANALYSIS AND INVERSION DETECTION

We calculated the population branch statistic (PBS; Yi et al., 2010) for all morphotypes using ANGSD, which describes the change in allele frequency in a population after divergence from the most recent common ancestor. Site allele frequency likelihoods were calculated using the settings -gl 2 -minMapQ 10 -minQ 10 -minInd 10 with the base in the Lake Trout genome assumed to be ancestral. Next, 2-dimensional site frequency spectra (2D-SFS) were calculated for all ecomorphotype pairs using realSFS. All 2D-SFS and site allele frequency likelihoods were used as priors for PBS calculations using realSFS. PBS was then reported in 100 KB sliding windows with 10 Kb offsets as well as 5 Mb windows with 10 Kb offsets using realSFS. These values were Z-transformed and visualized using ggplot2 (Wickham & Wickham, 2007) and circlize (Gu et al. 2014). One-hundred Kb windows with Z-transformed PBS values greater than 5 were considered to be potential targets of selection. PBS values in 5 Mb windows were also plotted using the R-package

circlize in order to visualize the locations of especially large islands of divergence. The extended PCAdapt model implemented in PCAngsd was also used to identify putatively adaptive loci using standard settings (Meisner et al., 2021). Scores from the PCAdapt test were converted to p-values using the script pcadapt.R (https://github.com/rosemeis/pcangsd) as in Meisner et al. (2021). SNPs were considered significant after Bonferroni correction at an alpha threshold of 0.05.

We used the auxiliary covariate model implemented in Baypass (Gautier 2015) and the C2 contrast statistic (Olazcuaga et al., 2020; also implemented in Baypass) to identify chromosomal regions associated with ecomorphotype. Allele frequency estimates for each sample collection were exported using ANGSD and converted to Baypass input format. Ecomorphotype was coded as a binary variable (-1 = absence, 1 = presence) for each sample collection. SNPs with Bayes Factors greater than 10 were considered to be associated with a given ecomorphotype in accordance with Jefferys' Rule (Jefferys, 1998). Additionally, SNPs with C2 contrast test p-values less than a Bonferroni corrected alpha of 0.05 were considered significant.

Islands of divergence were defined by dividing the genome into 1 Mb sliding windows offset by 100 Kb, identifying all windows with more than 2 SNPs with significant p-values from the C2 contrast test, and merging overlapping windows using bedtools merge (Quinlan and Hall, 2010). This process was repeated using C2 contrast test results for each ecomorphotype. Within each region, we defined the position of strongest association (POSA) as the position of the SNP with the largest -log10 p-value from each respective C2 contrast test. If multiple SNPs within a window had identical p-values that were also the highest in the region (i.e. multiple infinite -log10 p-values) then we

considered the POSA to be the mean of the positions for these SNPs. The total number of associated SNPs, density of associated SNPs, proportion of associated SNPs, and size of each region were calculated for each region for each ecomorphotype (see Supplemental Material 3.2). Bedtools closest (Quinlan and Hall, 2010) was used to identify the coding sequence (CDS) closest to each POSA and gene descriptions were extracted for each CDS from the feature table available with annotation release 100 for the RefSeq version of the SaNama_v1.0 genome assembly. The C2 contrast test results were chosen for defining associated regions because this proved to be the most conservative test for adaptive differences between ecomorphotypes.

We used the R package inveRsion (Cáceres et al. 2012) to identify putative chromosomal inversions. For this analysis we made use of genotypes called with gStacks that were filtered to exclude individuals and loci with greater than 10% missing data after requiring a genotype quality of 20 and more than 5 reads to call a genotype. This dataset was also filtered to exclude SNPs with minor allele frequencies less than 0.1. The resulting file was imputed using Beagle v5.2 before being used as input for inveRsion using the arguments blockSize=3, minAllele=0.1, maxSteps = 100, and 4 Mb windows. Overlapping regions of interest for which the Bayesian Information Criterion (BIC) was greater than 0 were merged using the function InvClust and putative inversions with a maximum BIC greater than 100 were retained for additional analysis. Coordinates for putative inversions were cross referenced with the locations of islands of divergence described above by comparing chromosomal positions.

GENE SET ENRICHMENT AND CANDIDATE GENES

Functional annotations were obtained for all coding sequences (CDS) in the Lake Trout genome using the Panzzer 2 public annotation server (Törönen et al. 2018). Translated coding sequences from Lake Trout Annotation Release 100 were accessed via the RefSeq FTP site for assembly SaNama_1.0 on April 14th, 2021 (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/016/432/855/GCF_016432855.1_SaNama_1.0/). We required a minimum query and subject coverage of 0.6, a minimum alignment length of 100, and sequence identity between 0.4 and 1.0 to retain homologous sequences. We output one DE per query sequence with a form factor cut-off of 0.2. Gene ontology (GO) term prediction was done using the Argot scoring function and we removed redundant GO terms using a Blast2GO threshold of 55. GO terms were output for all coding sequences located on assembled chromosomes that were associated with annotated proteins. This set of GO terms were used as the baseline for gene set enrichment analyses.

We specifically focused on CDS in close proximity to POSA identified within regions containing more than two significant SNPs because these associations are less likely to be produced by false positives. For each POSA, we identified the closest CDS, then subset this list to exclude CDS further than 100 Kb from the focal position. Gene set enrichment analysis (GSEA) was performed using the r-package topGO (Alexa & Rahnenführer, 2009) using the weight01 algorithm and with significance assessed using p-values from a Fischer's exact test. Significantly enriched GO terms associated with biological processes were identified for each ecomorphotype using a p-threshold of 0.05. We also extracted all CDS within 100 Kb of POSA and used these to generate a full list of candidate genes (see Supplementary Material 3.2).

RESULTS

BIOINFORMATICS

We obtained between 99,348 and 12,145,470 mapped reads for individuals with a mean mapped read count of 3,273,380. A total of 222 individuals and 225,700 SNPs distributed across Lake Trout chromosomes and unplaced scaffolds were retained for low-coverage analyses (NGSAdmix, PCAngsd, PCAdapt, Baypass, and PBS Scan). A total of 9,785 SNPs and 144 individuals were retained for inversion detection with inveRsion. A total of 181 individuals and 18,095 SNPs were retained in the gStacks dataset used for estimating diversity statistics and evaluating neutral and adaptive population genetic structure using DAPC, PCA and pairwise global FST. Of these SNPs, 12,037 were retained for analysis after imputation and thinning. Of these SNPs, 1,647 were retained in the smaller dataset used for AMOVA.

POPULATION STRUCTURE AND DIVERSITY

Principle components analysis conducted with PCAngsd suggested the existence of three groups corresponding to leans, humpers, and siscowets (Figure 3.3). The first principal component (PC1) separated humpers, leans, and siscowets and explained 2.89% of variance. The second axis (PC2) primarily separated leans from humpers and siscowets and explained 2.57% of variance. Multiple individuals exhibited scores that were intermediate between leans and humpers or siscowets and humpers; however, we did not identify individuals that were clearly intermediate between the leans and siscowet clusters. This suggests a relative lack hybridization between leans and siscowets (Ma & Amos, 2012). Results from PCA and DAPC using the complete, neutral, and adaptive marker sets produced nearly identical patterns of population structure (Supplemental Material 3.5).

The null hypothesis of panmixia between ecomorphotypes was rejected with a p-value of 0.001 for the AMOVA conducted along this stratum (Table 3.2). We estimated that 5.07% of variance is partitioned between ecomorphotypes, 3.65% exists within ecomorphotypes, and 91.28% exists within samples. We failed to identify any significant evidence of population structure associated with sampling location (p=0.736). The high proportion of variance within samples relative to within ecomorphotypes suggests that substantial gene flow occurs between ecomorphotypes.

Results from NGSadmix further support this conclusion (Figure 3.1D). We found that deltaK was maximized at K=3. Again, the clusters identified primarily correspond to fish identified as leans, humpers, and siscowets. Individual ancestry coefficients suggest that hybridization between leans and humpers and humpers and siscowets is common, but hybridization between leans and siscowets is rare (similar to PCA). One fish that was identified as a siscowet appears to have been a lean-humper hybrid and multiple fish identified as siscowets appear to have primarily humper ancestry. Five fish identified as humpers had primarily siscowet ancestry and one of these fish appears to be a lean-humper hybrid. We predict that some of the inconsistencies between individual ancestry coefficients and field identifications are due to misidentifications (see Muir et al. 2014).

As expected under the hypothesis that hybridization rates are increasing over time, we found that levels of humper ancestry increased within the Apostle Islands lean population between the 1960s and 1990s (Table 3.1). The most extreme change occurred between the 1980s and 1990s sampling periods for which the proportion of humper ancestry within the lean populations shifted from 0.5% to 18.9%. The proportion of siscowet ancestry varied between 0.2% and 1.4% across this same time period. High levels

of humper ancestry were also observed within fish identified as leans from Stannard Rocks and Isle Royale (43.0% and 29.1%, respectively) collected in the 1990s. The results from PCAngsd also suggest that the most divergent leans are from the earliest collections available (1960s and 1970s from the Apostle Islands). Genotype frequencies in lean collections from the Apostle Islands and Isle Royale match Hardy-Weinberg expectations, suggesting that these populations represent randomly mating hybrid populations. However, it should be noted that deviations from HWE were primarily detected in collections with the highest sample sizes. A significant excess of heterozygotes relative the HWE was detected at 45 loci in the Stannard Rock lean collection.

Similarly, PCAngsd results indicate that the most divergent siscowets were from the earliest collections. The siscowet ancestry proportions for the 1970s and 1980s siscowet collections from the Apostle Islands were 99.6 and 96.1%, respectively. The proportion of siscowet ancestry in collections from the 1990s (Stannard Rock, Isle Royale, and Whitefish Point) varied from 54.5 to 89.3%, with the majority of hybrid ancestry originating from the ancestral humper population (8% to 43.8%). The proportion of lean ancestry varied between 1.2% and 7.7% across siscowet collections, with the Isle Royale siscowet collection having the highest proportion of lean ancestry. Deviations from HWP were only detected in the siscowet collection from Isle Royale.

Similarly, the 1990s humper collection from Isle Royale also has the highest observed proportion of lean ancestry (11.5%). Levels of siscowet ancestry within the Isle Royale and Caribou Reef humper collections were similar (7.7 and 6.7%, respectively), despite differences in the level of lean ancestry (11.5% vs. 2%), suggesting that lean-humper hybridization might be more common near Isle Royale. Deviations from HWP,

primarily due to an excess of heterozygotes, were detected in the Isle Royale and Caribou Reef humper collections.

Considering all analyses, we conclude that ecomorphotypes represented sub-populations of a larger metapopulation during the 1990s and previous decades. Additionally, admixture between these subpopulations primarily involved hybridization between siscowets and humpers or leans and humpers. Additionally, levels of hybridization appear to have increased across time at the Apostle Islands sampling location. Furthermore, levels of admixture between ecomorphotypes varied across sampling locations.

OUTLIER ANALYSIS AND INVERSION DETECTION

Of the 225,700 SNPs tested for associations with ecomorphotype, 601 exhibited significant test results based on the C2 contrast test, Baypass auxiliary covariate model, and PCAdapt and were also located within a 100 Kb PBS window for which the Z-transformed value was greater than 3 (Figure 3.4; Supplemental Material 3.1). The C2 contrast test was the most conservative test, with 92.3% of the 601 detected SNPs yielding significant results based on 1 or more of the other tests and 39.8% of SNPs being confirmed by all other tests (Figure 3.5). Results from this test were used for delineating islands of divergence; however, results from the PBS scan with 5 Mb windows should also be noted. The PBS scan conducted using 5 Mb windows identified 6 large regions associated with siscowets on chromosomes Sna12, Sna13 Sna19, Sna25, Sna32, and Sna38. The same analysis identified 3 large regions associated with the humper ecomorphotype on chromosomes Sna12, Sna15, and Sna19. The PBS scan with 5 Mb windows also identified 6 regions associated

with the lean ecomorphotype on chromosomes Sna1, Sna9, Sna13, Sna16, Sna19, and Sna35 (Supplemental Material 3.4).

We identified 9 regions associated with the humper ecomorphotype, 50 regions associated with the lean ecomorphotype, and 74 regions associated with the siscowet ecomorphotype based on C2 contrast test results (Supplemental Material 3.2). Of these, 10 siscowet and lean associated regions contained SNPs with infinite p-values. When these regions were merged, we were left with 94 genomic regions associated with adaptive differences between ecomorphotypes. Ecomorphotype associated regions ranged in size from 1 to 6.4 Mb and covered 1.0%, 5.8%, and 7.8% of the genome for humpers, leans, and siscowets, respectively. Between 0 and 24.15% of each chromosome was covered ecomorphotype associated regions. The number of significant SNPs per Mb for these regions ranged from 1.58 to 29.03 with between 3 (the minimum threshold) and 90 significant SNPs in each region. We identified 42 coding sequences within 100 Kb of the POSA for humper-associated regions. These include CDS for FLVCR1-like, Gamma M2 -like, VASH2-like, tubby protein-like, LATS2-like, and interleukin-17D among others. The second strongest association with the humper ecomorphotype came from a group of SNPs within tubby protein-like, which is associated with multiple developmental processes including eye development. We identified 238 CDS within 100 Kb of POSA for lean associated regions. These include Wnt-5a, SOX6-like, RNF121, GDF8-like, cadherin-4-like, and vgll3 among others. Vgll3 has been found to be associated with sea-age at maturity in Atlantic Salmon (Barton et al. 2016). In this case, the POSA was 72.9 Kb away from vgll3 and 5 other candidate genes were in closer proximity. It is therefore unclear if vgll3 is the true target of

selection in this case. We identified 326 CDS near POSA for siscowet associated region. These include CDS for frizzled-1-like, DVL-2-like, ABHD8, and PARP14.

InveRsion identified 10 putative inversions (Table 3.3). The approximate coordinates (minimum start and maximum end) for 6 of these inversions overlapped detected islands of divergence. The most striking example comes from a putative inversion located between 68.12 and 83.05 Mb on chromosome Sna1. This inversion overlapped 9 initially detected islands of divergence and contains SNPs with strong significant associations with all three ecomorphotypes. These include a 6.4 Mb lean associated region containing 70 significant SNPs (67-73.4 Mb), some of which yielded infinite p-values. Another 2.6 Mb (75.7-78.3 Mb) lean associated region was also detected within the putative inversion region. This island of divergence contained SNPs with infinite p-values within the coding sequence for the transcription factor sox6-like. A full record of ecomorphotype associated regions and candidate genes is available in Supplemental Material 3.2.

GENE SET ENRICHMENT AND CANDIDATE GENES

Panzzer2 obtained functional annotations for 50,653 protein associated coding sequences located on Lake Trout chromosomes. We identified 72 GO terms that were over-represented in our set of lean-associated candidate genes. These include GO terms for regulation of eye photoreceptor cell development, negative regulation of non-canonical Wnt signaling, retina morphogenesis in camera-type eye, locomotor rhythm, adipose tissue development, blood vessel development, cellular response to decreased oxygen, articular cartilage development, and positive regulation of fast-twitch skeletal muscle fiber contraction, among others.

We identified 86 GO terms that were significantly over-represented in our set of siscowet-associated candidate genes. These include GO terms for entrainment of circadian clock by photoperiod, positive regulation of bone mineralization, negative regulation of gluconeogenesis, ureteric peristalsis, negative regulation of triglyceride biosynthetic process, axial mesoderm development, negative regulation of female gonad development, muscle tissue morphogenesis, and positive regulation of male gonad development.

We identified 26 GO terms that were significantly over-represented in our set of humper-associated candidate genes. These include GO terms for cellular response to hypoxia, regulation of organ growth, response to vitamin D, photoreceptor cell maintenance, and negative regulation of canonical Wnt signaling (see Supplemental Material 3.3).

DISCUSSION

The primary outcome of this study was the identification of multiple genomic regions associated with the adaptive diversification of Lake Trout ecomorphotypes in Lake Superior. Our ecomorphotype association analyses and gene set enrichment analyses demonstrate that Lake Trout ecomorphotypes are polygenic (as suggested by Perrault-Payette et al., 2017) and a multitude of biological processes are under differential selection between forms.

Additionally, we found that multiple genomic islands of divergence are associated with putative inversions. Future work should seek to experimentally validate the ten inversions identified here (Table 3.3) using long-read sequencing, linkage mapping, or chromatin conformation capture (Merot 2020). One of the most striking examples comes from chromosome Sna1. We detected 9 significant peaks of association with lean, siscowet,

and humper ecomorphotypes on this chromosome. All detected peaks were within the region spanning 58.9 and 83.6 MB and candidate genes included dennd2b, tubby-like, ataxin-1 like, ras2, hsd17b12b, and sox-6-like. This region of Sna1 was also associated with body size and condition factor in a previous QTL mapping study (Smith et al. 2020). Within this region, C2 statistic p-values were lowest for lean and humper C2 contrast tests. Of the 9 peaks identified, one was associated with humpers, five were associated with leans, and three were associated with siscowets. The gene in closest proximity to the peak within the humper associated region was tubby-like; a member of the tubby gene family which is associated with insulin signaling and metabolism in adipose tissue, retina development, and autosomal recessive retinitis pigmentosa (Mukhopadhyay & Jackson, 2011). Interestingly, this chromosome arm exhibits a pattern of delayed re-diploidization in Lake Trout (Smith et al. 2021).

The study conducted by Goetz et al. (2010) identified multiple genes associated with immune response that were differentially expressed between leans and siscowets in addition to multiple genes associated with lipid metabolism. C1q proteins exhibited some of the highest levels of differential expression between the two forms in this study. They proposed that these differences might be due to variable pathogen susceptibility associated with higher pathogen loads in warmer, shallow water, habitats occupied by leans. Multiple studies exploring genetic differentiation between sockeye salmon (Oncorhynchus nerka) ecotypes have also identified immune response associated genes as targets of divergent selection (Gomez-Uchida et al. 2011; McGlauflin et al. 2011, Larson et al. 2014). We identified a 3.3 MB region of chromosome Sna13 for which the POSA (Sna:13:29325600) was nearest to the gene C1q-like protein 2 for lean Lake Trout. This region overlapped a

siscowet associated region; however, the POSA (Sna:13:29222356) for siscowet was different and closer to an uncharacterized C3orf38 homolog. Other genes near the POSA included insig1, which is associated with the regulation of lipogenesis (Li et al. 2003). Numerous other immune response-associated candidate genes were identified in close proximity to POSA (i.e., < 50kb). These include interferon a3-like, vitamin D3 receptor A, complement C3-like, interferon regulatory factor 9-like, interleukin-12 subunit alpha, N-myc-interactor, complement factor B-like and Nkx-2-5, among others. Our study adds to a growing body of knowledge suggesting that ecotypic divergence might often be accompanied by selection on genes related to immune response.

The strongest signal of association on Sna16 came from a 6.2 MB lean-associated region for which the POSA overlapped the coding sequence for Wnt5a. The POSA contained SNPs with infinite log10 p-values for the lean C2 statistic association test. Wnt5a is one of the best studied non-canonical Wnt proteins (Kikuchi et al. 2012), and is associated with a diversity of processes related to embryonic and morphological development. These including axis elongation and adipogenesis (Kikuchi et al. 2012). The region in question contains 70 lean-associated SNPs, with 10.89% of all SNPs within the region being associated with the lean ecomorphotype. Wnt signaling involves binding of a Wnt protein by a Frizzled family receptor and the transduction of a biological signal into the cytoplasm via Dishevelled (Wodarz & Nusse 1998). This signal transduction pathway is highly conserved (Rothbäche et al., 1995) and facilitates a diversity of developmental and homeostatic processes. Interestingly, the strongest signal of association on Sna32 (also an infinite log10 p-value) came from a 3.2 MB siscowet-associated region for which the POSA was in closest proximity to a Frizzled 1-like protein. Additionally, another POSA located on

Sna36 was in close proximity to a Dishevelled homolog (DVL-2-like, 41.7 Kb). Additionally, GO terms related to positive or negative regulation of canonical and non-canonical Wnt signaling were found to be over-represented by all gene set enrichment analyses. For Siscowet, we identified a total of 4 candidate genes associated with Wnt signaling within 100 Kb of POSA (Frizzled-1-like, DVL-2-like, IFT80, and hdac1-B). For leans we identified 5 candidate genes associated with Wnt signaling within 100 Kb of POSA (Wnt-5a, DVL-2-like, IFT80, VANGL2, and NKX-2-5). Two Wnt signaling associated candidate genes were identified for humpers (NKX-2-5 and LATS2-like). Given the infinite p-values for Wnt-5a and Frizzled-1-like and the statistically significant over-representation of Wnt signaling associated GO terms, we conclude that adaptive differences between Lake Trout ecomorphotypes are partially associated with variation in proteins involved with Wnt signaling.

Furthermore, our results indicate that the breakdown of reproductive isolation between ecomorphotypes documented by Baillie et al. (2016) is likely associated with admixture between leans and humpers and siscowets and humpers. Our analysis of population genetic structure suggests that lean and siscowet ecomorphotypes hybridize with humpers; however, hybridization between siscowet and lean Lake Trout appears to be comparatively rare in the samples examined. Additionally, admixture coefficients calculated for a limited number of historical samples from Apostle Islands leans suggest a rapid increase in the amount of hybridization with humpers between the 1980s and 1990s. This suggests that decreases in genetic distance (Baillie et al. 2016) and phenotypic distinctiveness (Muir et al. 2014) are likely associated with hybridization between humpers and leans or humpers and siscowets. Admixture coefficients for historical samples

were generated using low-coverage data for greater than 200,000 SNPs and the methods employed are able to account for variation in data quality across samples (Skotte et al., 2013); however, these results should be validated using larger sample sizes and genotyping methodologies that are more appropriate for historical specimens (Rowe et al. 2011; i.e., targeted sequence capture, whole genome sequencing, or PCR based enrichment strategies).

Researchers have hypothesized that the humper ecomorphotype emerged as the result of introgression between lean and siscowet forms during periods of variable water level following the last Pleistocene glacial retreat (Burnham-Curtis, 1993). If this is the case, then it makes sense that reproductive isolation mechanisms would be weaker between leans and humpers or siscowets and humpers than between leans and siscowets, and this might explain observed patterns of hybridization. Future work should seek to identify loci associated with pre- and post-mating isolation between Lake Trout ecomorphotypes. Loci associated with pre-mating isolation could be identified using detailed information on spawning habitat use and spawning time at the individual level. Loci associated with post-mating reproductive isolation could potentially be identified by testing for evidence of viability selection (Cotto & Servedio, 2017) using genotypic ratios in families created using multiple ecomorphotypes (Luo & Xu, 2003).

The active reintroduction of ecomorphotype associated genetic variation could help facilitate the restoration of phenotypically diverse Lake Trout forms capable of utilizing the full diversity of available habitats in the Great Lakes (Edsall & Kennedy, 1995; Ebener, 1998; Krueger, Jones, & Taylor 1995). Lean Lake Trout are only able to utilize a small proportion of habitats and niches that Lake Trout historically occupied in the Great Lakes

(Edsall & Kennedy 1995). Additionally, lean Lake Trout hatchery strains derived from Lake Superior encompass only a small fraction of the total genetic variation that exists within the larger metapopulation (Page, 2001), and this has been identified as an impediment to restoration of wild populations (Bronte et al. 2003). The Klondike Reef strain, founded by humper Lake Trout from Lake Superior, currently represents the only non-lean strain stocked in the Great Lakes. This strain has exceptionally low eye-up rates in the hatchery (3-33%) and the process of domestication appears to have resulted in a significant reduction in genetic diversity (Salvesen, 2015). The extent to which adaptive variation and humper-associated alleles were maintained during the process of domestication is unclear. Additionally, the results presented here suggest that some level of introgression likely occurred in the wild prior to the founding of the Klondike Reef broodstock. The Klondike Reef strain has been stocked in Lakes Michigan (Larson et al., 2021) and Erie (Muir et al. 2012) and proposals have been made to reintroduce siscowet and humper Lake Trout to lakes where they were extirpated (Muir et al. 2012).

Knowledge of the genetic variation underlying ecomorphological variation could aid in the restoration of Lake Trout populations in the Great Lakes because captive bred populations could be screened for lean, siscowet, or humper associated alleles in order to ensure that variation associated with ecomorphological variation is not lost during domestication and reintroduction. This could be done concurrently with a genotyping program aimed at monitoring levels of inbreeding and genome-wide variation in hatchery populations (Flanagan et al., 2018). Given that lean Lake Trout are only able to utilize a small proportion of habitats and niches that Lake Trout historically occupied in the Great Lakes (Edsall & Kennedy 1995), the continued absence of siscowet and humper forms

outside of Lake Superior represents a direct impediment to the long-term goal of restoring the deep-water food web in the Great Lakes. The reintroduction of siscowet and humper Lake Trout would likely facilitate the exploitation of a greater diversity of niches; however, further research on the processes driving hybridization between forms is likely necessary.

APPENDIX

Figure 3.1: Panels A, B, and C display artists representations of humper, lean, and siscowet Lake Trout ecomorphotypes. Panel D displays NGSadmix results for the most likely K (K=3) based on the delta K method. Results from 10 runs were averaged to calculate individual ancestry proportions. The blue, red, and orange vertical bars correspond to individuals with humper, lean, and siscowet ancestry, respectively. Samples are labeled as leans, siscowets, and humpers according to field identifications from fisheries management professionals. See Table 3.1 for a breakdown of lean, siscowet, and humper ancestry proportions by sampling location and time period. Images are from Muir et al., (2021).

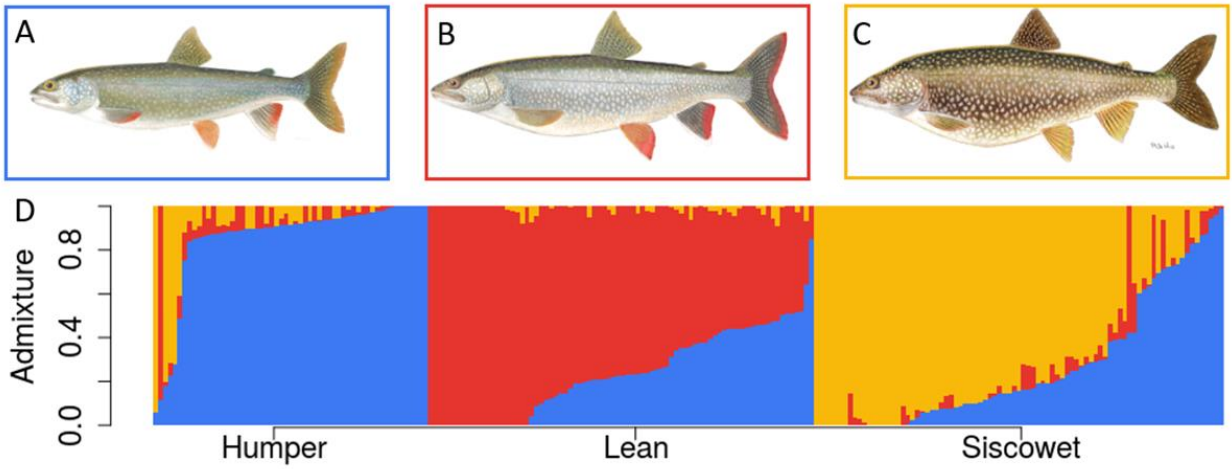

Figure 3.2: Map of the Great Lakes region with sampling locations labeled within Lake Superior. A list of the different ecomorphotypes that were sampled at each location is displayed below the label for each sampling location.

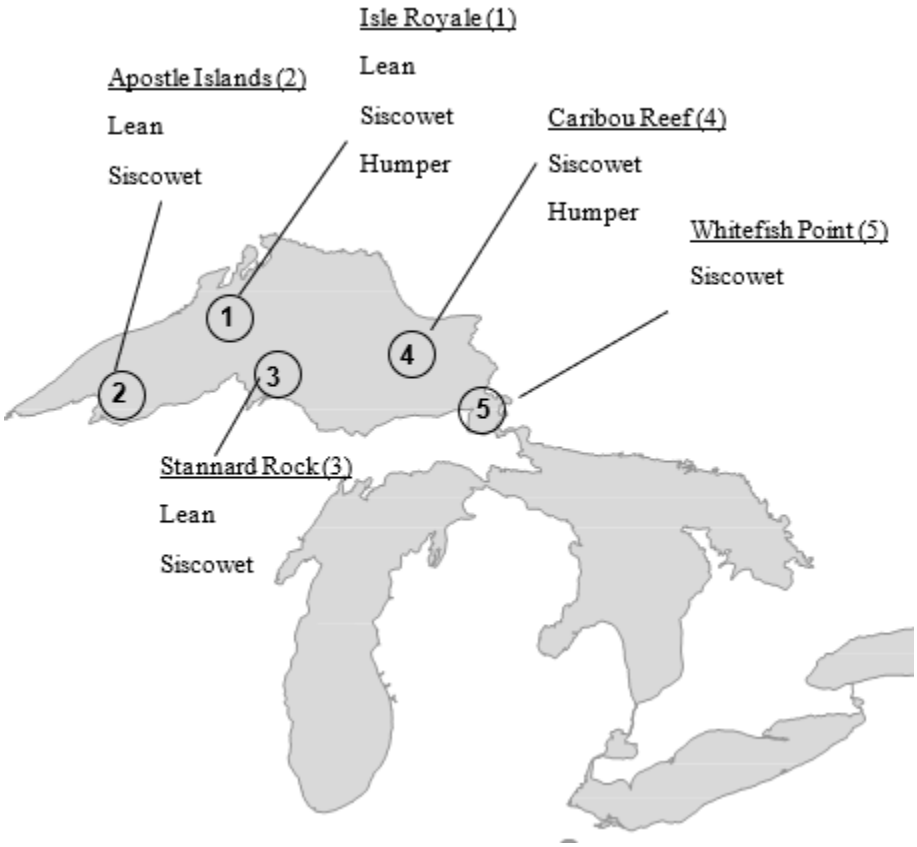

Figure 3.3: Population genetic structure in Lake Trout from Lake Superior in the 1990's and earlier evaluated using principal components analysis (PCA) conducted using PCAngsd and the full dataset (225,700 SNPs). Humpers, leans, and siscowets are identified using blue, red, and orange points respectively. Sampling locations are delineated with separate symbols as described in the legend. The first axis (y-axis here; Principal Component 1) describes 2.89% of variation and separates all three ecomorphotypes. The second axis (x-axis here; Principal Component 2) describes 2.57% of variation and primarily separates leans from the other two ecomorphotypes. Similar patterns were observed regardless of whether or not putative adaptive regions were excluded.

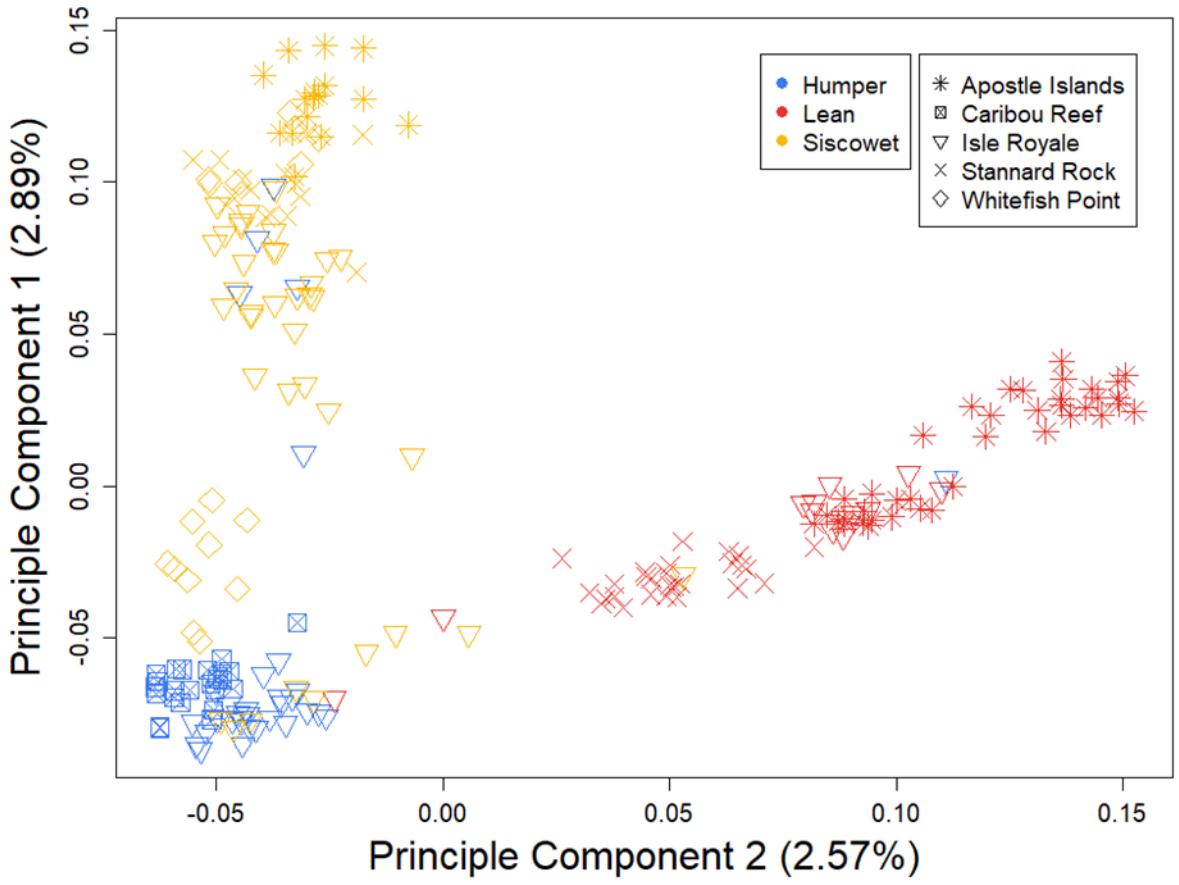

Figure 3.4: Venn diagram displaying overlap among the sets of ecomorphotype associated loci identified by the extended PCAdapt model implemented in PCAngsd, the C2 contrast test, Bayes Factors output by Baypass,, and SNPs in windows where the Z-transformed population branch statistic for the window was greater than 3 for one or more ecomorphotype. Of the SNPs detected by at least one test, 11.26% (601) were detected using every test.

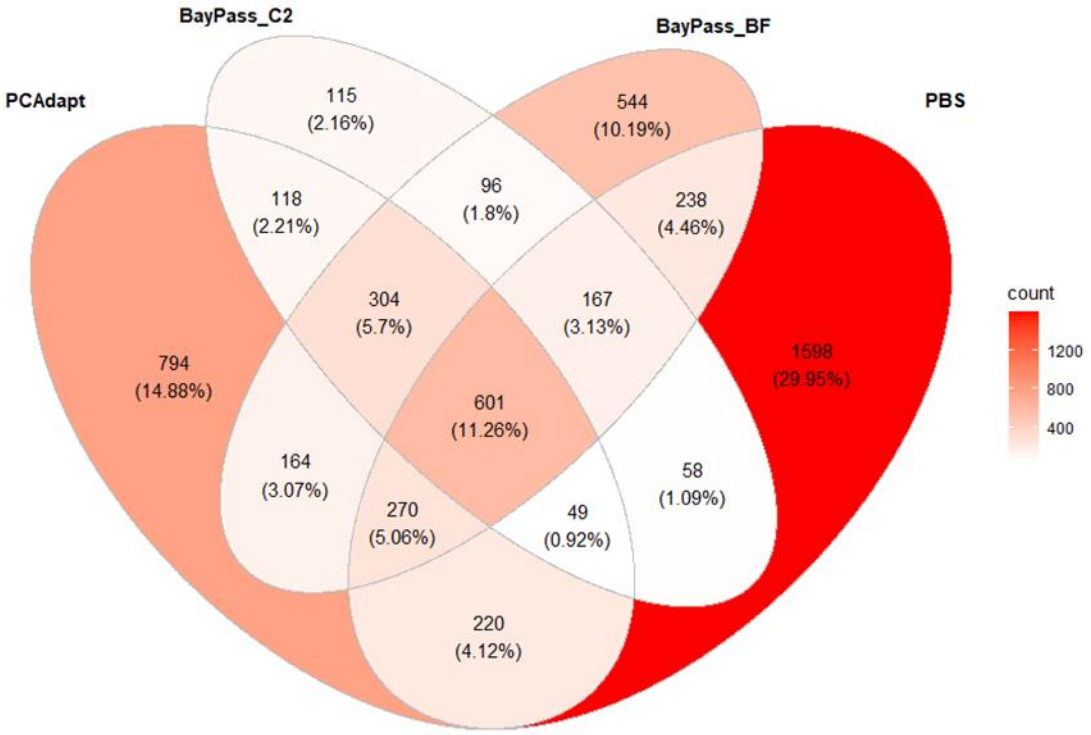

Figure 3.5: Circos plot displaying results from C2 statistic outlier scans conducted with Baypass for siscowet (outer ring, ring 1), humper (ring 2), and lean (inner ring, ring 3) associated loci. Each point represents a -log10 p-value for a single SNP. Each panel corresponds to a Lake Trout chromosome (Sna1-Sna42) and the x-axis corresponds to the physical position of each SNP on a respective chromosome. Coordinates are delineated as mega bases on the outside ring. SNPs with significant p-values at a Bonferroni corrected alpha of 0.05 are highlighted in orange for siscowet, blue for humpers, and red for leans.

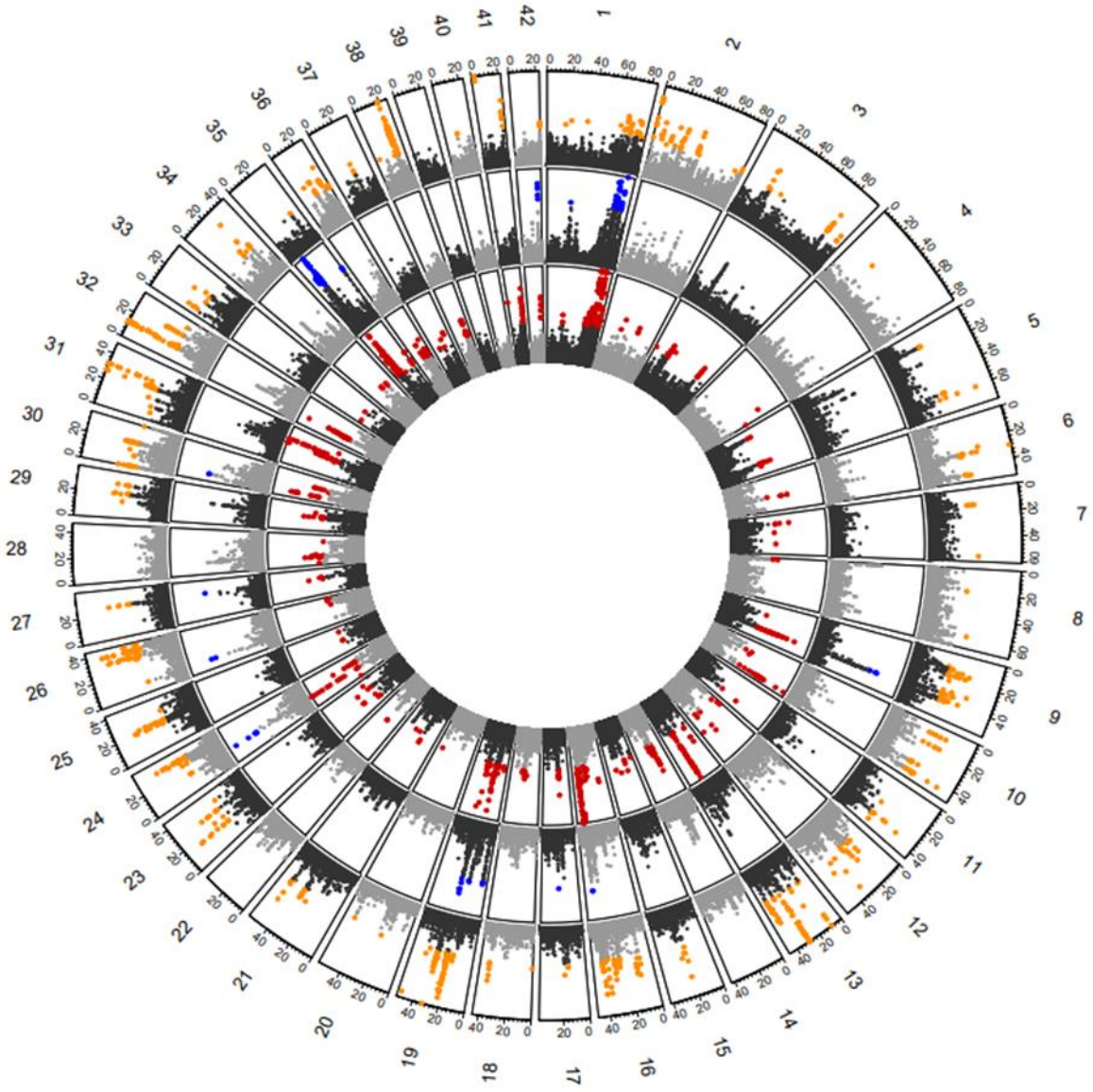

Table 3.1: Diversity and admixture proportions across collections. The table below lists sample sizes for the complete dataset (N-All) and the dataset after removing individuals with high levels of missing data (N-Sub). The number of polymorphic sites (MAF>0.05) for each group, mean FIS, and mean overserved and expected heterozygosities (H_O and H_E) are also listed for each group. These statistics were calculated using the filtered genotype set produced using gStacks. The last three columns list the mean individual ancestry coefficients calculated by NGSadmix for lean, humper, and siscowet clusters for each stratum. This analysis was conducted using the complete dataset processed with ANGSD and values were calculated using all available individuals.

Ecotype	Location	Decade	N-All	N-Sub	MAF > 0.05	F_{IS}	H_O	H_E	Lean Anc.	Sisc. Anc.	Hump. Anc.
Lean			**80**	**62**	**10908**	**0.02**	**0.31**	**0.32**			
	Apostle Islands										
		1960s	7	0	-	-	-		0.995	0.005	0.000
		1970s	7	1	-	-	-		0.988	0.012	0.000
		1980s	8	5	7846	0.33	0.26	0.39	0.981	0.014	0.005
		1990s	17	17	10651	-0.04	0.35	0.33	0.808	0.002	0.189
	Isle Royale										
		1990s	13	13	10173	-0.01	0.35	0.35	0.687	0.022	0.291
	Stannard Rock										
		1990s	28	26	10395	-0.04	0.33	0.32	0.551	0.018	0.430
Humper			**57**	**53**	**10814**	**-0.02**	**0.33**	**0.32**			
	Isle Royale										
		1990s	33	29	10755	-0.02	0.33	0.32	0.115	0.077	0.808
	Caribou Reef										
		1990s	24	24	10416	-0.04	0.35	0.34	0.020	0.067	0.913
Siscowet			**85**	**66**	**10866**	**0.02**	**0.31**	**0.31**			
	Apostle Islands										
		1970s	8	2	-	-	-		0.004	0.996	0.000
		1980s	8	7	9,230	0.18	0.28	0.35	0.030	0.961	0.010
	Stannard Rock										
		1990s	15	13	9776	-0.07	0.37	0.35	0.024	0.893	0.082
	Whitefish Point										
		1990s	17	14	9623	0.00	0.33	0.33	0.012	0.550	0.438
	Isle Royale										
		1990s	37	30	11130	-0.01	0.32	0.32	0.077	0.545	0.378

Table 3.2: Results from Analysis of Molecular Variance (AMOVA) testing for evidence of population structure associated with ecotype and sampling location (Stratum). Sigmas, percent of total variance ascribed to each partition, degrees of freedom, and p-values generated using the randomization test described in Excoffier et al. (1992) are listed.

Stratum	Variation	Sigma	%	DF	p-value
Ecotypes	Variation between ecotypes	14.96	5.07	2	0.001
	Variation within ecotypes	10.76	3.65	6	0.001
	Variation within samples	269.32	91.28	172	0.001
	Total variation	295.04	100.00	180	-
Locations	Variation between locations	-3.01	-1.03	4	0.736
	Variation within locations	24.67	8.48	4	0.001
	Variation within samples	269.32	92.55	172	0.001
	Total variation	290.99	100.00	180	-

Table 3.3: Coordinates for putative inversions detected using inveRsion using a maximum Bayesian Information Criteria (BIC) threshold of 100. Chromosome, minimum start and end coordinates, maximum start and end coordinates, size range, maximum BIC, and frequency estimates are provided. All chromosomal positions are listed in megabases (Mb).

Chromosome	Start (min)	Start (max)	End (min)	End (max)	Max. Size	Min. Size	Max BIC	Frequency
Sna1	68.12	78.99	72.16	83.05	14.93	6.82	140.48	0.45
Sna5	8.33	8.33	12.34	12.34	4.01	4.01	108.38	0.45
Sna5	58.37	68.76	62.72	72.9	14.53	6.04	487.67	0.49
Sna7	48.76	56.37	52.81	60.43	11.66	3.57	220.34	0.47
Sna30	3.28	8.89	7.35	13.16	9.87	1.54	115.54	0.51
Sna34	30.92	30.92	35.86	35.86	4.94	4.94	112.68	0.25
Sna37	24.79	26.55	28.9	30.56	5.76	2.35	262.69	0.52
Sna39	5.35	5.96	9.99	9.99	4.64	4.03	126.07	0.48
Sna40	16.22	18.95	20.56	23.39	7.17	1.61	185.41	0.49
Sna42	11.08	17.27	15.13	22.11	11.03	2.14	158.32	0.58

The following supplementary materials mentioned herein are too large to be usefully displayed here and available upon written request to the author.

Supplemental Material 3.1 – Results from all outlier tests

Supplemental Material 3.2 – Tables listing all candidate genes, POSA, CDS nearest to POSA, and coordinated for islands of divergence

Supplemental Material 3.3 – Significantly enriched GO terms

Supplemental Material 3.4 – PBS Scan in 5MB windows

Supplemental Material 3.5 – Population structure with complete, neutral, and adaptive marker sets

Supplemental Material 3.6 – FST matrices for complete, neutral, and adaptive marker sets

CHAPTER 4: EVOLUTION AFTER REINTRODUCTION – THE GENETIC BASIS FOR VARIATION IN FITNESS AMONG SOURCE POPULATIONS IN RECOVERING POPULATIONS OF AN AQUATIC TOP PREDATOR

ABSTRACT

Species extirpation events are often followed by large-scale reintroduction efforts that make use of individuals derived from *in situ* or *ex situ* source populations. These efforts are often unsuccessful. Source populations frequently have variable fitness in introduced environments and the biological processes that explain this variation in survival and reproductive success are rarely dissected successfully. We used a combination of F_{ST} outlier tests and local ancestry inference to identify genomic regions associated with inter-strain variation in fitness in the recovering Lake Trout (*Salvelinus namaycush*) population in Lake Huron. We identified 97 high-confidence F2 hybrids that were spawned by wild parents between 1998 and 2014. Local excesses of Seneca Lake origin ancestry along the haplotypes of these individuals indicated that elevated fitness of the Seneca Lake hatchery strain can likely be attributed to adaptive differences in 7 genomic regions. We also identified a single region on chromosome Sna35 in which selection favors Great Lakes origin haplotypes, suggesting that this locus is associated with local adaptation to Great Lakes environments. Further analysis indicated that these signals of selection were dependent on the genetic background of hybrid individuals. Specifically, Seneca strain ancestry was only significantly favored on chromosome 11 within F2 hybrids with Seneca Lake and Lake Michigan ancestry. We detected two separate genomic regions associated with increased Seneca strain fitness within F2 individuals for which the Great Lakes ancestry component was derived from Lake Superior origin strains. These regions were located on chromosomes Sna8 and Sna19. Within this group of individuals, a region of

chromosome Sna1 and a region of Sna35 both exhibited an excess of Great Lakes origin haplotypes, again suggesting that Great Lakes strains carry adaptive alleles at some loci. As expected, these admixture outlier regions overlapped with SNPs with significant signals of adaptive differentiation between hatchery strains. This collection of outlier SNPs was enriched for genes associated with swimming behavior and negative regulation of vascular wound healing, which suggests that variation in fitness between strains might be due to differences in their ability to avoid and survive Sea Lamprey (*Petromyzon marinus*) predation.

INTRODUCTION

The Anthropocene has been characterized by immense reductions in biodiversity and widespread local species extirpations (Cardinale et al. 2012), which have disproportionately affected freshwater taxa (Dudgeon et al. 2006; Ricciardi & Rasmussen, 1999). Our ability to mediate the consequences of this extinction crisis will largely depend on our ability to conserve intact populations and efficiently restore those that have declined or been lost (Hayward & Slotow, 2016). For species of disproportionally high economic, cultural, and ecological importance; local extirpation or population decline has often been followed by human-mediated restoration efforts to reintroduce taxa using individuals from remaining wild or captive populations, improve habitat, and address other issues that might hinder recovery. Given rapid declines in biodiversity, substantial collaborative species reintroduction efforts will undoubtedly become more common (IUCN/SSC, 2013); however, restoration programs relying on captive populations are often unsuccessful (Reading et al. 2013). Given that functional aquatic systems are essential for the continued availability of access to clean water, food, and energy for all humans

(Dudgeon et al. 2006), it is vital that we improve our understanding of the factors that underly the recovery of wild populations of aquatic (and terrestrial) species during the decades following the outset of human-mediated reintroduction efforts.

The international effort to restore extirpated wild Lake Trout (*Salvelinus namaycush*) populations in the Great Lakes of North America provides an excellent example of the sort of large-scale species recovery efforts that began during the second half of the 20th century, and will likely increase in frequency in the future (Muir et al., 2012). Lake Trout were an abundant top predator in the Great Lakes prior to European settlement of the Great Lakes region (Hansen, 1999). Following the basin-wide collapse of the Lake Whitefish commercial fishery in the early 20th century, fishing pressure was largely transferred to Lake Trout populations, which caused significant declines in abundance between 1930 and 1960 (Hansen, 1999). An invasive predator, the Sea Lamprey (Petromyzon marinus), also invaded the Great Lakes during this time, leading to further increases in adult mortality that ensured functional extirpation from all lakes except Lake Superior and a small isolated population in Lake Huron (Hansen, 1999).

The restoration program that commenced largely focused on Sea Lamprey population control, the creation of aquatic refuges, habitat improvement, and stocking juvenile Lake Trout from a diverse collection of domesticated hatchery strains that were founded by remnant wild Lake Trout in the region (Hansen, 1999; Muir et al., 2012). Hatchery strains originated from multiple locations: Green Lake, Wisconsin; Lewis Lake, Wyoming; Apostle Island, Isle Royale, and other locations in Lake Superior; Georgian Bay in Lake Huron; and Seneca Lake, New York (Muir et al., 2012). All strains except for the Seneca Lake strain were founded by individuals from Great Lakes populations, or

populations that were originally founded by stocking Great Lakes individuals (Krueger et al., 1983). Lake Trout populations in Lake Superior rebounded relatively quickly; however, recovery was not realized in other Great Lakes (Hansen, 1999). In other Great Lakes, the re-emergence of natural reproduction was hindered by high levels of Sea Lamprey predations on adults (Pycha et al., 1980), predation on juveniles by invasive alewife (Madenjian et al., 2008; *Alosa pseudoharengus*), reduced juvenile survival caused by thiamine deficiency (Fitzsimmons et al., 2010), and potentially reduced hatching success associated with PCB contamination (Mac and Edsall, 1991). Despite these complicating factors, wild reproduction was first observed in Lake Huron starting in the 1980s, with levels of natural recruitment gradually increasing through the present (He et al., 2012; Hansen ,1999). Evidence of natural recruitment was later observed in Lake Michigan (Hanson et al., 2013) and Lake Ontario (Gatch et al., 2021). However, the restoration of viable, self-sustaining, populations has only been achieved in Lake Superior and remains an important international fisheries management objective for all other lakes (Hansen, 1999).

Multiple lines of evidence suggest that hatchery Lake Trout strains that were introduced across the Great Lakes have variable fitness in the wild. Specifically, work by Scribner et al. (2018) and Larson et al. (2021) indicate that hatchery strains originating from Seneca Lake, New York contributed disproportionately to wild recruitment in Lakes Huron and Michigan. One potential explanation for this phenomenon is that variation is due to adaptive genetic differences between strains (He et al., 2016); however, the genetic basis underlying strain adaptive variation has not been examined. No attempts have been made to identify the specific loci that underly differences in fitness between strains in the wild due to the lack of genomic resources. Previous studies have found that the Seneca strain is

less prone to Sea Lamprey induced mortality (Schneider et al., 1996), have distinct patterns of seasonal habitat use (Bergstedt et al., 2003), and expresses different age-specific growth rates than Great Lakes origin hatchery strains (Elrod et al., 1996). Additionally, the native Lake Trout population in Seneca Lake (which founded the U.S. and Canadian Seneca hatchery strains) has coevolved in sympatry with Sea Lamprey (Bronte et al. 2007; Schneider et al. 1996), and lamprey predation is the primary source of adult mortality in many contemporary Great Lakes environments (Sitar et al., 1999). Seneca Lake origin individuals likely have unique behavioral or physiological adaptations that allow them to avoid or survive Sea Lamprey predation. These adaptations are likely a byproduct of predator-prey coevolution in ancestral environments, the signatures of which may be measured using recently developed genomic resources (Smith et al., 2020; Smith et al., 2021).

It is also critical to note that Great Lakes hatchery strains appear to have been founded by severely bottlenecked populations in many cases (Page et al., 2004; Guinand et al., 2012). The collapse of Lake Trout populations in the Great Lakes largely occurred over a single generation (Guinand et al., 2003), and population genetics theory predicts that rapid declines in effective population size will result in an increase in the frequency of some deleterious recessive alleles due to drift (Bortoluzzi et al., 2020). Domesticated environments are also often associated with low effective population size (Page et al., 2005; Marsden et al., 2016), and strong drift in the hatchery environment could have further exacerbated this issue (Fleming et al., 2001; Fernández & Caballero., 2001).

Genomic methods provide researchers with unprecedented opportunities to dissect the genetic basis for variation in fitness and ecologically relevant traits in captive and wild

populations. Previous work has suggested some level of inter-strain hybridization in recovering Lake Huron Lake Trout populations (Scribner et al., 2018), and hybridized populations provide unique opportunities for identifying genomic regions associated with variation in fitness (Pool, 2015; Leitwein et al., 2020). Hybridization beyond the F1 generation produces haplotypes with a mosaic of ancestry states, and the distribution of tracts of ancestry across the genome is expected to be distorted by natural selection (Secolin et al., 2019; Leitwein et al., 2020). Specifically, if certain alleles contribute to increased fitness and those alleles are at different frequencies in ancestral populations, then we expect selection to favor haplotypes originating from the ancestral population where beneficial alleles are at higher frequency (Oleksyk et al. 2010). In F2 hybrids and later generations, this pattern will manifest as a discrepancy between global (genome-wide) ancestry and locus-specific ancestry at loci associated with variation in fitness. Multiple studies have used this effect to document evidence of selection in human and animal populations (Kovach et al. 2016; Oziolor et al. 2019; Jones et al. 2018; Pierron et al. 2018). Strong linkage disequilibrium resulting from recent admixture (Admixture LD; ALD) is also beneficial in many cases because signals of selection will typically be readily apparent across large chromosome segments; however, these allelic correlations can potentially make it difficult to determine which genes are the true targets of selection (Smith & O'Brien, 2005). Some biological processes are also expected to produce locus specific excesses of hybrid ancestry (Runs of Hybridity; R_{HYB}; where two haplotypes are descended from different source populations) within hybrid individuals. For example, if a deleterious recessive allele is at elevated frequency in some ancestral populations, then

individuals with hybrid ancestry at this locus will have higher fitness on average than individuals with certain homozygous local ancestry (Kim et al. 2018).

The primary goal of this study was to identify genomic regions associated with variation in fitness in re-emerging populations of wild Lake Trout in Lake Huron. To this end, we had five main predictions for this study. (1) Given that multiple previous studies have shown that Seneca Lake origin Lake Trout contribute disproportionately to wild recruitment, we predicted that we would identify multiple genomic regions with a significant excess of Seneca origin alleles across F2 admixed individuals. (2) We also predicted that loci with distorted ancestry proportions would overlap, or be in close proximity to, loci associated with adaptive divergence between strains. (3) We predicted that adaptively diverged loci within regions associated with fitness variation would be disproportionately associated with biological processes related to the ability to survive or avoid Sea Lamprey predation. (4) We predicted that we might identify a comparatively small number of loci with an excess of haplotypes originating from Great Lakes hatchery strains in F2 hybrid individuals. And lastly, (5) if deleterious recessive alleles are at elevated frequency in some or all Great Lakes hatchery strains, or if certain loci are associated with hybrid vigor, we predicted we would identify multiple loci with an excess of R_{HYB} in F2 hybrid individuals. As a prerequisite to evaluating these predictions, we identified F1, F2, and purebred Lake Trout and characterized patterns of admixture among Seneca Lake and Great Lakes origin hatchery strains across time.

MATERIALS AND METHODS

LABORATORY METHODS AND SAMPLES

Samples were obtained from adults of eight domestic hatchery Lake Trout strains that were used to produce offspring that were stocked in Lake Huron throughout the course of reintroduction (Scribner et al., 2018). The history of hatchery strain development and captive rearing are described in Page et al. (2005). Adult samples included the Isle Royale (SIW; IR), the Marquette (SMD; MQ), Apostle Islands (SAW, AI), Green Lake (GLW; GL), Lewis Lake (LLW, LL), Canadian Seneca Lake (SLC; CS), U.S. Seneca Lake (SLW, SL), and Parry Sound (HPW, PS) strains (n=24 per strain). Samples were also obtained from state (Michigan Department of Natural Resources) and federal (U.S. Fish and Wildlife Service and U.S. Geological Survey) fisheries management agencies for wild born Lake Trout from Lake Huron collected between 2002 and 2017 (n=1146). Individuals were identified as being wild born based on the lack of clipped fins. Wild born samples were divided into three temporal categories corresponding to individuals born during early (n=210), middle (n=659), and late (n=277) stages of re-emerging natural recruitment (collected 2002-2004, 2009-2012, and 2016-2017). Samples were collected from 5 fisheries management units within the U.S. waters of Lake Huron – MH1, MH2, MH3, MH4, and MH5 (ordered from North to South). Samples from the three southern management units (MH3, MH4, and MH5) were combined due to low sample size in MH4 and MH5 (combined samples are referred to as MH3 samples hereafter). Ages for all wild born samples were determined by cooperating agency personnel who collected the samples using boney structures (otoliths or maxilla; Wellenkamp et al., 2015). All fish from Lake Huron were sampled at between 4 and 13 years of age and were born in the wild over a 19-year timespan (1994-2013).

DNA was extracted from fin tissue, scale tissue, or tissue scraped from maxilla using Qiagen DNeasy kits (QIAGEN, Inc., Germantown, MD) and the manufacturer recommended protocols. DNA quality and quantity were assessed using a Nanodrop 2500 or Nanodrop 1000 spectrophotometer. Samples with 260/280 or 260/230 ratios less than 1.7 or concentrations less than 15 ng/ul were cleaned and concentrated using a 1:1 (beads to DNA) Ampure XP clean-up (Beckman-Coulter). Prior to genotyping, double stranded DNA (dsDNA) was quantified using Quant-It Picogreen assays (Life Technologies). Samples were then diluted to a uniform concentration of 5ng/ul (dsDNA) in low EDTA TE (Teknova) before proceeding with library preparation.

Pst1 RAD libraries were prepared using the protocol described in Ali et al. (2016) with modifications described in Smith et al. (2020). Libraries were sheared to a mean fragment size of 350 bp after adapter ligation using a Covaris E220 Ultrasonicator (Covaris) and the manufacturer recommended protocol and were amplified for 11 cycles. Completed libraries were quantified using Quant-It Picogreen assays (Life Technologies) run in triplicate. The distribution of fragment sizes for libraries was measured using Genomic DNA Tapestation assays (Agilent) after shearing and amplification. Batches of 3 libraries were pooled in equimolar amounts and enriched for 58,889 variable RAD loci described in Smith et al. (2020) using a MyBaits v3 Custom Target Enrichment Kit (Arbor Biosciences, Ann Arbor, Michigan). Baits were allowed to hybridize to target loci for 16 hours and hybridization and wash reactions were carried out at a temperature of 65ºC. Target enriched libraries were amplified for 10 cycles using a KAPA Library Amplification Kit for Illumina using manufacturer recommended cycling conditions and cleaned twice

using 0.9:1 Ampure XP clean-ups (Beckman Coulter). Each pool (3 plates) was sequenced in a single HiSeq 4000 or HiSeq X sequencing lane using 2X150 base pair paired end-reads.

BIOINFORMATICS

The quality and quantity of sequencing data were initially assessed using FastQC (Andrews, 2010). We then re-oriented reads such that inline, individual specific, barcodes were located at the beginning of the first read using a custom Perl script. Reads for individual samples were then demultiplexed using the process_rad_tags program from Stacks v.2 (Rochette et al., 2019), with the restriction enzyme set to Pst1. PCR duplicates were then removed using the program clone_filter, also from Stacks v.2. We then used Trimmomatic v0.32 (Bolger et al., 2014) to clip sequencing adapters and trim reads whenever the mean base quality in 4bp sliding windows dropped below Q15. Paired-end reads were retained if both reads were longer than 50bp after trimming. Reads were then mapped to the RefSeq version of the Lake Trout genome (Smith et al., 2021) using bwa mem (Li, 2013) with default settings and resulting bam files were sorted using samtools sort (Li et al., 2009).

Genotypes were called using gStacks using the marukilow genotyping model (Maruki & Lynch, 2017), a minimum mapping quality of 10, a maximum soft clipping level of 0.2, and alpha thresholds of 0.05 for variant and genotype calling. Genotypes were retained if likelihoods for the called genotype and the second most likely genotype were different by one or more order of magnitude (GQ>10) using vcftools (Danecek et al., 2010). Loci were removed from the dataset if the minor allele was observed fewer than 3 times across all hatchery and wild individuals or if genotypes were called for fewer than 70% of

individuals with a genotype quality greater than 10. Individuals were then removed if the count of missing genotypes was greater than 2 standard deviations from the mean.

Finally, we required the minor allele frequency to be greater than or equal to 0.05 in at least one hatchery or wild collection and required that more than 50% of individuals had called genotypes in all collections in order to retain a locus. Confounded paralogous loci were then detected and removed using a modified version of HDplot (McKinney et al., 2017; https://github.com/edgardomortiz/paralog-finder;) that accounts for the presence of missing data. Confounded paralogous loci were removed using a D threshold of 5 and a maximum heterozygosity threshold of 0.55 after evaluating resulting plots of heterozygosity and allelic read ratios. The resulting VCF file was phased and imputed using Beagle v5.2 (Browning & Browning, 2016) prior to additional analysis.

GLOBAL ANCESTRY AND HYBRID CLASSIFICATION

We initially explored the genetic relationships among hatchery strains and wild born Lake Huron individuals using principal components analysis (PCA). PCA was conducted using the complete set of phased and imputed genotypes. The genomic covariance matrix was calculated in R (R Core Team, 2020; VanRaden, 2008) and the eigen function was used to compute eigenvectors. This analysis was initially conducted using all hatchery and wild born Lake Huron individuals but was repeated using only hatchery individuals after observing large differences in sample size between clusters in our initial analysis, which can potentially bias results (Burgos-Paz et al., 2014).

Hatchery records suggest that three of the hatchery populations stocked in Lake Huron and elsewhere are potentially admixed. These included the Apostle Islands, Marquette, and Green Lake strains (Krueger et al., 1983; Page et al., 2004). Additionally,

PCA suggested possible inter-strain hybridization in the collection of wild born Lake Huron individuals, as suggested by Scribner et al. (2018). Given these results, genome-wide (global) ancestry coefficients were calculated using ADMIXTURE (Alexander et al., 2009) for all K values between 1 and 20 for all hatchery origin individuals. The optimum value of K was determined using cross validation with the –cv option; however, ancestry plots were produced for all K values between 2 and 7. Ancestry coefficients were then projected using the -P option for all hatchery origin individuals for K values with the lowest cross-validation errors.

Individuals were assigned to hybrid categories using a simulation and classification-based approach inspired by Barker et al. (2019). Briefly, a total of 39 inter-strain hybrid categories were simulated using the hybridize function in the R package adegenet with genotypes from hatchery origin individuals used as input (Jombart, 2008). We simulated all possible F1 hybrids, as well as all potential backcrosses and double-backcrosses of Seneca and Great Lakes origin strains. We did not simulate F1 intercrosses because we ultimately differentiated these individuals from F1 hybrids using information on the distribution of R_{HYB} across individual genomes. Additionally, we did not simulate hybrids beyond the F3 generation. Parameters for a discriminant analysis of principle components (DAPC; Jombart et al., 2010) classification model were selected using 30-fold cross validation with the xvalDAPC function with between 10 and 100 principal components used for testing. Self-assignment probabilities were calculated for all hybrid categories and all hatchery populations and were visualized as a confusion matrix in order to verify that out-of-bag simulated samples could be accurately assigned to their respective hybrid category. The resulting model was used to predict hybrid categories for all wild born individuals and

calculate posterior probabilities of assignment to each respective group. Individuals were assigned to a hybrid category if their posterior probability of assignment was greater than 0.99.

LOCAL ANCESTRY INFERENCE

We used the program RFmix (Maples et al., 2013) to determine the ancestral origins of haplotype segments and identify R_{HYB} within the genomes of wild-born individuals. For this analysis, we sought to classify haplotype segments as Seneca or Great Lakes origin in order to identify genomic regions with elevated penetrance of Seneca alleles relative to null expectations. Individuals were included in one of the two haplotype reference panels (Great Lakes and Seneca Strain) if they were sampled directly from a hatchery or if their posterior probability of assignment to a hatchery strain (from DAPC) was greater than or equal to 0.99. For RFmix, we used a window size to 0.5 centimorgans, an n parameter of 5, a prior expectation of 3 generations since the onset of hybridization, and an interpolated genetic map as input (Smith et al., 2020). The genetic positions for genotyped markers were estimated by modeling genetic location from the Lake Trout linkage map (in centimorgans; Smith et al., 2020) as a function of physical location (in bp) using a loess model with a 2nd degree polynomial and a span of 0.2 (Rezvoy et al., 2007). Loess models were fit using the loess function from the R-package (R Core Team, 2020) stats. Genetic positions were then predicted from physical positions for genotyped loci. Markers at the beginning or end of chromosomes without interpolated map positions were considered to be 1 cM away from the nearest marker with an estimated map position.

Based on ancestry tracts inferred using RFmix, we calculated the proportion of Seneca and Great Lakes origin alleles (P_S and P_{GL}, respectively) and the proportion of the

genome composed of runs of hybridity (R_{HYB}) for each wild individual. Individuals that were initially identified as F1 hybrids by DAPC were reclassified as intercross hybrids if less than 90% of their genome was composed of R_{HYB} after evaluating the relationship between P_S and R_{HYB} relative to the expected values for F1 hybrids (0.5 and 1, respectively). Individuals initially identified as F2 backcrosses to Seneca were confirmed if less than 95% of their alleles originated from the Seneca Lake ancestral population, their DAPC posterior probability was greater than 0.9, and more than 10% of their genome was composed of R_{HYB}. Individuals initially identified as F2 backcrosses to Great Lakes strains were confirmed if the proportion of Great Lakes origin alleles was less than 95%, their DAPC posterior probability was greater than 0.9, and more than 10% of their genomes were composed of R_{HYB}. Haplotype ancestries were extracted for all high confidence intercrosses and F2 hybrids and these haplotypes were used to test whether or not certain loci were associated with hybrid vigor or differences in fitness between hatchery strains.

Within these advanced stage hybrids, we expected to observe an excess of haplotypes derived from the Seneca strain in regions where Seneca strain individuals carry alleles associated with elevated fitness. Conversely, we expect to observe a deficit of Seneca origin haplotypes in regions where Great Lakes origin alleles provide a fitness advantage. Expected distributions of our test statistics (P_S and P_{GL}) were generated by 1) concatenating haplotypes for all chromosomes onto single pseudo-haplotypes, 2) circularizing these artificial haplotypes, and 3) cutting haplotypes at a random location drawn from a discrete uniform distribution in order to re-linearize the genome (similar to Tang et al., 2007). This process was carried out for each of the two haplotypes within each individual and repeated 1,000 times. At each iteration, we calculated P_S and P_{GL} for each

locus in order to generate a distribution of our test statistics under the null hypothesis that ancestry states are randomly distributed across the genome. Significance was assessed using a permutation p-value and an alpha threshold of 0.01. We should note that a Bonferroni correction is inappropriate in this case, due to high levels of linkage disequilibrium in F2 hybrids. This process was repeated for all wild-born backcrosses and intercrosses combined (n=97), backcrosses to Great Lakes strains (n = 28), and backcrosses to the Seneca strain (n =31). The test was run separately for various F2 hybrid classes for two reasons. First, different alleles might only be associated with a fitness advantage on certain genetic backgrounds. Second, the theoretical maximum deviation between null and observed distributions of P_S and P_{GL} are inherently limited by genetic background, and we speculated that we would have higher power to detect extreme excesses of Seneca ancestry within individuals with genetic backgrounds that are primarily composed of Great Lakes origin alleles. We also tested for locus specific excesses and deficits of R_{HYB} using the same permutation procedure; however, only the combined dataset of intercrosses and backcrosses was considered and a vector of 0s and 1s (corresponding to presence or absence of R_{HYB}) was permuted rather than each of the two haplotypes for each individual.

ADAPTIVE DIVERGENCE BETWEEN STRAINS

We tested for evidence of adaptive divergence between hatchery strains using two methodologies. We first used the R-package pcadapt (Luu et al., 2017) to detect loci with significant differences in allele frequency between the clusters identified via PCA. Loci that were significantly associated with the first principal component (which separated the Seneca Strain from other strains) were considered to be associated with Seneca strain

divergence. Loci associated with the second principal component (which separated the Parry Sound strain from others) were considered to be associated with adaptive divergence of the Parry Sound strain from other strains. Significance was assessed at a p-threshold of 0.01.

We supplemented pcadapt results with those of the core model outlier test implemented in BAYPASS (Gautier, 2015). For this analysis, we also defined allele frequency contrasts based on the clusters identified by ADMIXTURE at K = 7, and used the p-values obtained from a C2 contrast test (Olazcuaga, et al., 2020) to determine which hatchery strains were primarily driving observed patterns of divergence, if any. All hatchery populations with the exception of the Marquette strain and Apostle Islands strain were split at this value of K, with the Apostle Islands and Marquette strains being identified as a single admixed cluster that likely shared adaptive variation. We also simulated a data set composed of 100,000 SNPs under the Nicholson et al. (2002) model of hierarchical population structure, which was parameterized with the population covariance matrix obtained from our initial run of BAYPASS. BAYPASS was then re-run in order to determine the expected distribution of the XtX differentiation statistics under neutrality. Significance was assessed at a p-threshold of 0.01 given the observation that p-values of this threshold were expected to be exceedingly rare without the action of selection based on simulations (<1% false positive rate). C2 contrast statistic p-values were determined to be significant at the same threshold. Results from these outlier tests for all loci are available in Supplementary Material 4.1.

GENE SET ENRICHMENT ANALYSIS

Gene ontology (GO) terms were identified for all coding sequences (CDS) in the Lake Trout genome using the Panzzer2 functional annotation server (Törönen et al. 2018). Translated coding sequences from Lake Trout Annotation Release 100 were accessed from the RefSeq FTP site on April 14th, 2021 (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/016/432/855/GCF_016432855.1_SaNama_1.0/). We required a minimum query coverage of 0.6, a minimum subject coverage of 0.6, a minimum alignment length of 100, and sequence identity between 0.4 and 1.0 to retain matching sequences. We output a single DE for each query sequence and required a form factor of 0.2. GO term prediction was done using the Argot scoring function, with redundant GO terms being removed using a Blast2GO threshold of 55. GO terms for coding sequences located on assembled chromosomes that were associated with annotated proteins were used as the baseline data set for gene set enrichment analysis.

We performed a gene set enrichment analysis (GSEA) using the runTest function from the R-package topGO, with the algorithm set to "weight01" and the statistic argument set to "fisher" (Alexa & Rahnenführer, 2009). We sought to identify enriched GO terms associated with SNPs that were outliers between hatchery strains and located within regions that were local ancestry outliers. SNPs were selected by dividing the genome into 500 Kb windows offset by 100 Kb. For each window, we tabulated the number of significant tests, identified windows with 2 or more significant tests, then merged overlapping windows using bedtools (Quinlan & Hall, 2010). For each window we selected the SNP with the largest -log10 p-value (across all BAYPASS and C2 contrast outlier tests),

determined which genes were within 100 Kb of this focal SNP, then selected the closest gene to each of these SNPs for GSEA.

RESULTS

BIOINFORMATICS

After filtering on genotype quality, minor allele count, and genotype missingness, we were left with 98,879 SNPs distributed across 49,940 RAD loci. Of these RAD loci, 30,366 contained a single SNP, 9,867 contained 2 SNPs, 3,922 contained 3 SNPs, 2,041 contained 4 SNPs, and 3,744 contained 5 or more SNPs. A total of 1,321 individuals remained in the dataset after removing samples with high levels of missing data. After requiring a population minor allele frequency of 0.05 and maximum population missingness of 0.5, we were left with 57,799 SNPs. After filtering out potential paralogous loci with HDplot (McKinney et al., 2017) and phasing and imputing genotypes with Beagle v5.2 (Browning & Browning, 2016), we were left with 41,989 SNPs that were used for subsequent analysis.

GLOBAL ANCESTRY AND HYBRID CLASSIFICATION

Principal components analysis conducted using the complete set of wild and hatchery origin individuals suggested the existence of three clusters corresponding to the two Seneca Lake origin strains, the Lewis Lake strain, and all other Great Lakes origin strains (Figure 4.1A). The majority of wild born individuals appeared to cluster with the Lewis Lake and Seneca Lake populations (Figure 4.1A). A number of wild born individuals exhibited principal component scores that were intermediate distances between the three clusters, indicating the existence of inter-strain hybrids (Ma & Amos, 2012). The first principal component, which separated the Seneca origin strains from all other populations, explained 5.73% of variance. The second principal component, which split the Lewis Lake

strain from other Great Lakes origin strains, only explained 0.88% of variance. The principal components analysis conducted using only hatchery origin individuals also separated populations into three clusters; however, the Parry Sound strain was split from the other Great Lakes origin populations rather than the Lewis Lake strain. The first and second principal components from this analysis explained 6.6% and 3.34% of variance, respectively.

Cross validation errors for various values of K for ADMIXTURE were minimized at K=3, which corresponds to the separation of all Lake Superior and Lake Michigan origin strains, Seneca Lake origin strains, and the Parry Sound strain derived from Lake Huron (Figure 4.2). Lake Michigan origin ancestry was separated at K=4. As expected, the Green Lake and Lewis Lake strains appear to be almost entirely of Lake Michigan ancestry, while the Marquette strain and Apostle Islands strains appear to have a mixture of Lake Michigan and Lake Superior ancestry at this value of K. The two Lake Michigan origin strains (Lewis Lake and Green Lake) are separated at K=5, and these results suggest that the Lake Michigan ancestry in Apostle Islands and Marquette strains is derived from the Green Lake strain, which is consistent with hatchery records (Krueger et al., 1983). The Apostle Islands and Marquette strains form a distinct but apparently admixed cluster at K=6 and a number of Apostle Islands origin individuals are assigned to a separate cluster at K=7.

We assigned 1,147 wild individuals to hybrid categories using DAPC models built for the purpose of hybrid classification. Confusion matrices suggested that individuals could be assigned to F1 and F2 categories with 100% certainty; however, 0-30% of purebred individuals from Great Lakes strains were miss-assigned as F3 backcrosses to Great Lakes populations. F3 backcrosses to the Seneca strain were identified with theoretical accuracies

between 98 and 100%. Of the 25 individuals identified as F3 backcrosses to Great Lakes origin populations, 12 were from the earliest collection (collected 2002-2004) and it is extremely unlikely that hybridization would have proceeded to the F3 generation by this point in time. Lake Trout typically mature at an age of 6 or 7 (Sitar et al., 2014) and the first Seneca origin fish stocked in Lake Huron were from the 1984 year-class. F1 hybrids therefore could have been spawned as early as 1990 or 1991, and F2 hybrids could have been spawned as early as 1996 assuming a minimum age at maturity of 6. Given this result and the large proportion of purebred individuals that were classified as F3 backcrosses, all F3 backcrosses were re-classified as purebred individuals based on their primary hatchery strain ancestry. Similarly, putative F3 backcrosses to Seneca were re-classified as individuals of primarily Seneca Lake ancestry. No individuals that were putatively identified as F3 backcrosses were used for detecting signals of adaptation.

776 (67.65 %) of wild born individuals were identified as Seneca strain origin fish. 109 (9.5%) individuals were classified as Great Lakes origin fish. 50 (45.8%) of these individuals had genotypes that were consistent with F1 hybrids between the Lewis Lake strain and other Great Lakes origin strains. 34 of these 50 fish appear to be F1 hybrids between the Lewis Lake strain and the Marquette strain, 13 were Lewis Lake – Apostle Islands F1 hybrids, 2 were Lewis Lake - Green Lake F1 hybrids, and 1 was a Lewis Lake – Isle Royale hybrid.

A total of 181 individuals were initially identified as F1 hybrids between the Seneca strain and Great Lakes origin strains and 38 of these individuals were reclassified as F2 intercrosses after evaluating the distribution of R_{HYB} across their chromosomes (see Figure 4.3). This left 143 high confidence Seneca – Great Lakes F1 hybrids (12.46% of all wild fish

sampled). 14 (9.7%) of the 143 remaining F1 hybrids were produced by Seneca Lake and Apostle Islands strain parents, 3 (2.1%) were produced by Green Lake and Seneca Lake origin parents, 2 (1.39%) were produced by Isle Royale and Seneca origin parents, 95 (66.43%) were produced by Lewis Lake and Seneca Lake origin parents, and 28 (19.5%) were produced by Marquette and Seneca Lake origin parents. F1 individuals were born as early as 1994 in management units MH2 and MH3, but not until much later in MH1 (2002).

F2 intercross individuals (n=38, 3.3% of all wild fish) were primarily of Seneca Lake and either Marquette (21.05%) or Lewis Lake (68.4%) ancestry. Based on these results, F2 intercross individuals were born as early as 1996 in management unit MH3, 1997 in MH2, and 1998 in MH1. One F2 intercross had primarily Seneca and Green Lake ancestry, one had primarily Apostle Islands and Seneca Lake ancestry, and 2 had primarily Isle Royale and Seneca Lake ancestry.

We identified 44 F2 backcrosses to Seneca strain parents. These individuals were born as early as 1999 in MH2 and MH3. The 2003 year-class was the earliest in which F2 backcrosses to Seneca were observed in MH1, suggesting that admixture proceeded more slowly in the Northernmost management unit. The Great Lakes component of these individual's ancestries were primarily from the Marquette strain for 17 (38%) fish and primarily from the Lewis Lake strain for 14 (41.8%) fish. The Great Lakes origin ancestry for 7 (15.9%) of these fish most likely originated from the Apostle Islands strain. The Great Lakes origin ancestry for 4 (9.09%) of these fish most likely originated from the Isle Royale strain. The Great Lakes origin ancestry for 2 (4.54%) of these fish most likely originated from the Green Lake strain. F2 backcrosses with Isle Royale, Apostle Islands, Marquette,

Lewis Lake, and Green Lake ancestry were born as early as 1999, 2000, 1999, 2000, and 2006, respectively.

We identified 31 F2 backcrosses to Great Lakes strain parents. F2 backcrosses to the Lewis Lake strain were born as early as 1994 and 1995 in MH1 and MH3, respectively. This result was surprising given that 1996 was the earliest year class we expected to observe F2 hybrids. Assuming these individuals were not miss-classified, this suggests that Seneca Lake origin fish can mature as young as age 5. None of the simulated F2 backcrosses to Great Lakes origin parents were miss-classified; however, it is theoretically possible that these wild born individuals are actually of purebred Great Lakes origin fish. The Great Lakes origin ancestry component was mostly derived from the Lewis Lake strain for 18 fish (58.1%), the Marquette strain for 10 fish (32.2%), the Green Lake strain for 2 fish (6.4%), and the Apostle Islands strain for 1 fish (3.2%). All F2 backcross samples from year-classes spawned after 1998 were of Seneca Lake and either Lewis Lake or Marquette strain ancestry.

DAPC results indicated that F2 backcrosses to Seneca origin parents could be identified with perfect accuracy and these fish were spawned as early as 1999. Assuming an average age at maturity of 6 or 7 (Sitar et al., 2014), this suggests that the first F1 hybrids were born in 1992 or 1993 in Lake Huron and that 2005 or 2006 would likely be the first year that F3 hybrids would be spawned (~2011 for F4 and ~2017 for F5). The first F2 intercross individuals were sampled from the 1996 year-class. This suggests that F1 hybridization began as early as 1989 or 1990, which is feasible if the first Seneca origin fish that were planted in Lake Huron matured at age 5 or 6 (Madenjian et al., 1998). The observation that F2 backcrosses to Great Lakes origin parents were sampled from the 1994

year-class complicates matters to some extent. If these individuals are indeed F2 backcrosses, then this suggests that their Seneca origin grandparent and F1 hybrid parent both reached sexual maturity at an age of 4 or 5, which is theoretically possible for Lake Trout (Martin & Olver, 1980). We conclude that the first F1 hybrids between Seneca Lake and Great Lakes origin strains were spawned sometime between 1989 and 1993 and that the first wild F2 hybrids were born sometime between 1994 and 1999.

LOCAL ANCESTRY

Locus specific ancestries were estimated for 1,321 hatchery and wild born individuals. Global ancestry proportions (calculated as the proportion of alleles from Seneca Lake vs the Great Lakes hatchery strains) were consistent with expectations for purebred, F1, F2 intercross, and F2 backcross individuals (Figure 4.4). We chose to assign haplotypes to Great Lakes vs Seneca Lake ancestral populations (rather than to individual hatcheries) because ADMIXTURE results indicated that many of the Great Lakes origin populations likely shared haplotypic variation due to historical admixture among some strains. The existence of the same haplotypes in multiple Great Lakes hatchery strains would likely lead to spurious results from the haplotype classification model of RFmix. Rather, we classified F2 hybrids into two groups – hybrids for which the Great Lakes ancestry component was from Lake Michigan origin strains (mostly Lewis Lake, but also Green Lake) and hybrids for which the Great Lakes ancestry component was primarily from Lake Superior origin strains (mostly Marquette, but also the Apostle Islands and Isle Royale strains). These different groupings were used to determine if signals of selection and excesses or deficit of R_{HYB} were dependent on the source of Great Lakes genetic background.

Tests for elevated penetrance of Seneca origin haplotypes identified a total of 6 genomic regions for which Seneca origin alleles appear to provide a fitness advantage in Lake Huron (Figure 4.5A and 4.5C). Values of P_S were significantly elevated on chromosomes Sna11, Sna19, and Sna29 and were significant regardless of whether or not we evaluated backcrosses to Great Lakes populations or all F2 hybrids collectively (Figure 4.5A and 4.5C). Peaks on chromosomes Sna3 and Sna34 were only significant when backcrosses to Great Lakes populations were considered separately, suggesting that Seneca origin alleles at these loci only provide a fitness advantage when the genetic background is primarily composed of alleles from Great Lakes populations (Figure 4.5A).

Interestingly, the significant region on Sna3 also exhibited a significant excess of R_{HYB} relative to null expectations suggesting that this signal of selection is driven by elevated hybrid fitness (Figure 4.6). This region of Sna3 also overlaps a quantitative trait locus that explains variation in skin pigmentation in other Lake Trout populations (Smith et al., 2020); however, the exact trait or traits that are under selection in this case are unknown. An excess of hybrid genotypes was also detected on one arm of Sna8; and this region was slightly below our significance threshold for detecting loci with an excess of Seneca haplotypes. One region with a significant deficit of R_{HYB} was also detected on chromosome Sna21, suggesting this locus might be associated with decreased hybrid fitness (Figure 4.6).

We identified a single genomic region on chromosome Sna35 where F2 hybrids exhibited a significant excess of haplotypes originating from Great Lakes populations (Figure 4.5B). Admixture outlier regions ranged in size from 2.29 – 15.37 Mb (Table 4.1), reflecting extensive admixture linkage disequilibrium expected in collections of F2

individuals (Smith & O'Brien, 2005). Collectively, these results indicate that selection primarily favors Seneca origin alleles, that some Great Lakes origin individuals carry alleles associated with increased fitness on chromosome Sna35, and that two regions of the genome are potentially associated with hybrid vigor in Seneca – Great Lakes hybrids.

Results were quite different when individuals were split based on whether or not their Great Lakes ancestry component was derived from a Lake Superior or Lake Michigan origin strains. Specifically, the highly significant excess of Seneca alleles on Sna11 was only apparent within F2 hybrids for which the primary Great Lakes ancestry was from a Lake Michigan strain (most often the Lewis Lake strain; Figure 4.7A). This signal of selection was significant regardless of whether or not Seneca backcrosses were included in the analysis. Great Lakes origin haplotypes were not found to be at significantly elevated frequency in any cases within these individuals (Figure 4.7A; the Sna35 regions detected above was just below significance thresholds). Conversely, hybridized individuals with a Great Lakes ancestry component primarily from Lake Superior strains (most often the Marquette strain) exhibited locus specific excesses of Great Lakes origin alleles on chromosome Sna35 (the same region detected in our initial tests) and in a region of chromosome Sna1 (Figure 4.7B). No excesses of Seneca origin alleles were detected when all of these individuals were analyzed collectively; however, a large fraction of these individuals were backcrosses to Seneca (50%) and this likely minimized the maximum observable deviation from the null hypothesis. Two loci with a significant excess of Seneca alleles were detected when backcrosses to Seneca were dropped from this analysis (Table 4.1). These included a region of chromosome 8 (the same locus showing an excess of hybridity in Figure 4.6B) and a

region of chromosome 19 (the same locus with an excess of Seneca alleles in Figure 4.5A and 4.5C).

Additionally, within the Lake Michigan ancestry F2 collection, we observed a significant excess of R_{HYB} on 4 chromosomes (Sna3, Sna5, Sna8, and Sna36; Figure 4.8A; Table 4.1). Regions on chromosomes Sna12 and Sna31 were slightly below significance thresholds. Furthermore, chromosome-wide deficits of R_{HYB} were encountered within these individuals on Sna21 and Sna35 (Figure 4.8A). Within the Lake Superior ancestry F2 collection, deviations from the expected distribution of R_{HYB} were only detected on a single arm of chromosome Sna1, and this region exhibited an excess of R_{HYB} relative to null expectations (Figure 4.8B).

ADAPTIVE DIFFERENCES BETWEEN STRAINS

We identified 1,017 SNPs with signals of adaptive divergence between strains based on the BAYPASS core model. Based on the C2 contrast tests, we identified 846 loci with highly differentiated allele frequencies in the Seneca strain, 733 loci with highly differentiated allele frequencies in the Parry Sound strain, 550 in the Marquette and Apostle Islands strains, 546 in the Isle Royale strain, 126 in the Lewis Lake strain, and 207 in the Green Lake strain. A total of 68 outlier loci were detected by the BAYPASS core model and located within genomic regions exhibiting excesses and deficits of R_{HYB} and P_S (local ancestry outlier regions). We also identified a total of 79 Seneca strain C2 contrast test outlier loci within these regions. These were located on chromosomes Sna1 (n=1), Sna3 (n =7), Sna8 (n=24), Sna11 (n=11), Sna21 (n=22), Sna23 (n=1), Sna35 (n=12), and Sna36 (n= 1). Pcadapt detected 885 SNPs associated with the first principal component

(separating the Seneca strain from others) and 317 SNPs associated with the second principal component (separating the Parry Sound strain from others).

The strongest signals of divergence between the Seneca strain and others were located on chromosomes Sna1, Sna11, Sna14, and Sna19. The strongest signal of divergence identified overall was located on chromosome Sna8 (BAYPASS -log10p= 9.73), was primarily driven by divergence of the Parry Sound strain from all other strains, and this locus overlapped a region showing an excess of R_{HYB} in wild born F2 hybrids and a region with a significantly elevated frequency of Seneca origin haplotypes in F2 hybrids with Marquette/Lake Superior ancestry.

GENE SET ENRICHMENT ANALYSIS

For our GSEA conducted on outlier SNPs within regions associated with fitness in the wild, the most highly significant GO terms were swimming behavior ($p < 0.001$; 2 of 72 genes detected, 0.04 genes expected; Table 4.2) and negative regulation of vascular wound healing ($p = 0.002$, 1 of 4 genes detected, 0 genes expected; Table 4.2). The two genes associated with swimming behavior were unconventional myosin-IXAb (MYO9AB) and carbonic anhydrase-related protein 10-like (CA10-L). CA10-L is located on chromosome Sna3 within a group of overlapping windows that exhibited significant excesses of R_{HYB} and a significant excess of Seneca ancestry (Figure 4.5A, Table 4.1). MYO9AB is located on chromosome Sna21 in a region that exhibited a deficit of R_{HYB}. The gene associated with negative regulation of vascular wound healing was chemokine-like protein TAFA-5 (TAFA5). TAFA5 is located within the region of chromosome Sna8 that exhibited an excess of Seneca ancestry when the primary Great Lakes ancestry component was derived from Lake Superior origin strains.

DISCUSSION

Our first prediction was that we would identify multiple genomic regions with an excess of Seneca Lake origin haplotypes in wild born F2 hybrids relative to the null expectation that Seneca and Great Lakes ancestry would be randomly distributed across hybrid genomes. Results from permutation tests supported this prediction. Specifically, we initially identified a significant excess of Seneca origin haplotypes in F2 hybrids on 6 chromosomes, while an excess or Great Lakes origin haplotypes was only observed on chromosome Sna35. Interestingly, we found that Seneca Lake origin alleles were favored by selection on different chromosomes depending on individual genetic backgrounds. Wild born F2 individuals with ancestry from a Lake Superior origins strain, primarily Marquette and also Seneca Lake exhibited a significant excess of Seneca origin haplotypes in two regions on chromosomes Sna8 and Sna19 (Table 4.1). This suggests that Seneca origin alleles are favored by selection relative to Marquette origin alleles at these two loci. The gene TAFA5 is located within the fitness associated region of chromosome Sna8, is in close proximity to outlier loci between strains, and is known to be associated with regulation of vascular wound healing. Wild born F2 individuals with ancestry from a Lake Michigan origin strain, primarily Lewis Lake, only exhibited an excess of Seneca origin haplotypes on Sna11 (Table 4.1; Figure 4.7A). These results suggest that Seneca Lake origin alleles are favored by selection relative to Lewis Lake origin alleles in this region of Sna11.

Results from the gene set enrichment were highly supportive of our prediction that loci with signals of adaptive divergence between hatchery strains that were also located within local ancestry outlier regions would be in close proximity to genes associated with biological processes related to the ability to avoid or survive Sea Lamprey parasitism

(predictions 3; Table 4.2). Numerous F_{ST} outlier loci between strains were identified within chromosome regions associated with variation in fitness in the contemporary Lake Huron environment, which is consistent with predictions 2 and 3. The high level of significance found for the GO (gene ontology; e.g., function) term for regulation of vascular wound healing was particularly striking given that Sea Lamprey are likely the primary source of vascular wounds in Great Lakes Lake Trout (Sitar et al., 1999). The GO term with the highest significance was for swimming behavior (Table 4.2). It is possible that the two genes associated with this GO term (MYO9AB and CA10-L) are also under selection due to predation by Sea Lamprey on adult Lake Trout or predation on juveniles by other species, such as alewives.

We expected to find that Great Lakes origin alleles would be favored by selection in wild hybrids at a small number of loci. Our initial permutation tests identified a region on chromosome Sna35 in which Great Lakes origin alleles were favored by selection (Figure 4.5B), which was consistent with this prediction. Additionally, within F2 hybrid individuals with Marquette and Seneca Lake strain ancestry, we also identified an excess of Great Lakes origin haplotypes near the centromere of Sna1. Another region of Sna1 was previously found to be associated with ecomorphological variation in Lake Trout and contains a putative chromosomal inversion (Smith et al. 2021 – in prep; see Chapter 4); however, it does not appear that these regions overlap. This suggests some Great Lakes origin alleles are favored by selection relative to those originating from Seneca Lake; however, the effect is only apparent at two loci versus the 7 regions where we detected an excess of Seneca haplotypes. This result is similar to what has been observed in hybridized populations of Rainbow Trout (*Oncorhynchus mykiss*) and Westslope Cutthroat

(*Oncorhynchus clarkii*). In this case, previous studies had found that fitness (reproductive success) was negatively correlated with Rainbow Trout ancestry (Muhlfeld et al. 2009). Kovach et al. (2016) found that a larger proportion of loci had elevated frequencies of Westslope Cutthroat origin alleles than Rainbow Trout origin alleles in hybridized populations.

Finally, we predicted that we would potentially identify multiple loci with an excess of R_{HYB} in F2 hybrid individuals due to heterosis (Crow, 1948). We found this to be the case in 4 genomic regions and in 3 of these regions the effect was primarily driven by an elevated frequency of hybrid genotypes in crosses between Seneca Lake and Lewis Lake origin fish (Figure 4.8A). This implies that differences in fitness between the Seneca Lake strain and Lewis Lake strain could be due to elevated frequencies of deleterious recessive alleles in the Lewis Lake strain at certain loci (Kim et al., 2018). We were also surprised to find chromosome-wide deficits of R_{HYB} on Sna21 and Sna36 in this same collection of individuals. This suggests that hybrid genotypes on these two chromosomes might be associated with decreased fitness relative to individuals with homozygous local ancestries (e.g., outbreeding depression or lower fitness of hybrid individuals; Allendorf & Luikart, 2009). This effect was only observed in hybrids with Lake Michigan (primarily Lewis Lake) and Seneca Lake ancestry.

Furthermore, differences in fitness between the Seneca Lake strain and Lake Michigan origin strains (primarily Lewis Lake but also Green Lake), and the Seneca strain and Lake Superior origin strains (primarily Marquette but also Apostle Islands and Isle Royale) appear to be associated with different factors and different sets of loci. Overall, it appears that Seneca origin haplotypes are favored by selection (Figure 4.5, Table 4.1). The

vast majority of wild born individuals with Great Lakes ancestry originated from the Marquette or Lewis Lake strains. Within F2 hybrids with Marquette and Seneca ancestry, Seneca origin haplotypes are favored on chromosomes Sna8 and Sna19. The signal of selection on Sna8 appears that it may be associated with elevated hybrid fitness given the excess of R_{HYB} at this locus (Figure 4.7A, Figure 4.5B).

Marquette/Lake Superior origin haplotypes are favored by selection on Sna1 and Sna35, but these loci were only detected when backcrosses to Seneca parents were included in the analysis. It is possible that Marquette origin alleles are only favored at these loci when the majority of the genetic background originates from Seneca Lake. Another region near a telomere of Sna1 also displayed a significant excess of R_{HYB} within these individuals; however, these two regions do not overlap. An excess of RHYB was observed on multiple chromosomes in Lewis Lake – Seneca hybrids (n=4), indicating that hybrid genotypes are favored to homozygous Lewis Lake origin haplotypes at numerous loci. Additionally, Seneca origin haplotypes are at significantly elevated frequency on Sna11 within these individuals and Great Lakes origin alleles are not at significantly elevated frequencies at any loci. We conclude that the Seneca strain caries variation on Sna11, and potentially other chromosomes, that provides a significant adaptive advantage relative to the Lewis Lake strain in the contemporary Lake Huron environment.

Some of the strongest signals of selection between strains were located on Sna8 and Sna11. The signal of selection on Sna8, which yielded the highest -log10 p-values observed genome-wide, was primarily associated with divergence of the Parry Sound strain from all other strains. Interestingly, this signal of selection overlaps a region with significantly elevated Seneca ancestry in F2 hybrids, particularly in Marquette-Seneca hybrids, and

significantly elevate R_{HYB}. This could be biologically significant given that the wild Parry Sound population was the only population outside of Lake Superior to avoid extirpation, this region is associated with elevated fitness in the contemporary Great Lakes environment, and the Parry Sound strain carries unique genotypes in this genomic region. It is possible that this unique variation provided a fitness advantage in Parry Sound Lake Trout, which could have allowed them to avoid extirpation; however, this is speculative. The overlap from these two tests could also be a coincidence. For instance, one of the strongest signals of selection separating the Seneca strain from other strains was also located on Sna8 (~20 Mb from the signal of selection associated with Parry Sound divergence) and it is possible that selection is favoring Seneca origin alleles at this locus rather than at the Parry Sound outlier region.

Our results support the hypothesis that elevated contributions of the Seneca Lake strain to wild recruitment are at least partially due to alleles that contribute to increased fitness. The methods employed here are typically only accessible for species with an available linkage map and ideally a linkage map and physical genome assembly. However, given these resources (Smith et al., 2020; Smith et al., 2021, respectively), these sorts of analyses lay the groundwork for the identification of genetic factors that underly differences in fitness between ancestral populations that have been intentionally or unintentionally introduced into novel environments (Leitwein et al., 2020).

Results collectively indicate that multiple genetic factors influence relative fitness among Lake Trout hatchery strains that are actively being used to restore native populations in Lake Huron and other Great Lakes. For instance, within Lewis Lake – Seneca Lake hybrids, we found evidence for elevated fitness of Seneca origin alleles in one genomic

region, heterosis associated with genotypes in 3 regions, and evidence for reduced fitness of hybrid genotypes across two chromosomes. This represents one of few studies to use genomic resources to link adaptive differences between introduced source populations with signals of elevated fitness in reintroduced populations. Results indicate that local remnant populations might not be best suited to recolonize habitats in cases where extirpation was ultimately caused by the emergence of novel selective pressures (i.e., pressures from non-native predators). However, our results also indicate that local remnant populations used for restoration might carry adaptive genetic diversity despite depressed recruitment over the course of reintroduction. Overall, our results highlight the considerable nuance associated with variation in reproductive contributions between source populations during population reintroduction and recovery.

APPENDIX

Figure 4.1: Principal components analysis (PCA) for all hatchery (colored points) and wild born (transparent grey points) individuals (A). Results from a PCA conducted using only hatchery origin individuals are displayed in panel B. The first principal components are displayed on the y-axes and the second principal components are displayed on the x-axes. The existence of multiple individuals with intermediate scores (A) in the wild suggests the existence of inter-strain hybrids.

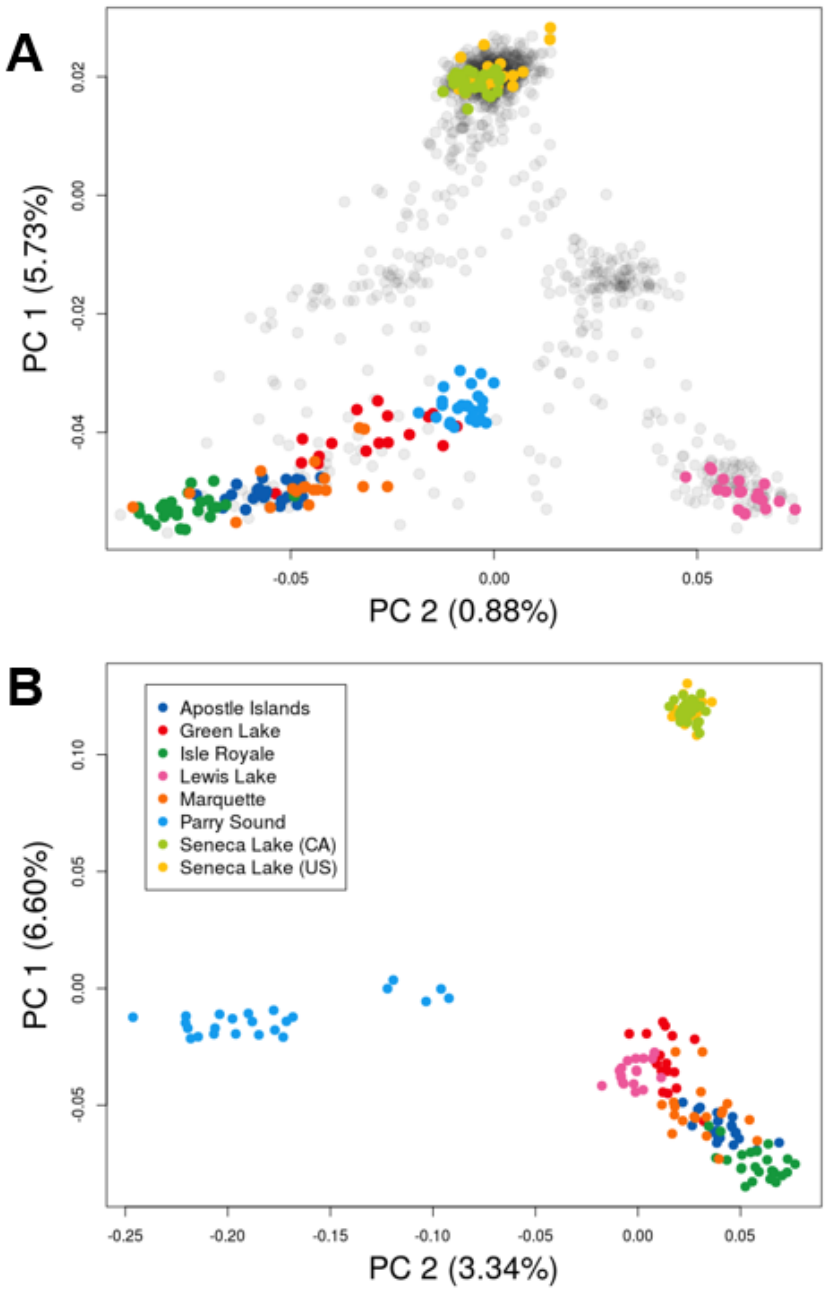

Figure 4.2: Ancestry coefficients reported by the individual based clustering program ADMIXTURE for all K values between 2 and 7 (top to bottom). Cross validation error was minimized at a K value of 3; however, error rates were only slightly lower than for K=2 and K=4. The two Seneca Lake origin hatchery populations (Canadian Seneca, SC; U.S. Seneca, SU) are clearly separated from Great Lakes origin strains at K=2. The Parry Sound Strain is split out at K=3. At K=4, strains with known Lake Michigan ancestry (Lewis Lake, LL; Green Lake, GL) and admixed Lake Superior strains (Marquette, MQ; and Apostle Islands, AI) are separated from the Isle Royale strain, which is believed to be entirely of Lake Superior origin.

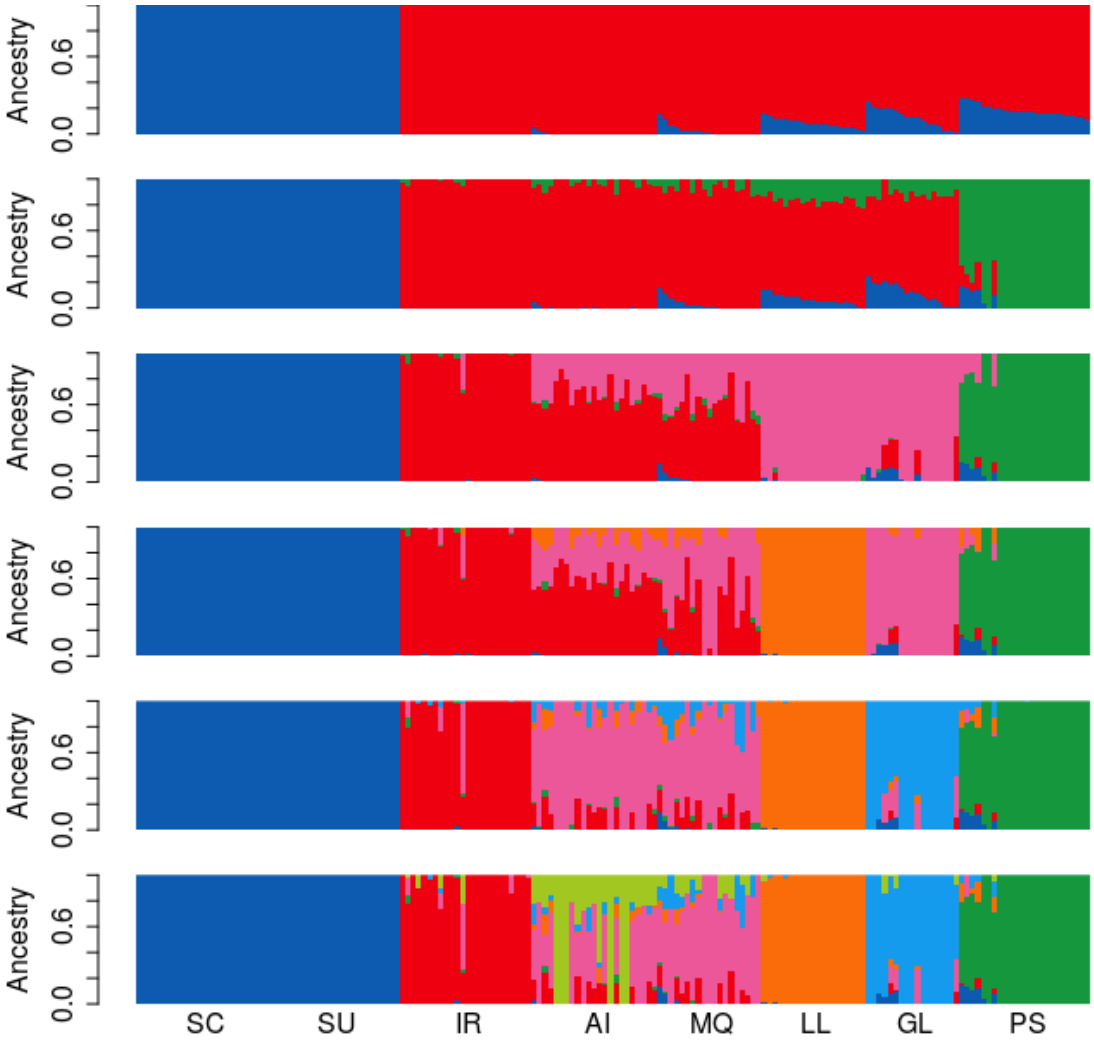

Figure 4.3: The figure below displays example local ancestry inference results for 3 individuals (2 stacked haplotypes per individual) identified as F1 hybrids (F1), F2 intercrosses (F2 (IC)), F2 backcrosses to Seneca Lake origin strains (F2(SL)), and F2 backcrosses to Great Lakes origin strains (F2(GL)). Red blocks correspond to haplotype blocks originating from Seneca Lake, while pale yellow blocks correspond to haplotypes originating from the Great Lakes.

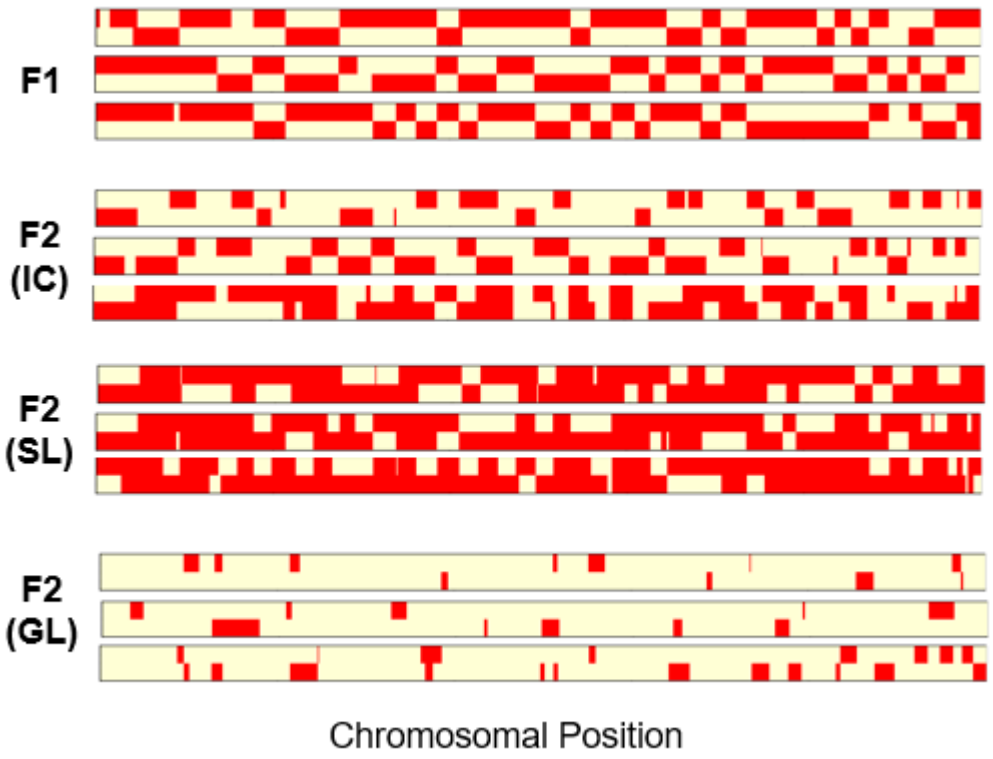

Figure 4.4: Panels A, B, C, D, E, and F display the proportion of Seneca (blue fill) and Great Lakes (red fill) origin alleles within wild individuals identified as pure-bred Seneca (A), purebred Great Lakes (B), F1 hybrids (C), F2 intercrosses (D), F2 backcrosses to Seneca (E), and F2 backcrosses to Great Lakes origin strains (F). All F2 intercrosses were initially identified as F1 hybrids by discriminant analysis of principle components, then reclassified based on the proportion of their genome composed of runs of hybridity. Each individual is represented by a vertical bar and the y-axis displays the proportion of alleles derived from Seneca Lake versus Great Lakes origin populations based on local ancestry inference.

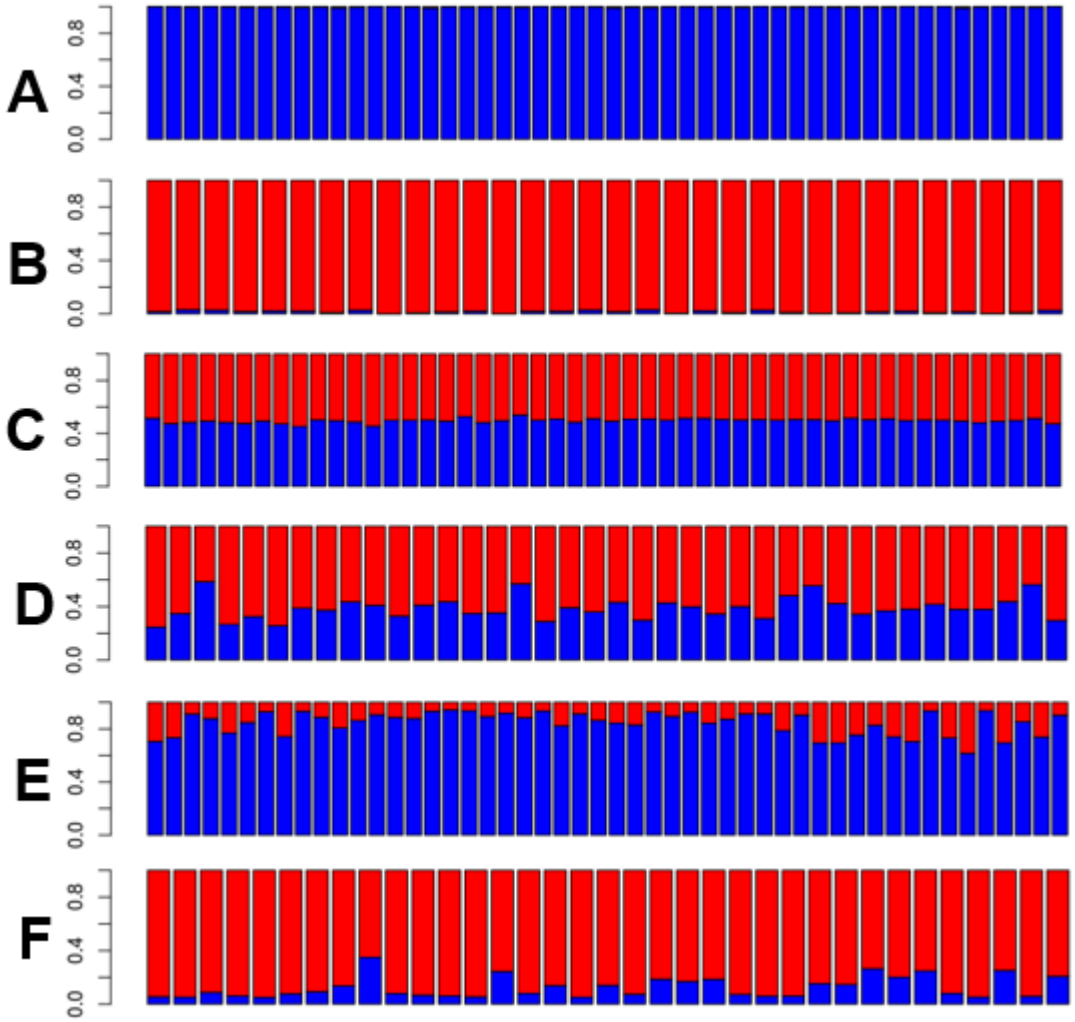

Figure 4.5: Manhattan plots displaying the -log10 p-values for tests for an excess of Seneca Lake origin haplotypes (Panels A and C) and tests for an excess of Great Lakes origin haplotypes (Panel B). P-values were calculated using a permutation procedure that randomized the order of ancestry blocks along haplotypes for each F2 backcross and intercross. The test in panel A was conducting using only individuals that were identified as backcrosses to Great Lakes origin strains in order to maximize statistical power. Panel B displays a test for locus specific excess of Great Lakes origin haplotypes. Panel C displays results from a test that was equivalent to the one displayed in panel A; however, all backcross and intercross individuals were used in the test rather than just backcrosses to Great Lakes strains. Our threshold for significance ($p < 0.01$) is shown with a dashed red line. Each point corresponds with the -log10(p-value) for a single SNP and chromosomes are demarcated using alternating orange and blue points in order to improve readability.

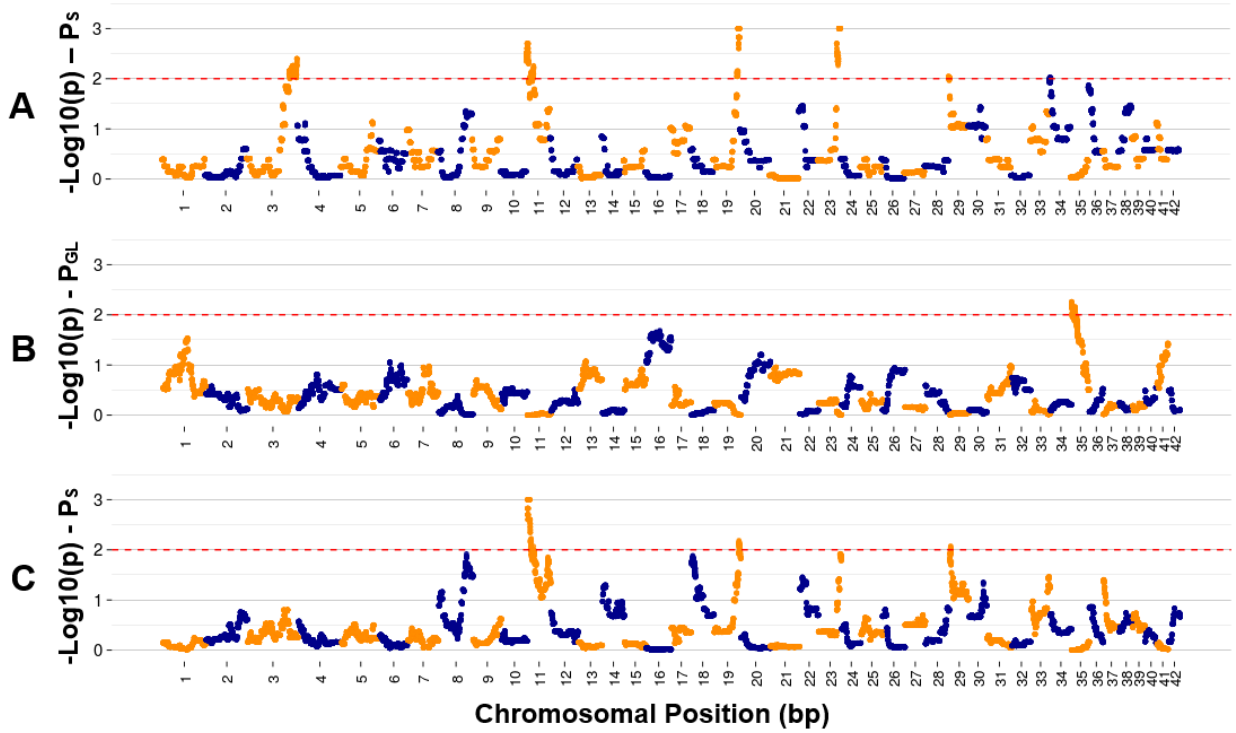

Figure 4.6; Manhattan plots displaying the -log10 p-values for tests for deficits (A) and excesses (B) of runs of hybridity at certain loci within the genomes of all high confidence F2 hybrid individuals. Each point corresponds with the -log10(p-value) for a single SNP. Chromosomes are demarcated using alternating orange and blue points. The significance threshold ($p < 0.01$) is demarcated with a dashed red line. Lake Trout chromosomes Sna1-Sna42 are listed in sequential order on the x-axis.

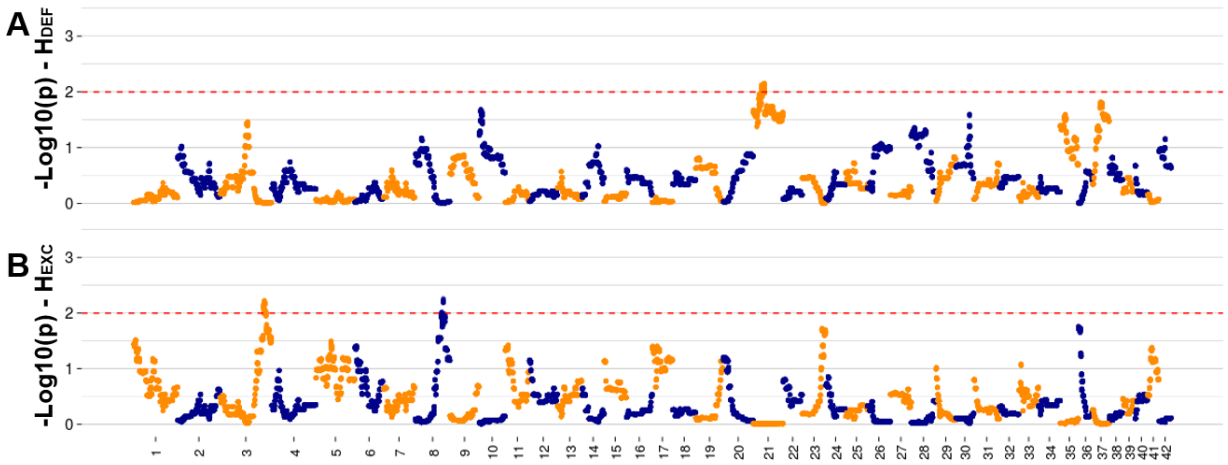

Figure 4.7: Manhattan plots below display the negative and positive log10 p-values for tests for an excess or deficit of Seneca origin haplotypes when the Great Lakes ancestry component originates from Lake Michigan (A; primarily Lewis Lake, but also Green Lake) versus Lake Superior (B; primarily Marquette, but also Isle Royale and Apostle Islands) origin strains. In both plots, negative values (log10 p-values) on the y-axis correspond to results from a test for a deficit of Seneca origin alleles relative to null expectations. Positive values (-log10 p-values) correspond to tests for an excess of Seneca origin alleles relative to null expectations.

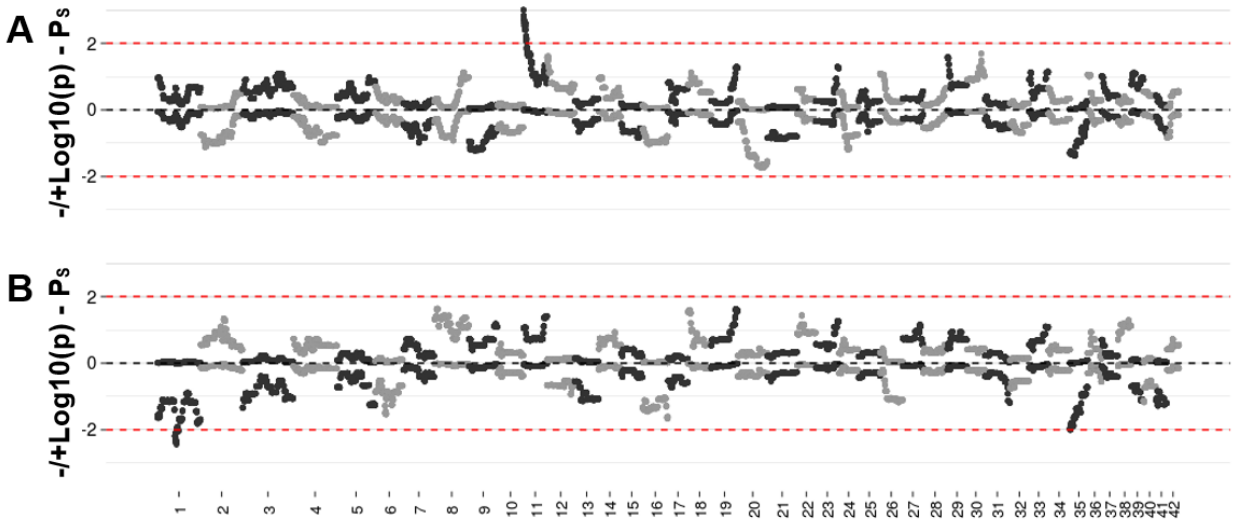

Figure 4.8: Manhattan plots below display the negative and positive log10 p-values for tests for an excess or deficit of runs of hybridity when the Great Lakes ancestry component originates from Lake Michigan (A; primarily Lewis Lake, but also Green Lake) versus Lake Superior (B; primarily Marquette, but also Isle Royale and Apostle Islands) origin strains. In both plots, negative values (log10 p-values) on the y-axis correspond to results from a test for a deficit runs of hybridity relative to null expectations. Positive values (-log10 p-values) correspond to tests for an excess of runs of hybridity relative to null expectations.

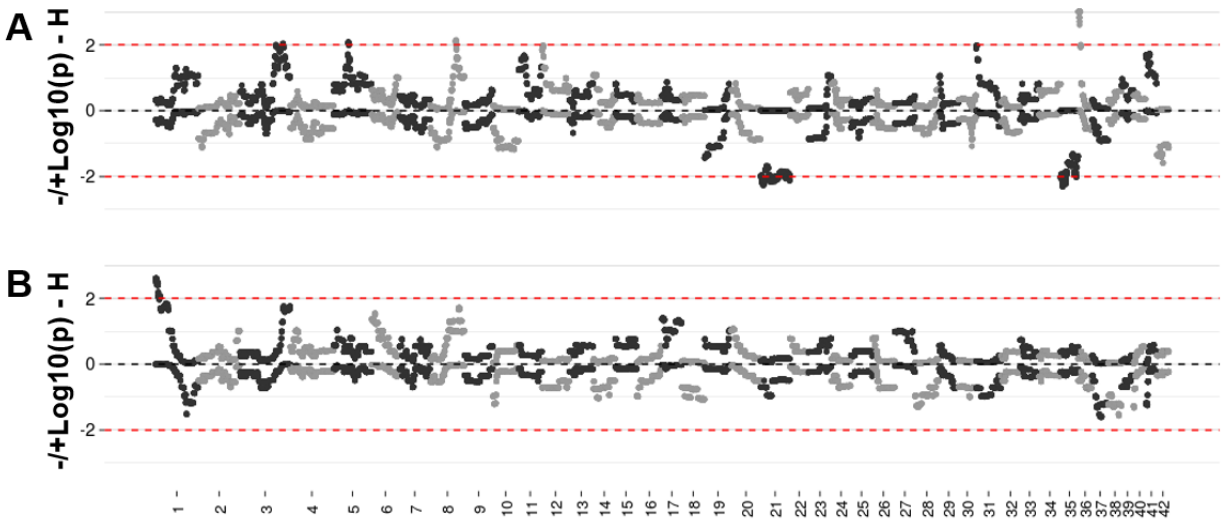

Table 4.1: Coordinates for all regions containing excessive Seneca origin alleles (PSL Excess), excessive Great Lakes origin alleles (PGL Excess), and excesses or deficits of runs of hybridity (RHYB). Chromosome, start and end coordinates (in base pairs), region size (in megabases), the pattern of deviation from null expectations, and the genetic background on which the test was significant are listed. Background corresponds to whether or not the hybridized individuals examined were Lake Superior/Marquette – Seneca hybrids or Lake Michigan/Lewis Lake – Seneca hybrids. Overlapping regions with significant results from multiple tests are highlighted with grey boxes with alternating shades.

Chromosome	Start (bp)	End (bp)	Size (Mb)	Pattern	Background
Sna1	120951	9445264	9.32	R_{HYB} Excess	Superior/MQ
Sna1	34142351	41234775	7.09	P_{GL} Excess	Superior/MQ
Sna3	70265148	84986049	14.72	R_{HYB} Excess	Michigan/LL
Sna3	82789449	98161317	15.37	P_{SL} Excess	All
Sna3	82789449	87823164	5.03	R_{HYB} Excess	All
Sna5	28777276	30034292	1.26	R_{HYB} Excess	Michigan/LL
Sna8	727518	66614250	65.89	P_{SL} Excess	Superior/MQ
Sna8	49680960	51550957	1.87	R_{HYB} Excess	Michigan/LL
Sna8	50829274	54524842	3.70	R_{HYB} Excess	All
Sna11	237125	14743612	14.51	P_{SL} Excess	All
Sna11	237125	7647607	7.41	P_{SL} Excess	Michigan/LL
Sna19	46557722	52115887	5.56	P_{SL} Excess	All
Sna19	48674091	52115887	3.44	P_{SL} Excess	Superior/MQ
Sna21	73533	56736687	56.66	R_{HYB} Deficit	Michigan/LL
Sna21	15502176	21555077	6.05	R_{HYB} Deficit	All
Sna23	37335254	44681299	7.35	P_{SL} Excess	All
Sna29	244320	2531863	2.29	P_{SL} Excess	All
Sna34	607198	3250148	2.64	P_{SL} Excess	All
Sna35	75348	6811709	6.74	P_{GL} Excess	All
Sna35	75348	30967710	30.89	R_{HYB} Deficit	Michigan/LL
Sna35	216671	1614225	1.40	P_{GL} Excess	Superior/MQ
Sna36	63903	5038774	4.97	R_{HYB} Excess	Michigan/LL

Table 4.2: The top 5 most highly significant gene ontology (GO) terms obtained from SNPs under differential selection between hatchery strains that were also located within regions associated with elevated fitness in the wild based on local ancestry inference. The GO identification number, GO term, number of annotated genes associated with each GO term, number of significant genes, number of expected significant genes, and p-value associated with a Fischer's exact test are listed for each GO term. The 'Annotated' column corresponds to the total number of annotated genes in the genome associated with a given GO term.

GO ID	Term	Annotated	Significant	Expected	p-value
GO:0036269	Swimming Behavior	72	2	0.04	0.00073
GO:0061044	Negative Regulation of Vascular Wound Healing	4	1	0	0.00221
GO:0035702	Monocyte Homeostasis	5	1	0	0.00276
GO:0070100	Negative Regulation of Chemokine-Mediated Sign. Pathway	5	1	0	0.00276
GO:0036015	Response to Interleukin-3	7	1	0	0.00386

Supplemental Material 4.1: This supplementary material contains results from all outlier tests conducted and is available upon written request to the author.

CHAPTER 5: HIGH-THROUGHPUT AND COST-EFFECTIVE GENOTYPING RESOURCES FOR LAKE TROUT (SALVELINUS NAMAYCUSH)

ABSTRACT

We present two novel genotyping panels for Lake Trout (*Salvelinus namaycush*) that utilize genotyping-in-thousands sequencing (GTSeq) and restriction site-associated DNA capture (Rapture). The GTSeq panel targets 300 loci with high minor allele frequencies in Great Lakes Lake Trout populations, along with two sex diagnostic loci, and was developed in collaboration with researchers at the Idaho Department of Fish and Game and Pacific States Marine Fisheries Commission. The Rapture panel makes use of biotinylated RNA baits to enrich Pst1 RAD libraries for 5011 variable RAD loci. These include neutral loci with high minor allele frequencies in Great Lakes populations, loci associated with ecomorphological variation in Lake Superior, and loci showing evidence of adaptive divergence between Lake Trout hatchery strains and wild populations. Sex diagnostic loci were also included on the Rapture panel; however, these loci yielded inconsistent genotypes. One of the two GTSeq sex diagnostic markers was consistently called across individuals and matched visually determined sex in 96.5% of samples. The two panels provide concordant estimates of basic population genetic summary statistics, including measures of genetic diversity and inter-population variance in allele frequency (F_{ST}). Estimates of F_{ST} were correlated among panels and were generally comparable with estimates obtained from previous microsatellite-based studies. Results from discriminant analysis of principal components indicate that all panels are effective for assignment tests in Great Lakes populations. Leave-one out mixture simulations suggest that both panels will be suitable for mixed stock analysis in the Great Lakes; however, the Rapture panel

produced less biased strain contribution estimates for some hatchery strains. This was particularly true when Rapture haplotype genotypes were used for analysis rather than those for SNPs. Both panels were useful for genotyping historical samples; however, the GTSeq panel performed more favorably in this regard and neither panel consistently produced genotypes for samples collected prior to 1969 (>50 yrs old).

INTRODUCTION

The maturation of high-throughput next generation sequencing (NGS) technologies has had profound impacts on the field of conservation genetics (Primmer, 2009; Garner et al., 2016; Hunter et al., 2018). In the case of Atlantic Salmon (Salmo salar) and Pacific salmonids (*Oncorhynchus sp*.), the ability to genotype hundreds to millions of polymorphic loci at consistently decreasing cost (Ali et al., 2016; Campbell et al., 2015; Mardis, 2017) has led to an increase in the precision of conventional population genetic analyses, while also enabling researchers to address qualitatively novel questions related to the genomic basis for variation in fitness and ecologically important traits (Waples, Naish, & Primmer, 2020; Allendorf et al. 2022).

NGS technologies have frequently been operationalized to address routine questions that are relevant to fish and wildlife management and conservation (e.g., Garner et al., 2016). This often requires the development of relatively small genotyping panels (100s to 1000s of loci) that can be easily deployed at large scale for evaluations of population structure, admixture, stock composition, inbreeding, and parentage (Komoroske et al., 2019; Sard et al., 2020; Euclide et al., 2021; May et al., 2021). Initially, many of these panels made use of moderate-throughput quantitative PCR (qPCR) assays that were developed using consensus sequences obtained from restriction site associated DNA (RAD; Baird et

al., 2008; Ali et al. 2016) sequencing or other NGS methods (Amish et al., 2012; Campbell et al., 2012; Roffler et al., 2016; Stetz et al., 2016; Liu et al., 2016). For example, Larson et al. (2014) used RAD sequencing to develop a panel of 96 single nucleotide polymorphisms (SNPs) with elevated variance in allele frequency between Chinook Salmon (Oncorhynchus tshawytscha) populations from western Alaska. This panel increased the number of identifiable reporting groups and assignment accuracy when used for genetic stock identification (GSI). In contrast, many recently designed panels have made use of targeted, sequence-based, genotyping methodologies where SNP genotypes are called directly from sequencing data (Beacham et al., 2020a; Hargrove et al., 2021; Reid et al., 2021).

Genotyping-In-Thousands sequencing (GTSeq; Campbell et al., 2015) and restriction site associated DNA capture (Rapture; Ali et al., 2016) have emerged as two cost-effective options for sequence-based, high-throughput genotyping in non-model species (Meek & Larson, 2019). GTSeq provides a methodology for genotyping hundreds of SNPs in thousands of samples using a multiplex PCR-based enrichment strategy (Campbell et al. 2015). This genotyping method has been widely adopted for parentage-based tagging, genetic stock identification, hybrid identification, and mixed stock assessments for a variety of fish species (Elliot et al., 2018; Barclay et al., 2019; Beacham et al., 2020; Bootsma et al., 2020; Hargrove et al., 2021) and these panels have enabled cost-effective genotyping at range-wide scales. For example, Beacham et al. (2020b) used a GTSeq panel created for coho salmon (*Oncorhynchus kisutch*) to develop a genetic stock identification baseline composted of 57,982 individuals collected from reference populations across the species range. This same panel was used to conduct a parentage-based tagging study on 6,391 fish collected in a mixed stock fishery in British Columbia, Canada and assign fish to their

hatchery of origin. Notably, hatchery assignments were 100% accurate for a subset of 308 fish that were marked with coded-wire tags (Beacham et al., 2020a). GTSeq panels are typically limited to genotyping fewer than 500 loci (Meek & Larson, 2019), which is more than adequate for many population genetic applications.

Although GTSeq was initially conceived as a method for genotyping SNPs (Campbell et al., 2015), recent studies have demonstrated that the usefulness of these datasets is magnified when multiple SNPs within a target region are physically phased (using sequencing reads) into 'microhaplotype' markers. Microhaplotype loci have more than two alleles (i.e., haplotypes) and can increase the accuracy of parentage-based tagging (Baetscher et al. 2019) and genetic stock identification (McKinney et al., 2017).

Rapture combines the high sample multiplexing capacity of RAD sequencing (Baird et al., 2008; Davey et al., 2013; Andrews et al., 2016) with in-solution targeted sequence capture (Jones & Good, 2016) to enrich multiple pooled RAD libraries for loci that are known to be polymorphic based on previous experiments (see Ali et al., 2016; Meek & Larson, 2019; and Stahlke et al., 2021 for additional details). Rapture panels have been developed for multiple species of conservation concern including Rainbow Trout (*Oncorhynchus mykiss*; Ali et al., 2016), Sea Lamprey (*Petromyzon marinus*; Sard et al., 2020), and Westslope Cutthroat (*Oncorhynchus clarkii*; Strait et al., 2021), among others (Euclide et al., 2021; Komoroske et al., 2019; Reid et al., 2021).

Rapture panels can be used to target a much larger number of polymorphic loci than GTSeq panels (Meek & Larson, 2019). These panels are therefore more useful population genetic and quantitative genetic applications that benefit from high marker density including genome-wide association studies (GWAS; Barson et al., 2015), genomic

prediction (Ødegård et al., 2014), local ancestry inference and admixture mapping (Maples et al., 2013; Hoggart et al., 2004), scans for selection (Catchen et al., 2017), estimation of individual inbreeding coefficients (Kardos et al., 2016), or projects requiring multiple marker types such as species-diagnostic, high minor allele frequency (MAF; e.g. for relatedness assessment), and high F_{ST} loci for monitoring adaptive variation or for population assignment (e.g., detecting hatchery strays, introgression or long-distance dispersers). For example, Smith et al. (2020) created a Rapture panel targeting 58,889 Pst1 RAD loci in Lake Trout (*Salvelinus namaycush)* that was later used to generate a high-density linkage map and identify loci associated with variation in body size, condition factor, morphology, and pigmentation. This same panel was also used for characterizing the genetic basis for ecomorphological variation in the species and identifying fitness associated loci using local ancestry inference (see Chapters 3 & 4). Margres et al. (2018) developed a 15,898-locus panel that was used to identify loci associated with transmissible cancer resistance in Tasmanian devils (*Sarcophilus harrisii*). Multiple studies have also demonstrated that Rapture panels can be applied across divergent taxa (Reid et al., 2021; Komoroske et al., 2019) and in systems where divergent taxa actively hybridize (Strait et al., 2021). Similar flexibility was previously recognized for other methods based on hybridization capture (e.g., exon capture; Cosart et al., 2011).

Hybridization-based target enrichment strategies tend to have lower genotyping efficiency than multiplex-PCR based methods (Mamanova et al., 2010). Therefore, a smaller proportion of sequencing reads are expected to align to target regions for Rapture (and other hybridization-capture methods) than for GTSeq (and other multiplex-PCR based methods). For this reason, the operationalization of Rapture panels requires some

knowledge of the relationships between individual read counts, error rates, and the proportion of target loci that are successfully called (see Figure 3 in Reid et al., 2021). Microhaplotype genotypes can also be called using paired-end sequencing data generated using RAD sequencing protocols (RAD haplotypes); however, this is a relatively new development (Rochette et al., 2019). It is therefore feasible that Rapture data could be used to call microhaplotype genotypes at hundreds or thousands of target loci at low cost, but this has not been empirically demonstrated to our knowledge.

The primary goal of this study was to develop a Rapture genotyping panel for monitoring and research on native and invasive Lake Trout populations. Lake Trout have experienced severe declines in diversity, distribution, and abundance in their native range (Hansen 1999). Ready-to-use genotyping resources would be valuable for addressing a number of questions related to Lake Trout reintroduction and conservation. For example, multiple hatchery populations have been introduced to the Great Lakes and modern genotyping resources would be useful to enhance accuracy of mixed stock analysis (Scribner et al., 2018), individual assignments to population (or hatchery strains) of origin (Larson et al., 2021), and the identification of inter-strain hybrids (McDermid et al., 2020) which were historically based on microsatellite loci. Additionally, intentional and unintentional Lake Trout introductions in the Western United States have had severe deleterious impacts on native aquatic communities (Tronstad et al., 2015; Ruzycki et al., 2003; Koel et al., 2005) and suppression efforts represent a significant expense for state and federal fisheries management agencies (Syslo et al., 2013). A low-cost, high-throughput, genotyping panel would enable managers to monitor the effective number of breeders (Nb; Waples & Do, 2008); and other population genetic parameters over the

course of suppression. Luikart et al. (2021) recently found that sampling hundreds or thousands of independent loci dramatically improves power to detect population declines using Nb estimates from multiple cohorts. Nb monitoring could also be used to detect early signs of population expansion or contraction in the native range (Tallmon et al., 2012). We further sought to determine the quantity of sequencing data needed per individual in order to produce high quality genotypes with a minimal fraction of missing data and low error rates.

Additionally, we provide a description of a GTSeq panel that was developed in collaboration with researchers at the Idaho Department of Fish and Game (IDFG) and Pacific States Marine Fisheries Commission (PSMFC), which includes a subset of markers targeted by the Rapture panel and a number of microhaplotype loci. We compare estimates of population genetic summary statistics, genotype call rates for samples of varying age, and patterns of population genetic structure between the two panels. We also evaluate the usefulness of these two panels for mixed stock analysis and population assignment. We explored the utility of microhaplotype genotypes called using Rapture data for mixed stock analysis and evaluations of population genetic structure, with the prediction that these data would allow for less biased estimates of strain contributions and more accurate descriptions of population genetic structure.

MATERIALS AND METHODS

GTSEQ AND RAPTURE PANEL DESIGN

The GTSeq and Rapture panels were designed using data generated with the 58,889 locus Rapture panel described by Smith et al. (2020). These included data for individuals from multiple hatchery strains stocked in the Great Lakes, multiple Lake Trout ecotypes

(Muir et al., 2014), individuals from linkage mapping families (Smith et al., 2020), and wild born Lake Trout from Lake Huron (Scribner et al., 2018). Genotyped hatchery populations included the U.S. and Canadian Seneca Lake strains (SLW; CAN), the Parry Sound strain (HPW); the Isle Royale strain (SIW), the Apostle Islands strain (SAW), the Marquette strain (SMD), the Lewis Lake strain (LLW), and the Green Lake strain (GLW). Lean, siscowet, and humper ecotypes from Lake Superior were also genotyped.

The collection of candidate GTSeq loci included 671 RAD loci with high minor allele frequencies (>0.05 in one or more hatchery population), consistently high call rates across samples (> 80% call rate), no evidence for deviations from Hardy-Weinberg proportions in any hatchery population, and allele read ratios near 0.5. Deviations from Hardy-Weinberg proportions were identified using the exact test implemented in VCFtools (Danecek et al., 2011; Wigginton et al., 2005). SNPs were removed if p-values were significant (alpha = 0.05) after Bonferroni correction in one or more hatchery populations. SNPs were also removed if the allelic read ratio (Allele Balance) across all heterozygous samples was less than 0.4 or greater than 0.6. We then selected SNPs between 50 and 130 base-pairs from each Pst1 cut-site in order to provide sufficient flanking sequence for GTSeq primer design. Of the remaining loci, 300 were included because they were found to be informative for differentiating Lake Trout hatchery populations and ecotypes.

Hatchery strain and ecotype informative SNPs were identified by cross referencing the list of remaining SNPs with a list of outlier loci from discriminant analysis of principle components conducted using the R-package (R Core Team, 2020) adegenet (DAPC; Jombart et al. 2010; Jombart et al., 2008). DAPC was performed by grouping samples by ecotype or hatchery strain, identifying the optimal number of principle components and discriminant

functions for separating groups using cross validation with the function xvalDapc, and selecting markers with loadings in the top 5%. Ecotype and hatchery informative markers were then thinned on an increment of 10 megabases (Mb). We also preferentially selected 170 RAD loci containing 2 or more high confidence SNP loci, which could conceivably be phased into microhaplotypes (Baetscher et al., 2019). We then selected random loci from our set of high confidence SNPs in order to fill gaps larger than 5 Mb whenever possible. This process was repeated until no additional gaps larger than 5 Mb could be filled. GTseq primers were designed for 300 of these 671 loci by personnel at the Idaho Department of Fish and Game (see acknowledgements). Two additional sex diagnostic loci were included on the final panel for a total of 302 loci (Supplemental Material 5.1).

For the Rapture panel, we required that loci pass the minor allele frequency, allele balance, call rate, and Hardy-Weinberg filters described above; however, ecotype and hatchery strain informative loci were thinned on an increment of 1 Mb, rather than 10Mb, in order to increase marker density. Additionally, supplemental variable RAD loci were included within 1 Mb of loci with high DAPC loadings in order to allow greater power for identifying signals of selection. Additional loci within regions associated with adaptive divergence between ecotypes and hatchery strains, and variation in fitness between hatchery strains were also included based on results of previous and ongoing research (see Chapters 3 & 4). Three putative sex diagnostic loci were also included. Gaps larger than 1 Mb were filled using the same methodology used to fill gaps for the GTseq panel. A total of 6,377 consensus sequences were submitted to Arbor Biosciences (Ann Arbor, Michigan, U.S.A) for bait design. Consensus sequences were 160 bp in length and were designed to be adjacent to the Pst1 cut-site for each RAD locus; however, they did not include the cut-site

itself. Four overlapping 80 bp baits were designed for each RAD locus by Arbor Biosciences (1.7X tiling density). Baits were included for sex diagnostic loci and loci exhibiting strong signals of selection if they passed Arbor Biosciences' 'moderate' filtering criteria and were less than 25% repeat masked regardless of the number of baits that were retained (115 loci, 415 baits).

Additional Rapture loci associated with adaptive differences between ecotypes and hatcheries and loci putatively associated with fitness in the wild were included if all 4 baits passed the 'moderate' filtering criteria, had fewer than two off-target alignments with melting temperatures greater than 65 ºC, fewer than two off target alignments with melting temperatures between 60 and 65 ºC, and were less than 25% repeat masked (2,028 loci, 8,112 baits). Additional polymorphic loci were included if all 4 baits passed the 'strict' filtering criteria, had fewer than one off-target alignment with a melting temperature greater than 65 ºC, fewer than two off target alignments with melting temperatures between 60 and 65 ºC, and contained 0 repeat masked bases. In total, 19,999 baits were selected for 5,011 variable loci (Supplemental Material 5.2).

LIBRARY PREPARATION AND SEQUENCING

Pst1 BestRAD libraries (Ali et al., 2016) were prepared for 755 Lake Trout samples from hatchery and wild populations from the Great Lakes. These included 24 individuals from the U.S. Seneca Lake strain, 24 individuals from the Canadian Seneca Lake strain, 24 individuals from the Lewis Lake strain, 24 individuals from the Green Lake strain, 24 individuals from the Parry Sound strain, 24 individuals from the Apostle Island strain, and 24 individuals from the Isle Royale strain. Additionally, we genotyped 242 siscowet, lean, and humper Lake Trout collected from Lake Superior. These included 96 siscowet collected

from Isle Royale, Caribou Reef, Whitefish Point, and Stannard Rock; 86 leans collected from Isle Royale, the Apostle Islands, Stannard Rock, and Caribou Reef; and 60 humpers collected from Caribou Reef and Isle Royale. Additionally, we included 95 samples collected from Lake Michigan and Lake Superior during each decade between the 1940s and 1980s. DNA was extracted from scale samples for these individuals using the bead-based extraction protocol described by Ali et al. (2016). Two plates (190 samples) of un-fin clipped, wild born, Lake Trout from Lake Huron were also included (Scribner et al., 2018). Hatchery samples, historical samples, and a subset of 96 wild born fish from Lake Huron were also sent to IDFG for genotyping using the GTSeq panel.

RAD libraries were prepared using the protocol from Ali et al. (2016) with modifications described in Smith et al. (2020). Libraries were quantified using QuantIt Picogreen assays (Life Technologies, Carlsbad, California, U.S.A) run in triplicate before pooling equal amounts of DNA from each library (8 libraries total). Pooled libraries (4 libraries per pool) were enriched for 5,011 variable Pst1 RAD loci using a MyBaits v5 Custom Target Enrichment Kit (Arbor Biosciences, Ann Arbor, Michigan, U.S.A). Baits were allowed to hybridize to targets for 16 hours. Hybridization reactions were performed at 65 ºC and wash reactions were performed at between 66 and 68 ºC in a benchtop dry bath. Target enriched pools were each amplified for 12 cycles using a KAPA Library Amplification Kit for Illumina (KAPA Biosciences, Wilmington, Massachusetts, U.S.A) using manufacturer recommended PCR conditions. Amplified DNA was purified twice using 0.9:1 (bead:DNA ratio) Ampure XP clean-ups (Beckman-Coulter, Brea, California, U.S.A) and eluted in low-EDTA TE buffer. Each pool was quantified using QuantIt Picogreen Assays run in triplicate (Life Technologies, Carlsbad, California, U.S.A) before combining pools. The

pooled library was sequenced in 2 Illumina HiSeqX lanes (Illumina, San Diego, California, U.S.A) using 2X150 paired-end reads and a 5% phiX spike-in. Sequencing was carried out by the Novogene Corporation (Beijing, China).

BIOINFORMATICS

For the Rapture dataset, paired-end reads were re-oriented such that Pst1 cut-sites and sample specific barcodes were located at the beginning of the first read (see Smith et al., 2020 for details). Reads were then demultiplexed using process_radtags implemented in Stacks v2 (Rochette et al., 2019). At this point, PCR duplicates were removed using clone_filter and sequencing adapter contamination was removed using Trimmomatic v0.32 (Bolger et al., 2014). Trimmomatic was also used to trim reads whenever the mean base quality across a sliding window of 4 bp dropped below Q15. Reads were then mapped to the Lake Trout genome (Smith et al., 2021; Genbank accession: GCA_016432855.1) using bwa mem (Li, 2013) with standard settings. Resulting bam files were sorted using samtools sort (Li et al., 2009), and filtered to remove secondary, supplementary, improperly paired, and low-quality alignments (MQ<10) using samtools view. At this point, the number of retained mapped reads was calculated for each sample using samtools flagstat. Genotypes were called using gStacks using the maruki_low genotyping model (Maruki & Lynch, 2017) and genotypes were exported to VCF format using the populations program from Stacks. Genotypes were set to missing if they were called with fewer than 5 reads or if the likelihoods for the two most likely genotypes were different by less than two orders of magnitude (GQ<20). We then required a minimum of 3 observations of the minor allele and removed individuals and loci with greater than 70% missing data. All filtering was done using VCFtools (Danecek et al., 2011). We then produced a separate dataset containing RAD

haplotype (e.g., microhaplotype) genotypes using the populations program with arguments set to --min-samples-overall 0.7, --filter-haplotype-wise, and --min-mac 3. Loci were retained if more than one allele was observed across hatchery populations and the least frequent allele was observed more than twice in at least one hatchery population. The locations of recovered Rapture and GTSeq loci were visualized using the R package quantsmooth (Oosting et al., 2005; Figure 5.1).

POPULATION STRUCTURE AND DIVERSITY

We characterized genetic diversity within and among Lake Trout hatchery strains using the R-package diveRsity (Keenan et al., 2013). We computed mean observed and expected heterozygosity, mean F_{IS}, the standard deviation of F_{IS}, the number of polymorphic SNP markers, the average number of genotyped individuals per locus, and allelic richness for each hatchery population. For both marker sets, we tested for deviations from Hardy-Weinberg proportions for each locus in each hatchery population using Chi-Square tests as implemented in diveRsity. Significance was assessed at a Bonferroni corrected p-value of 0.05 and the number of loci showing deviations from Hardy-Weinberg proportions was determined for each panel. We also calculated pairwise F_{ST} between all pairs of hatchery populations using the R function fastDivPart and repeated this process for the GTSeq, Rapture SNPs, and Rapture haplotype datasets. Pairwise F_{ST} estimates were compared with those obtained between hatchery populations in a recent microsatellite study (Scribner et al., 2018).

Additionally, we used DAPC to characterize the extent to which the three marker sets (GTSeq loci, Rapture SNPs, and Rapture haplotypes) were useful for assessing population structure and individual assignments to strain of origin. DAPC was conducted

using 20 principal components and 6 discriminant functions. We conducted a second run of DAPC for each marker panel using settings obtained from 30-fold cross validation with xvalDapc that minimized out-of-bag misclassification rates. Individuals were visualized along the first two linear discriminant functions and population assignment results were converted to confusion matrices for each marker panel (Hendricks et al., 2018).

ERROR RATE ESTIMATION AND CALL RATES

We used the R-package Whoa (https://github.com/eriqande/whoa; see Hendricks et al. 2018 and Stahlke et al., 2021) to estimate genotype error rates and characterize the relationship between read depth and error rate for Rapture SNP genotypes. The filtered VCF file was split by population and genotypes for each of the 8 hatchery populations were used as input for the function.

Based on estimated dataset-wide miss-call rates, we determined the number of SNPs that were called with greater than 5X, 7X, 10X, 15X, and 20X coverage for each individual. We then fit a loess model describing the number of genotyped loci at greater than some level of coverage as a function of mapped read count using to loess function from the R-package stats (R Core Team, 2020). We estimated the absolute minimum number of mapped reads needed per individual as the point at which more than 80% of loci would be genotyped at 5X or greater coverage according to the loess model. We estimated the optimal number of mapped reads as the point at which greater than 95% of loci were predicted to be genotyped at greater the 7X coverage according to the loess model.

Additionally, the number and proportion of loci with called genotypes was determined for each of the samples collected between the 1940s and 1980s. These values

were compared using boxplots for the Rapture SNPs and GTSeq datasets in order to gain insights about the relationship between call rate and sample age.

SEX DIAGNOSTIC LOCI

We determined the reliability of sex diagnostic loci (Smith et al., 2020) by genotyping a subset of 144 hatchery origin fish of known sex using both genotyping panels. We determined the number of individuals for which sex was correctly determined based on these markers for each hatchery population.

MIXED STOCK ANALYSIS

We determined the suitability of the three marker sets (GTSeq, Rapture SNP, and Rapture haplotype) and our reference panel for mixed stock analysis in the Great Lakes by carrying out leave-one-out simulations (Anderson et al., 2008) conducted using the R-package rubias (https://github.com/eriqande/rubias). Based on results from DAPC, we assigned individuals to 3 reporting groups. These included Lake Superior and Lake Michigan derived strains, Lake Huron derived strains, and Seneca Lake derived strains, with individual hatchery strains nested within these reporting groups. For each marker set, we preformed 1000 mixture simulations with 200 individuals in each mixture using the function assess_reference_loo. The accuracy of mixed stock assessment was determined by preforming a simple linear regression between observed and expected mixture proportions across all simulations for each hatchery population while keeping the intercept fixed at 0. Hatchery populations for which the slope of this relationship was greater than 1.05 or less than 0.95 were considered to be systematically over or under represented across simulated mixtures.

RESULTS

BIOINFORMATICS

We generated 1.908 billion reads (954.4 million pairs) across all individuals that were genotyped using the Rapture panel. Of these, 42% of reads survived after demultiplexing, trimming, duplicate removal and mapping. We obtained a mean of 1.06 million mapped reads per individual across all samples and a mean of 1.19 million reads per individual after excluding historical samples. Genotypes were called at 2,787,409 loci. We were left with 9,560 high confidence SNPs located on 5,006 RAD loci after filtering on genotype quality, genotype read depth, minor allele counts, and missingness (Table 5.1). RAD haplotype genotypes were called for 8,979 loci. Of these, 4,514 loci passed filtering criteria and were retained for additional analysis. Of these, 2,765 loci had 3 or more alleles across hatchery populations, 931 had 4 or more alleles, 307 had 5 or more alleles, and 36 had 6 or more alleles. The locations of Rapture and GTseq markers across the genome are shown in Figure 5.1.

POPULATION STRUCTURE AND DIVERSITY

Between 274 and 299 GTSeq SNP loci were polymorphic in hatchery populations and mean allelic richness varied between 2.022 and 2.155. Between 4 and 11 SNPs included on this panel showed significant deviations from Hardy-Weinberg proportions. Between 4,515 and 5,904 Rapture SNPs were polymorphic in each hatchery population, mean allelic richness varied from 1.569 to 1.742, and between 0 and 7 markers exhibited significant deviations from Hardy-Weinberg expectations. Between 3,180 and 3,904 Rapture haplotype markers were polymorphic in each hatchery population, mean allelic

richness varied between 1.746 and 1.977, and between 1 and 5 markers showed evidence of deviating from Hardy-Weinberg expectations in hatchery populations (Table 5.1).

Levels of misclassification were low overall (typically less than 10%). Mis-assignments typically involve the populations LLW, GLW, SMD, SAW, and SIW which are all of Lake Superior and Lake Michigan origin. Patterns of population genetic structure indicated by DAPC were similar across marker sets. The best separation between Lake Michigan and Lake Superior origin hatchery populations was obtained using the Rapture haplotypes dataset (based the first 2 discriminant functions; Figure 5.2). This marker set also produced patterns of population genetic structure that were highly similar to those observed using a much larger collection of loci (58,889 loci, see Chapter 4). In all cases, the Seneca Lake and Parry Sound strains were easily differentiated from strains originating from Lake Michigan and Lake Superior (Figure 5.2). The first two discriminant functions obtained using the Rapture haplotypes dataset suggest that this marker set has the highest power to differentiate hatchery populations originating from Lake Superior and Lake Michigan (SIW, SAW, SMD, GLW, and LLW). These 5 populations form a single profuse cluster in the DAPC conducted using GTSeq genotypes. Four of these populations were difficult to differentiate based on the first two discriminant functions obtained using the Rapture SNPs dataset, where the LLW population forms a distinct cluster in this case.

Assignment results from DAPC, which evaluate all 6 discriminant functions rather than just the first 2, indicate that all three panels will be effective for assigning individuals to their hatchery strain of origin, with between 88% and 100% of individuals being correctly assigned (Table 5.2). DAPC results indicate that Parry Sound and Seneca Lake should be considered as distinct reporting groups. DAPC was able to assign individuals to

these populations with 100% accuracy for all marker sets and DAPC parameters examined. Lake Michigan and Lake Superior origin hatchery populations could likely be considered as a separate reporting group. The most frequent mis-assignments within this group were between SMD, LLW, GLW, and SAW. A small number of Apostle Islands (SAW) and Green Lake (GLW) origin fish were consistently assigned to the Marquette strain (SMD; 5-13%; Table 5.2).

Pairwise F_{ST} estimates obtained using the three marker sets were highly correlated (Figure 5.7). Correlation coefficients (R^2) between the Rapture haplotypes, Rapture SNPs and GTSeq panels ranged from 0.968 to 0.997. Correlation coefficients between the marker sets evaluated here and the F_{ST} estimates from Scribner et al., (2018) ranged from 0.870 to 0.904 (Figure 5.7; Table 5.5). Microsatellite derived FST estimates were most strongly correlated with those from the GTSeq panel (R^2 = 0.904). Additionally, microsatellite derived estimates were lower on average by 0.031, 0.033, and 0.026 than for the Rapture SNPs, Rapture haplotypes, and GTSeq datasets, respectively.

ERROR RATE ESTIMATION AND CALL RATES

The mean dataset-wide error rate across hatchery populations was 0.711% with estimates ranging between 0.024 and 1.59% across populations (Figure 5.3). Requiring more than 5 reads to call a genotype led to small, but notable, decrease in the prevalence of heterozygote miss-call errors (from allelic drop out). Specifically, the mean mis-call rate after requiring a minimum of 7, 10, 12, 15, and 20 reads to call a genotype were estimated to be 0.74%, 0.46%, 0.43%, 0.42%, and 0.39%, respectively. We conclude that a minimum read depth of 5X is suitable for calling genotypes using this panel, although more than 7X coverage is highly preferable.

We found that approximately 446,684 mapped reads are needed per individuals to call genotypes at greater than 5X coverage for more than 80% of loci (Figure 5.4, our minimum threshold). This value increased to 1,047,129 reads if we required 7X coverage to call genotypes at more than 95% of loci. Assuming 42% of reads are ultimately mapped and pass other filters (see Results above), then no fewer than 1,059,594 individual raw reads (529,797 paired end reads) should be generated for each individual. For this genotyping panel, we suggest that 3 or 4 96-well sample plates be multiplexed in each sequencing lane (288-384 samples) in order to obtain high quality genotypes at the majority of targeted loci for all samples and a dataset-wide error rate less than 1%. As many as 5 or 6 plates could likely be multiplexed in a single HiSeq lane in cases where variation in DNA quality across samples is minimal (i.e. recently collected samples).

Sample age had a dramatic effect on the number of called genotypes for both panels (Figure 5.5; Table 5.3). For both panels, there was a strong negative correlation between call rate and sample age. Call rate declines substantially for samples collected prior to 1969 for both panels. The number of genotyped loci steadily declined as a function of time since sample collection for the Rapture panel. We were able to call 70.3%, 49.0%, 38.5%, 24.5%, 24%, 6%, and 1.2% of recovered SNPs on average for samples collected in the 1980s, 1970s, 1960s, 1950s, and 1940s, respectively, using Rapture. Conversely, the GTSeq panel achieved relatively high call rates for samples collected during the 1960s and later (77.9-93.4% on average); however, mean call rates declined to 47.8 and 2.2% for samples collected in the 1950s and 1940s.

SEX DIAGNOSTIC LOCI

Sex diagnostic loci were not consistently recovered by the Rapture panel and only one of the two sex diagnostic loci included on the GTSeq panel yielded consistent genotype calls (Sna_sex_bia). The sex diagnostic marker (Sna_sex_bia) determined the correct sex for 96.52% of individuals across all populations examined. Sex was determined with 100% accuracy for samples from Seneca Lake, Apostle Islands, and Isle Royale hatchery populations. Sex determination accuracy ranged from 91.6% to 95.8% for Green Lake, Lewis Lake, and Marquette hatchery strains, suggesting that this marker is not in perfect linkage with the Lake Trout sex determination locus in some populations.

MIXED STOCK ANALYSIS

Leave one-out mixture simulations indicated that the Rapture haplotypes dataset will produce the least biased estimates of strain contributions (Figure 5.6; Table 5.4). All panels were effective for estimating overall contributions from the three proposed reporting groups. Bias is minimized with the Rapture haplotypes dataset. However, contributions from the Apostle Islands strain were consistently underestimated in simulations conducted using the Rapture SNPs and GTSeq markers (Figure 5.6, Table 5.4). The Rapture haplotypes dataset also demonstrated a tendency to underestimate contributions from the Apostle Islands strain by approximately 7% on average; however, the level of bias was far less severe than for the Rapture SNP and GTSeq marker sets. Contributions from the Marquette and Green Lake Strain were also consistently overestimated based on simulations conducted using the Rapture SNPs and GTSeq datasets. All marker panels produced minimally biased estimates of strain contributions for the Seneca Lake, Parry Sound, Isle Royale, and Lewis Lake populations (observed vs.

expected slope < 1.05 and > 0.95). Our baseline reference populations were composed of 24 samples from each hatchery strain in this case and bias could likely be reduced further for all panels by expanding this sample set.

DISCUSSION

Both panels and all marker sets examined show great promise for the low-cost genotyping of Lake Trout populations in the Great Lakes and elsewhere. The panels will facilitate research and genetic monitoring, including routine mixed stock assessments, genetic stock identification, pedigree reconstruction, and the early detection of population decline using estimates of the effective number of breeders from multiple cohorts (e.g., Luikart et al., 2021). Both panels will allow individual assignment to identify the source population of origin of invasive Lake Trout (e.g., in western North America; Martinez et al. 2009) and potentially illegally-harvested trout from populations closed to fishing (e.g., Primmer et al. 2000).

The leave one out simulation analysis presented here indicated that the Rapture haplotype markers will be the preferable tool for mixed stock assessment in the Great Lakes (Figure 5.6; Table 5.4). We recovered 2,765 RAD haplotype loci with more than 2 alleles (haplotypes) even though this panel was not intentionally designed for genotyping these highly polymorphic loci. A carefully designed Rapture panel could conceivably be used to genotype 500-50,000 RAD haplotype loci with 3 or more alleles (at no additional cost compared to panels with bi-allelic loci).

DAPC analyses demonstrated that these panels can be used to accurately assign individuals to their hatchery strain of origin; however, we observed occasional misassignments between SMD, SAW, and GLW. This is not entirely unexpected given the

history of these three strains. The SMD strain was supplemented with gametes from the SAW and GLW strains in the late 1960s (Krueger et al., 1983; Page et al., 2004). Also, although the GLW strain was originally founded by Lake Trout derived from Green Lake, Wisconsin (which was established using fish from Southern Lake Michigan), a small number of Lake Superior origin fish (1 female and 7 males) were used to found the brood stock evaluated here (Kincaid et al. 1993). Additionally, this strain was cross-bred with fish from SAW in 1971 (Kreuger et al., 1983). The shared ancestry between these strains likely explains occasional mis-assignments between GLW, SAW, and SMD. Similarly, low levels of mis-assignment between the Lewis Lake (LLW) and Green Lake (GLW) strains are likely the result of shared Lake Michigan ancestry (Page et al. 2004).

These panels will also be useful for monitoring the more than 200 invasive populations in the Western U.S. (Martinez et al. 2009), which were largely derived from source populations with similar genetic composition to some of the hatchery populations mentioned here. Lake Trout populations in the state of Montana, for instance, are believed to have originated from Seneca Lake, Lake Michigan, Lake Superior, Lac LaRonge, and the Lewis Lake hatchery strain. Similarly, invasive populations in Utah were founded by individuals from New York and Lakes Michigan, Huron, and Superior (see Crossman et al., 1995 for a review on this topic). Lake Trout are also spreading across river networks and colonizing waterbodies from multiple invasive source populations (Martinez et al. 2009) and these genotyping panels could be extremely valuable for identifying source populations that should be prioritized for suppression. Globally, Lake Trout have been introduced across five continents and many of these introductions have resulted in naturally reproducing populations; however, the origin of introduced populations is often

unclear (Crossman et al., 1995). The highly polymorphic marker panels presented here could be used to determine to origins of these introduced populations at relatively low cost.

Our analysis of sequencing error rates and genotype call rates for the Rapture panel suggests that between 288 and 384 individuals should be sequenced in a single HiSeqX sequencing lane (assuming 350-400 million paired-end reads per lane) to ensure depth of coverage sufficient to minimize error rates. This translates to an estimated sequencing cost of between $3.90 and $5.21 per sample (assuming $1500.00 per sequencing lane). As many as 6 plates (576 samples) could likely be run in a single lane in cases where variation in sample quality is minimal. Assuming each capture reaction costs $243.75 and 288 or 384 individuals are multiplexed in each capture reaction, then the total cost of the hybridization capture reaction is either $0.64 and $0.85 per individual. DNA extraction, DNA quantification, and BestRAD library preparation can be done for less than $5.00 per sample (Stahlke et al., 2021), so the total per sample cost for this Rapture panel should be between $9.54 and $11.06. This is comparable to the cost of running some 16 and 24 locus microsatellite panels (Puckett, 2017), like those that have been used for previous genetic research on Lake Trout (Scribner et al., 2018; Larson et al., 2021; McDermid et al., 2020). However, it is important to note that this cost estimate does not include the added costs associated with bioinformatics and other data analysis.

Meek and Larson (2019) previously suggested the per sample cost for 500-10,000 locus Rapture panels to be approximately $15.00. We feel that this represents a reasonable estimate after incorporating variation in panel size, multiplexing schemes, reagent costs, and sequencing costs across laboratories (including labor for extraction and library preparation). For comparison, the authors estimated the cost of genotyping up to 500 loci

using GTSeq to be around $6.00 per sample. In the case of these two panels, this suggests that the cost per genotype for the Rapture panel is nearly an order of magnitude less than for GTSeq. However, it is important to note that Rapture typically requires higher quality DNA and the library preparation and data analysis process is slightly more difficult compared with GTSeq (Meek & Larson, 2019). Rapture is also limited to targeting Pst1 and Sbf1 RAD loci (however other restriction enzymes could be used), while GTSeq does not have this restriction.

Our inability to consistently recover sex diagnostic loci using Rapture also speaks to the potential difficulty of designing RNA baits for some loci. Alternative strategies exist for the genetic determination of sex in Lake Trout (Smith et al., 2020); however, it would be preferable if sex determination and genotyping could be performed with a single panel. Alternative RNA baits could be designed for putative sex diagnostic loci; however, this would require ordering an entirely new set of baits and success would not be guaranteed. Adding additional target loci to existing GTSeq panels is also potentially costly given that empirical testing of the new panel is recommended (Meek & Larson, 2020).

The relatively consistent genotype call rates from GTSeq for samples collected as early as 1969 (Table 5.3, Figure 5.5), supports the idea that GTSeq is the better of the two methods for dealing with low quality DNA. This makes sense given that locus-specific PCR is relatively sensitive and reliable when using low quality DNA. However, it is also important to note that neither panel could effectively genotype a large proportion of loci in samples that were collected more than 50 years prior to extraction (earlier than 1969). Alternative genotyping methods will likely be preferable for these applications. For instance, conventional targeted sequence capture (Jones & Good, 2016) and whole genome

sequencing (Staats et al., 2013; Parejo et al., 2020) offer two strategies that have been shown to be effective for genotyping ancient and historical samples. SNP genotyping based on qPCR (e.g., Fluidigm arrays with TaqMan assays) is perhaps the most reliable approach for low quality samples, however only 96-192 loci are often genotyped (Campbell and Narum, 2008).

In conclusion, the primers provided here for GTSeq and the bait sequences for Rapture panels will greatly simplify the process of conducting genetic research and monitoring on Lake Trout populations across their native and introduced ranges (Supplemental Material 5.1; Supplemental Material 5.2). The recent publication of multiple genomic resources for Lake Trout, including a genome assembly (Smith et al., 2021) and linkage map (Smith et al., 2020), will also facilitate additional research. Overall, these resources will allow researchers and fisheries managers to gain novel insights pertaining to the ecology and management of this important fish species. This study illustrates how other species could benefit from development of low-cost reliable genotyping approaches (GTSeq or Rapture) for research and monitoring to advance understanding and manage native and invasive populations of conservation concern.

APPENDIX

Figure 5.1: Mapping locations for 300 genotyping-in thousands sequencing (GTSeq) markers (A; red ticks) and 9252 restriction site associated DNA capture (Rapture) markers (B; blue ticks). Each horizontal black line represents a chromosome and the position of the tick is the physical position of a marker. 309 Rapture loci located on unplaced scaffolds (unanchored chromosomal segments) are not shown here. The two sex diagnostic markers from the GTSeq panel are also not shown here because they are also located on unplaced scaffolds (although they are likely located on chromosome 4 based on previous studies – see Chapter 1).

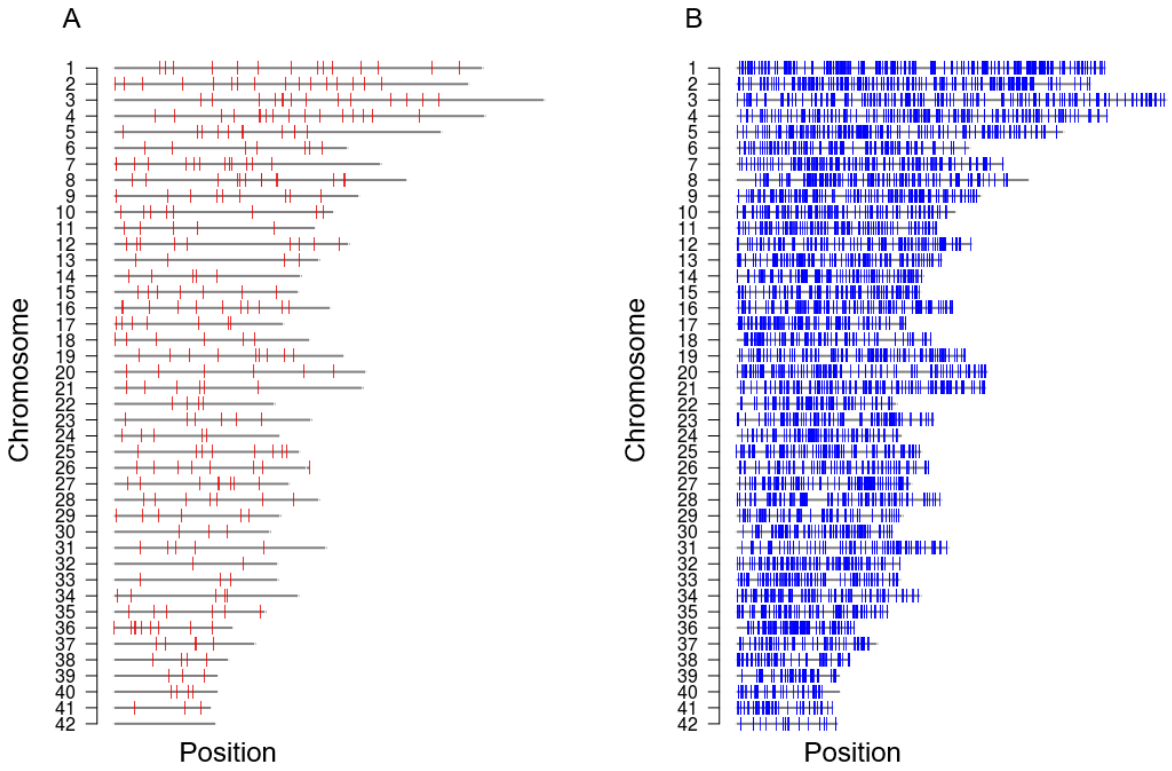

Figure 5.2: Discriminant analysis of principle components (DAPC) plots for the Rapture haplotype (A), Rapture SNP (B), and GTSeq SNP (C) datasets. DAPC was conducted using 20 principal components and k-1 (n = 6) discriminant functions. The first linear discriminant is listed on the x-axis of each plot and the second linear discriminant is listed on the y-axis for each plot. Seneca Lake individuals (SLW; light blue points) and Parry Sound (HPW; red points) are easily identifiable using all genotyping methods. Populations that are known to be admixed (SAW, SMD, and GLW) are located between the Isle Royale strain (SIW; Lake Superior origin; orange points) and Lewis Lake strain (LLW; Lake Michigan origin; dark green points) clusters in the DAPC conducted using Rapture haplotype data (A). This pattern is not readily apparent in the DAPC plots for figure panels B and C.

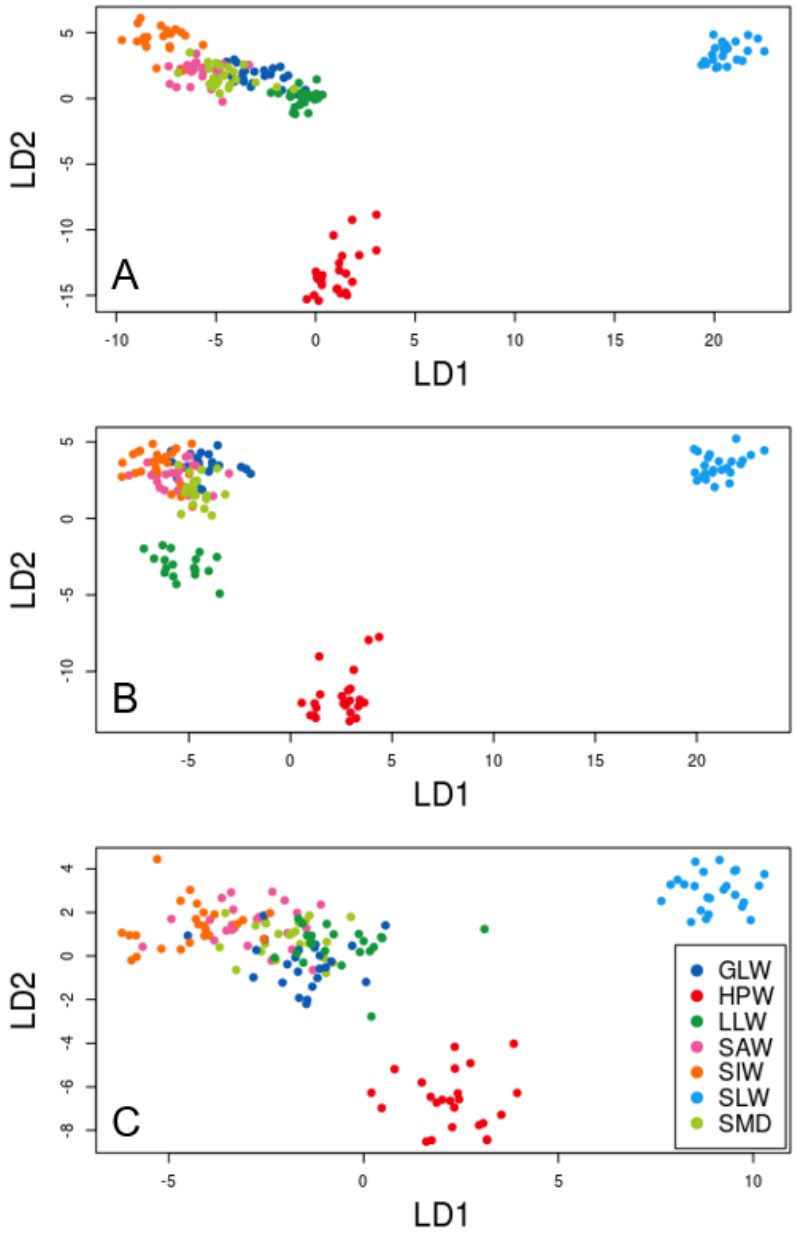

Figure 5.3: Heterozygote miss-call rates (e.g., caused by allelic drop out) estimated using the R-package "whoa." Points correspond to mean posterior estimates of the miss-call rate for the Lewis Lake (LLW), U.S. Seneca Lake (SLW), Marquette (SMD), Apostle Islands (SAW), Isle Royale (SIW), Green Lake (GLW), Parry Sound (HPW), and Canadian Seneca Lake (CAN) hatchery strain collections. 95% credible intervals are also displayed. The dashed black line corresponds to the average posterior mean estimate across populations (0.711%).

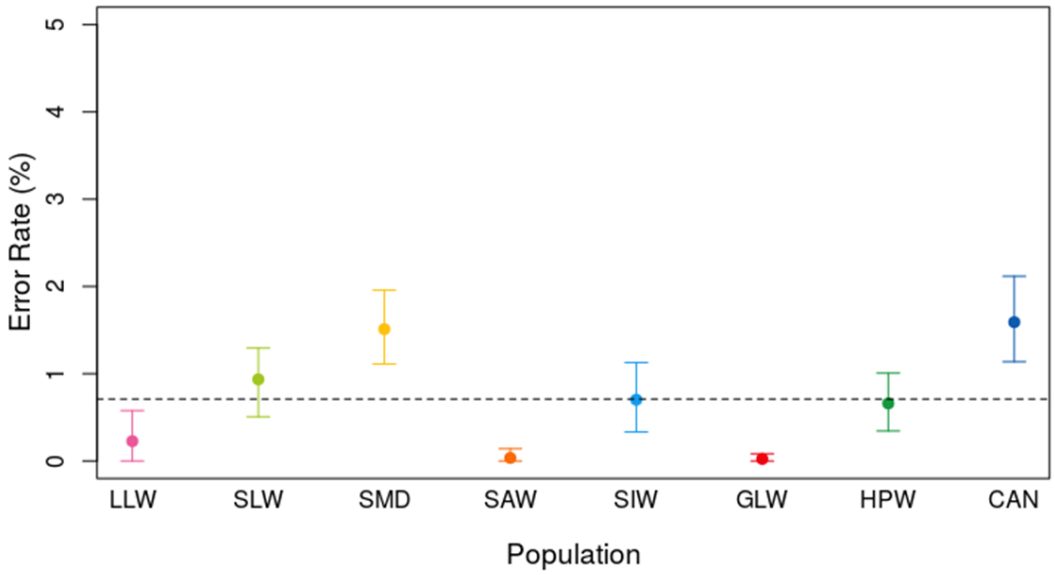

Figure 5.4: Number of mapped reads required to call genotypes for some proportion of consistently recovered single nucleotide polymorphisms (SNPs) with greater than 5X, 7X, 10X, 15X, and 20X coverage (see legend). Horizontal dashed lines correspond to thresholds for recovering 80 and 95% of SNPs. Vertical dotted lines correspond to suggested read count thresholds for calling greater than 80% of loci at 5X coverage and higher or greater than 95% of loci at 7X coverage and greater. A minimum of 446,684 mapped reads (223,342 read pairs) are required to call more than 80% of SNPs at greater than 5X coverage. A minimum of 1,047,129 mapped reads (approximately 523,564 read pairs) are required to call more than 95% of SNPs at greater than 7X coverage.

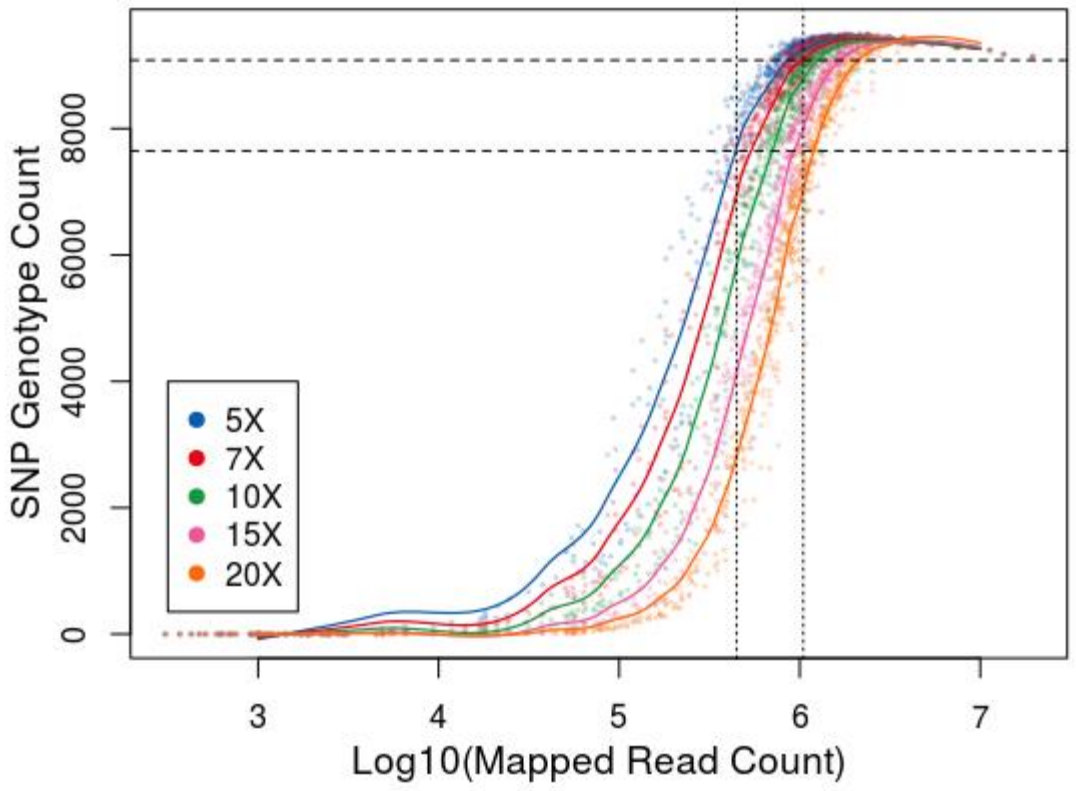

Figure 5.5: Boxplots displaying the number of called SNPs for historical scale samples using the Rapture (orange boxplot, upper panel) and GTSeq panels (blue boxplot, lower panel). Both panels are able to genotype samples collected up to 51 years ago at a fairly large number of SNPs; however, the number of called SNPs drops off substantially in collections from before 1969. SNP call rates steadily decline over time for the Rapture genotyping panel. Call rates are relatively consistent for the GTseq panel for samples collected prior to 1969, then decline substantially. Rapture targets so many SNPs that it provides more genotypes back to 1979.

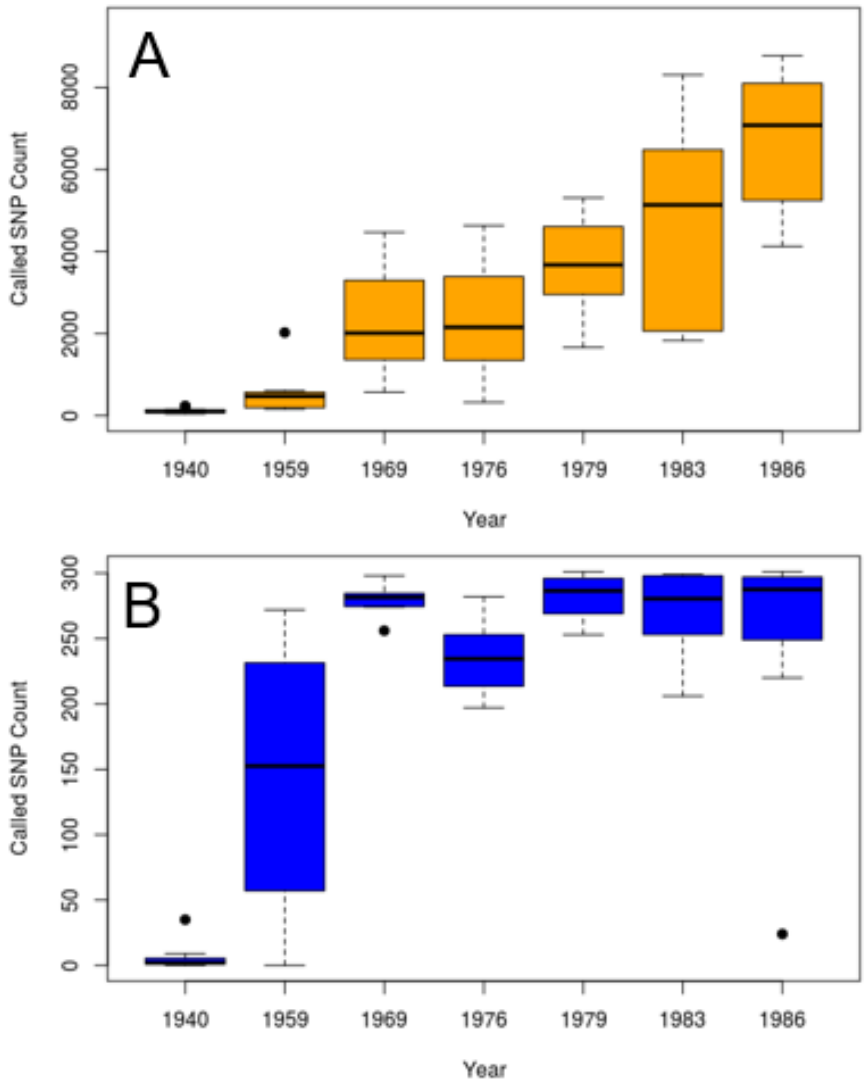

Figure 5.6: Mixture proportions from leave-one out simulations conducted using Rubias (y-axis) versus estimated mixture proportions for the Rapture haplotype dataset (A), the Rapture SNPs dataset, and the GTSeq dataset (C). Different colored points correspond to simulated and estimated mixture proportions for 7 different hatchery populations. The expected relationship (intercept = 0, slope =1) is marked with a black dashed line. Deviations from this line suggest that the marker panel will systematically over or underestimate mixture proportions for a given strain.

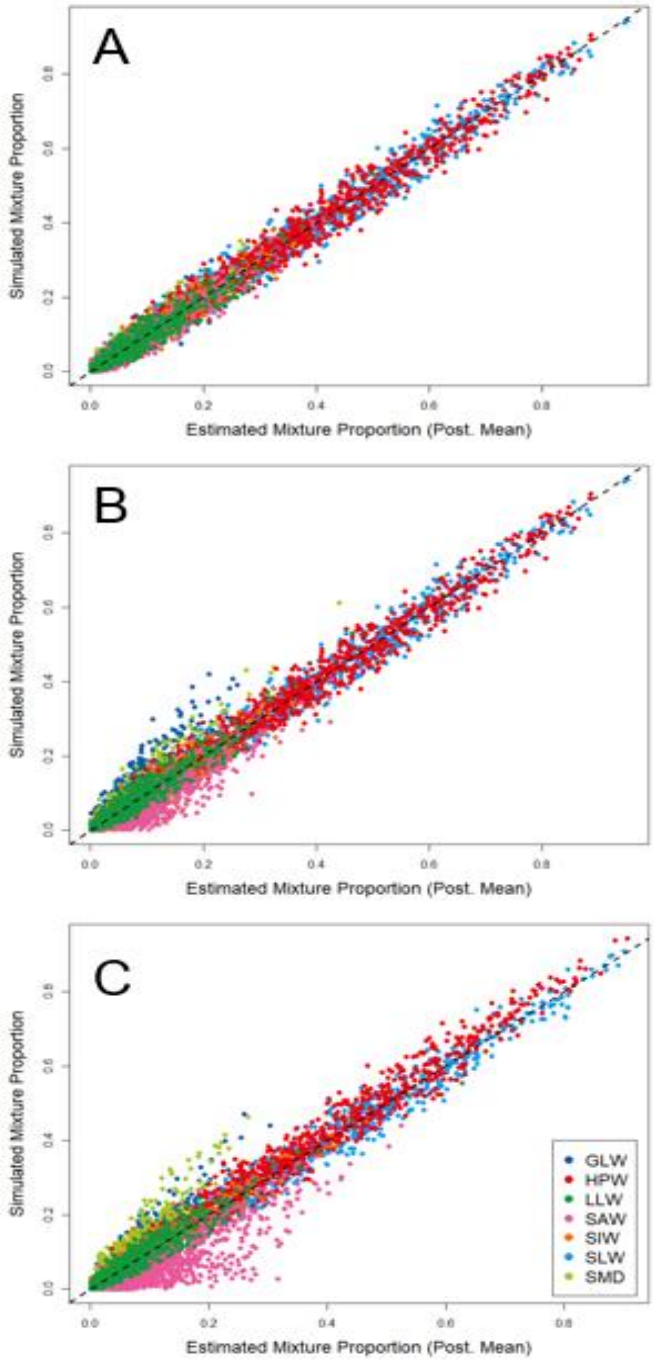

Figure 5.7: Comparison of pairwise FST estimates generated using Rapture haplotypes, Rapture SNPs, GTSeq, and 15 microsatellites (results from Scribner et al., 2018). Blue points correspond to pairwise FST estimates between hatchery populations originating from the proposed Superior/Michigan reporting group. Red points correspond to pairwise estimated between Superior/Michigan populations and the Parry Sound (HPW; Huron reporting group) population. Green points correspond to pairwise estimates between Superior/Michigan populations and the U.S. Seneca Lake strain (SLW; Seneca reporting group). The black point corresponds with the pairwise FST calculated between the Seneca Lake strain and the Parry Sound strain.

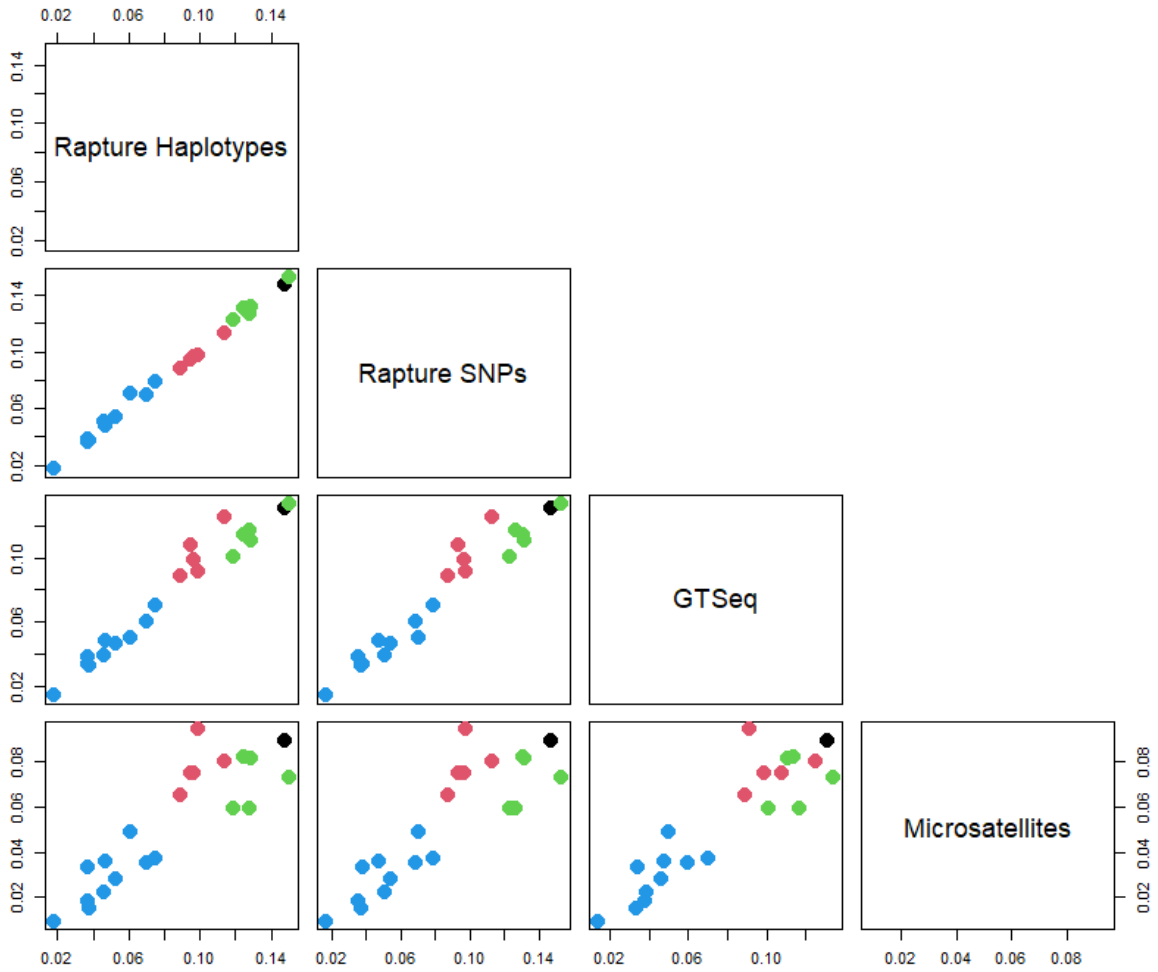

Table 5.1: Population genetic summary statistics for the GTSeq SNPs, Rapture SNPs, and Rapture haplotype datasets for 7 U.S. hatchery populations of Lake Trout. Abbreviations for population names used elsewhere in the manuscript are listed in the Abrv. column. The average number of individuals genotyped per locus (No. Ind.), number of polymorphic loci (No. PM), mean allelic richness, mean observed heterozygosity, mean expected heterozygosity, mean FIS, standard deviation of FIS, and the number of loci with significant deviations from Hardy-Weinberg expectations are also listed for each hatchery population.

Panel	Population	Abrv.	No. Ind.	No. PM	Mean(AR)	Mean(H_O)	Mean(H_E)	Mean(F_{IS})	SD(F_{IS})	No. HWE Dev.
GTSeq	Seneca Lake	SLW	23.98	274	2.022	0.353	0.332	-0.052	0.249	11
	Apostle Islands	SAW	23.94	299	2.155	0.413	0.390	-0.051	0.225	6
	Marquette	SMD	22.58	296	2.154	0.421	0.396	-0.049	0.229	4
	Green Lake	GLW	22.89	294	2.130	0.416	0.379	-0.085	0.220	9
	Lewis Lake	LLW	21.76	295	2.117	0.416	0.381	-0.076	0.246	6
	Isle Royale	SIW	23.87	298	2.163	0.415	0.388	-0.057	0.216	5
	Parry Sound	HPW	23.47	291	2.094	0.388	0.357	-0.073	0.209	7
Rapture SNPs	Seneca Lake	SLW	23.61	4515	1.569	0.177	0.174	-0.015	0.204	0
	Apostle Islands	SAW	23.27	5904	1.742	0.220	0.216	-0.023	0.196	1
	Marquette	SMD	18.25	5664	1.724	0.211	0.209	-0.011	0.222	0
	Green Lake	GLW	19.47	5312	1.669	0.207	0.198	-0.040	0.212	0
	Lewis Lake	LLW	18.54	5336	1.681	0.209	0.202	-0.028	0.226	0
	Isle Royale	SIW	23.50	5874	1.738	0.217	0.214	-0.014	0.203	7
	Parry Sound	HPW	23.47	5117	1.638	0.191	0.189	-0.015	0.197	0
Rapture Haplotypes	Seneca Lake	SLW	23.56	3180	1.746	0.216	0.213	-0.009	0.212	2
	Apostle Islands	SAW	23.29	3902	1.977	0.264	0.259	-0.020	0.197	3
	Marquette	SMD	20.84	3860	1.954	0.256	0.254	-0.008	0.215	1
	Green Lake	GLW	21.07	3632	1.875	0.246	0.237	-0.028	0.216	1
	Lewis Lake	LLW	21.93	3722	1.871	0.235	0.238	0.017	0.239	1
	Isle Royale	SIW	23.47	3904	1.973	0.262	0.258	-0.011	0.210	3
	Parry Sound	HPW	23.53	3488	1.825	0.233	0.228	-0.021	0.190	5

Table 5.2: Confusion matrices produced with DAPC using the Rapture haplotypes (A and D), Rapture SNPs (B and E), and GTSeq (C and F) datasets. Matrices A, B, and C correspond to results obtained from the DAPC model that minimized out-of-bag misclassification according to 30-fold cross validation for between 1 and 100 retained principal components and 6 retained discriminant functions. Matrices D, E, and F correspond to results from DAPC when the number of retained principal components was fixed to 20 and the number of discriminant functions was set to 6.

A

Known \ Assigned	LLW	GLW	SMD	HPW	SAW	SIW	SLW
LLW	1.00	0.00	0.00	0.00	0.00	0.00	0.00
GLW	0.04	0.92	0.04	0.00	0.00	0.00	0.00
SMD	0.00	0.00	0.91	0.00	0.09	0.00	0.00
HPW	0.00	0.00	0.00	1.00	0.00	0.00	0.00
SAW	0.00	0.00	0.08	0.00	0.92	0.00	0.00
SIW	0.00	0.00	0.00	0.00	0.04	0.96	0.00
SLW	0.00	0.00	0.00	0.00	0.00	0.00	1.00

D

Known \ Assigned	LLW	GLW	SMD	HPW	SAW	SIW	SLW
LLW	1.00	0.00	0.00	0.00	0.00	0.00	0.00
GLW	0.04	0.92	0.04	0.00	0.00	0.00	0.00
SMD	0.04	0.00	0.96	0.00	0.00	0.00	0.00
HPW	0.00	0.00	0.00	1.00	0.00	0.00	0.00
SAW	0.00	0.00	0.13	0.00	0.88	0.00	0.00
SIW	0.00	0.00	0.00	0.00	0.04	0.96	0.00
SLW	0.00	0.00	0.00	0.00	0.00	0.00	1.00

B

Known \ Assigned	LLW	GLW	SMD	HPW	SAW	SIW	SLW
LLW	1.00	0.00	0.00	0.00	0.00	0.00	0.00
GLW	0.00	0.95	0.05	0.00	0.00	0.00	0.00
SMD	0.00	0.00	1.00	0.00	0.00	0.00	0.00
HPW	0.00	0.00	0.00	1.00	0.00	0.00	0.00
SAW	0.00	0.00	0.13	0.00	0.88	0.00	0.00
SIW	0.00	0.00	0.00	0.00	0.04	0.96	0.00
SLW	0.00	0.00	0.00	0.00	0.00	0.00	1.00

E

Known \ Assigned	LLW	GLW	SMD	HPW	SAW	SIW	SLW
LLW	1.00	0.00	0.00	0.00	0.00	0.00	0.00
GLW	0.00	0.95	0.05	0.00	0.00	0.00	0.00
SMD	0.00	0.00	1.00	0.00	0.00	0.00	0.00
HPW	0.00	0.00	0.00	1.00	0.00	0.00	0.00
SAW	0.00	0.00	0.13	0.00	0.88	0.00	0.00
SIW	0.00	0.00	0.00	0.00	0.04	0.96	0.00
SLW	0.00	0.00	0.00	0.00	0.00	0.00	1.00

C

Known \ Assigned	LLW	GLW	SMD	HPW	SAW	SIW	SLW
LLW	0.95	0.00	0.05	0.00	0.00	0.00	0.00
GLW	0.00	0.88	0.13	0.00	0.00	0.00	0.00
SMD	0.00	0.00	0.96	0.00	0.04	0.00	0.00
HPW	0.00	0.00	0.00	1.00	0.00	0.00	0.00
SAW	0.00	0.00	0.08	0.00	0.92	0.00	0.00
SIW	0.00	0.00	0.00	0.00	0.00	1.00	0.00
SLW	0.00	0.00	0.00	0.00	0.00	0.00	1.00

F

Known \ Assigned	LLW	GLW	SMD	HPW	SAW	SIW	SLW
LLW	0.95	0.00	0.05	0.00	0.00	0.00	0.00
GLW	0.00	0.88	0.13	0.00	0.00	0.00	0.00
SMD	0.00	0.00	0.96	0.00	0.04	0.00	0.00
HPW	0.00	0.00	0.00	1.00	0.00	0.00	0.00
SAW	0.00	0.00	0.08	0.00	0.92	0.00	0.00
SIW	0.00	0.00	0.00	0.00	0.00	1.00	0.00
SLW	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 5.3: Mean, minimum, and maximum call rates for samples collected at 7 time points between 1940 and 1983. Call rate is the proportion of loci among individuals with sufficient data to call a genotype.

Collection Period	GTSeq (302 loci)	Rapture (9560 loci)
1940	0.022 (0.000, 0.116)	0.012 (0.005, 0.025)
1959	0.478 (0.000, 0.901)	0.061 (0.017, 0.212)
1969	0.925 (0.848, 0.987)	0.240 (0.060, 0.467)
1976	0.779 (0.652, 0.934)	0.245 (0.034, 0.485)
1979	0.934 (0.838, 0.997)	0.385 (0.174, 0.555)
1983	0.897 (0.682, 0.990)	0.490 (0.191, 0.869)
1986	0.825 (0.079, 0.997)	0.703 (0.432, 0.918)

Table 5.4: Regression analyses describing relationships between mixture proportions simulated using the leave-one out method in the Rubias program versus estimated mixture proportions. The slope of the relationship for each hatchery population is used as an indication of upward or downward bias in mixed stock analysis. Estimates of stock contribution are unbiased if the slope is equal to one. 95% confidence intervals for each slope estimate are included in parenthesis. Populations for which strain contributions are expected to be downwardly biased (slope < 0.95) are delineated with a (-) symbol. Populations for which strain contributions are expected to be upwardly biased (slope > 1.05) are delineated with a (+) symbol.

		Slope (Simulated vs. Estimated)					
Reporting Group	**Population**	**Rapture (SNP)**		**Rapture (Haplo.)**		**GTSeq**	
Seneca	Seneca Lake	0.995 (0.990, 1.000)		0.995 (0.990, 1.000)		0.994 (0.989, 0.999)	
Huron	Parry Sound	0.996 (0.991, 1.001)		0.996 (0.991,1.001)		1.042 (1.037, 1.047)	
Superior/Michigan	Apostle Islands	0.744 (0.729, 0.759)	-	0.931 (0.919, 0.943)	-	0.625 (0.609,0.641)	-
	Isle Royale	0.984 (0.972, 0.995)		1.005 (0.994, 1.017)		1.003 (0.991, 1.015)	
	Marquette	1.103 (1.080, 1.127)	+	1.009 (0.996, 1.021)		1.406 (1.376, 1.436)	+
	Green Lake	1.454 (1.427, 1.482)	+	0.989 (0.977, 1.000)		1.216 (1.189, 1.243)	+
	Lewis Lake	1.017 (1.005, 1.029)		0.969 (0.958, 0.981)		1.019 (1.006, 1.031)	
All Simulations		0.993 (0.990, 0.996)		0.994 (0.991, 0.996)		1.002 (0.998, 1.006)	

Table 5.5: Pairwise FST estimated for Rapture haplotypes (A), Rapture SNPs (B), GTSeq markers (C), and pairwise FST estimates calculated by Scribner et al. (2018) using 15 microsatellites. The lowest values are delineated with dark blue shading, while the highest values are delineated with dark red shading.

A

Rapture Haplotypes

	GLW	LLW	HPW	SAW	SIW	SLW
LLW	0.061					
HPW	0.099	0.097				
SAW	0.047	0.053	0.095			
SIW	0.070	0.075	0.114	0.037		
SLW	0.125	0.129	0.148	0.128	0.150	
SMD	0.037	0.046	0.089	0.018	0.038	0.119

B

Rapture SNPs

	GLW	LLW	HPW	SAW	SIW	SLW
LLW	0.071					
HPW	0.098	0.097				
SAW	0.048	0.054	0.094			
SIW	0.069	0.079	0.113	0.036		
SLW	0.131	0.132	0.147	0.127	0.153	
SMD	0.038	0.051	0.088	0.017	0.037	0.123

C

GTSeq

	GLW	LLW	HPW	SAW	SIW	SLW
LLW	0.050					
HPW	0.091	0.099				
SAW	0.048	0.046	0.108			
SIW	0.060	0.070	0.125	0.038		
SLW	0.114	0.111	0.131	0.117	0.134	
SMD	0.034	0.039	0.089	0.014	0.033	0.101

D

MicroSats (Scribner et al., 2018)

	GLW	LLW	HPW	SAW	SIW	SLW
LLW	0.049					
HPW	0.094	0.075				
SAW	0.036	0.028	0.075			
SIW	0.035	0.037	0.080	0.018		
SLW	0.082	0.081	0.089	0.059	0.073	
SMD	0.033	0.022	0.065	0.009	0.015	0.059

Supplemental Material 5.1: This file contains primer sequences for the Lake Trout GTseq panel

Supplemental Material 5.2: This file contains bait sequences for the 5011 locus Lake Trout RAD-Capture panel

CONCLUDING REMARKS

The resurgence of wild Lake Trout populations in Lake Huron (and other Great Lakes) provides an unprecedented opportunity to explore the population genetic factors that underly species recovery following extirpation and human mediated reintroduction. Additionally, Lake Trout express an exceptional diversity of ecomorphological variation, making them an optimal study species for exploring the genomic basis for adaptive radiation and incipient speciation. Over the course of this dissertation, I have created a suite of genomic resources that will be fundamental for future genomic research on Lake Trout. These include a high-density linkage map, a chromosome-anchored genome assembly, and three genotyping panels for the species.

These resources were used to address two questions relevant to Lake Trout conservation and reintroduction in the Great Lakes. In Chapter 3, I used a combination of low-coverage and conventional genotyping methodologies to identify a large number of loci associated with adaptive differences and reproductive isolations between Lake Trout ecomorphotypes that inhabit Lake Superior. I found that ecomorphotype associated loci are widely distributed across the genome and patterns of population genetic structure suggested that gene flow primarily occurred between leans and humpers and humpers and siscowets prior to a large-scale genetic homogenization event that occurred in the late 1990s or early 2000s. Results from time series samples suggested that levels of gene flow between ecomorphotypes increased substantially between the 1980s and 1990s.

In Chapter 4, I used local ancestry inference methodologies to identify genomic regions with an over- or under-abundance of haplotypes derived from the Seneca Lake ancestral population in a collection of 97 wild-born F2 hybrid Lake Trout from Lake Huron.

This analysis identified 7 genomic regions with an excess of alleles derived from the Seneca strain relative to the null expectation that ancestry tracts would be randomly distributed across hybrid genomes. I also identified two regions with an excess of alleles derived from Great Lakes strains, indicating that native populations contain some genetic variation that provides a fitness advantage in the contemporary environment. Interestingly, I identified multiple chromosomes with excesses and deficits of R_{HYB} suggesting that heterosis and outbreeding depression also explain some variation in fitness among wild-born Lake Trout in Lake Huron.

Results from these empirical studies shed light on the genetic factors associated with variation in fitness during the re-emergence of wild recruitment in Lake Huron, the genetic basis for ecomorphological variation in Lake Superior, and the factors that led to the breakdown of reproductive isolation between ecomorphotypes during population recovery. These studies provide important information that is relevant to Lake Trout conservation; however, the resources we have developed will likely have the greatest long-term impact on Lake Trout research. The Lake Trout linkage map and genome are of particular importance in this respect. These resources open up new avenues for research that were not accessible for non-model species just a few years ago; will be help to improve our understanding of evolution after autotetraploid genome duplication; and will be foundational to all future genomic research on Lake Trout.

REFERENCES

REFERENCES

Adams, C. E., D. Fraser, F. A. Huntingford, R. B. Greer, C. M. Askew, and A. F. Walker. "Trophic polymorphism amongst Arctic charr from Loch Rannoch, Scotland." Journal of Fish Biology 52, no. 6 (1998): 1259-1271.

Adams, Dean C., and Erik Otárola-Castillo. "geomorph: an R package for the collection and analysis of geometric morphometric shape data." Methods in Ecology and Evolution 4, no. 4 (2013): 393-399.

Ahrenstorff, T. D., Hrabik, T. R., Stockwell, J. D., Yule, D. L., & Sass, G. G. (2011). Seasonally dynamic diel vertical migrations of Mysis diluviana, coregonine fishes, and siscowet lake trout in the pelagia of western Lake Superior. Transactions of the American Fisheries Society, 140(6), 1504-1520.

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). RAD capture (Rapture): flexible and efficient sequence-based genotyping. Genetics, 202(2), 389-400.

Alexa, A., & Rahnenführer, J. (2009). Gene set enrichment analysis with topGO. Bioconductor Improv, 27, 1-26.

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

Allendorf, F. W., & Thorgaard, G. H. (1984). Tetraploidy and the evolution of salmonid fishes. In Evolutionary Genetics of Fishes. (pp. 1-53). Boston, MA: Springer.

Allendorf F.W., & Danzmann, R. G. (1997). Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. Genetics, 145(4), 1083-1092.

Allendorf, F. W., & Luikart, G. (2009). Conservation and the genetics of populations. John Wiley & Sons.

Allendorf, F.W., Funk, W.C., Aitken, S.N., Byrne, M., Luikart, G. (2022). Conservation and the Genomics of Populations. 3rd Ed. Oxford University Press.

Alfonso, Noel R. "Evidence for two morphotypes of lake charr, Salvelinus namaycush, from Great Bear Lake, Northwest Territories, Canada." Environmental Biology of Fishes 71, no. 1 (2004): 21-32.

Anderson, E. C., R. S. Waples, and Steven T Kalinowski. 2008. "An Improved Method for Predicting the Accuracy of Genetic Stock Identification." Can J Fish Aquat Sci 65:1475–86.

Andrews, S. (2010). Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data. URL: https://www. bioinformatics. babraham. ac. uk/projects/fastqc.

Andrews, Simon. "FastQC: a quality control tool for high throughput sequence data. Version 0.11. 5." Babraham Institute, Cambridge, UK http://www. bioinformatics. babraham. ac. uk/projects/fastqc (2014).

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

Anglès d'Auriac, M. B., H. A. Urke, and T. Kristensen. "A rapid qPCR method for genetic sex identification of Salmo salar and Salmo trutta including simultaneous elucidation of interspecies hybrid paternity by high-resolution melt analysis." Journal of fish biology 84, no. 6 (2014): 1971-1977.

Amish, S. J., Hohenlohe, P. A., Painter, S., Leary, R. F., Muhlfeld, C., Allendorf, F. W., & Luikart, G. (2012). RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. Molecular Ecology Resources, 12(4), 653-660.

Arendt, Jeffrey D. "Adaptive intrinsic growth rates: an integration across taxa." The quarterly review of biology 72, no. 2 (1997): 149-177.

Ayala, Diego, Rafael F. Guerrero, and Mark Kirkpatrick. "Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito." Evolution: International Journal of Organic Evolution 67, no. 4 (2013): 946-958.

Baetscher, D. S., Anderson, E. C., Gilbert-Horvath, E. A., Malone, D. P., Saarman, E. T., Carr, M. H., & Garza, J. C. (2019). Dispersal of a nearshore marine fish connects marine reserves and adjacent fished areas along an open coast. Molecular Ecology, 28, 1611–1623.

Baillie, S. M., Muir, A. M., Scribner, K., Bentzen, P., & Krueger, C. C. (2016). Loss of genetic diversity and reduction of genetic distance among lake trout Salvelinus namaycush ecomorphs, Lake Superior 1959 to 2013. Journal of Great Lakes Research, 42(2), 204-216.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS One, 3(10), e3376.

Balon, E. K. (1980). Charrs, salmonid fishes of the genus Salvelinus. Boston, MA: Kluwer.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology, 19(5), 455-477.

Barclay, A. W., Evenson, D. F., & Habicht, C. (2019). New Genetic Baseline for Upper Cook Inlet Chinook Salmon Allows for the Identification of More Stocks in Mixed Stock Fisheries: 413 Loci and 67 Populations. Alaska Department of Fish and Game, Division of Sport Fish, Research and Technical Services.

Barker, A. M., Adams, D. H., Driggers III, W. B., Frazier, B. S., & Portnoy, D. S. (2019). Hybridization between sympatric hammerhead sharks in the western North Atlantic Ocean. Biology letters, 15(4), 20190004.

Barra, V., and D. Fachinetti. "The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA." Nature Communications 9, no. 1 (2018): 4340.

Barría, A., Christensen, K. A., Yoshida, G., Jedlicki, A., Leong, J. S., Rondeau, E. B., & Yáñez, J. M. (2019). Whole genome linkage disequilibrium and effective population size in a coho salmon (Oncorhynchus kisutch) breeding population using a high-density SNP array. Frontiers in Genetics, 10, 498.

Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., ... & Primmer, C. R. (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. Nature, 528(7582), 405-408.

Barton, Bruce A., Car1 B. Schreck, and Lesley D. Barton. "Effects of chronic cortisol administration and daily acute stress on growth, physiological conditions, and stress responses in juvenile rainbow trout." Diseases of aquatic organisms 2, no. 3 (1987): 173-185.

Bay, Rachael A., Noah Rose, Rowan Barrett, Louis Bernatchez, Cameron K. Ghalambor, Jesse R. Lasky, Rachel B. Brem, Stephen R. Palumbi, and Peter Ralph. "Predicting responses to contemporary environmental change using evolutionary response architectures." The American Naturalist 189, no. 5 (2017): 463-473.

Beacham, T. D., Jonsen, K., McIntosh, B., Sutherland, B. J., Willis, D., Lynch, C., & Wallace, C. (2020a). Large-scale parentage-based tagging and genetic stock identification applied in assessing mixed-stock fisheries and hatchery brood stocks for coho salmon in British Columbia, Canada. Canadian Journal of Fisheries and Aquatic Sciences, 77(9), 1505-1517.

Beacham, T. D., Wallace, C., Jonsen, K., McIntosh, B., Candy, J. R., Rondeau, E. B., ... & Withler, R. E. (2020b). Accurate estimation of conservation unit contribution to coho salmon mixed-stock fisheries in British Columbia, Canada, using direct DNA sequencing for single nucleotide polymorphisms. Canadian Journal of Fisheries and Aquatic Sciences, 77(8), 1302-1315.

Benaglia T, Chauveau D, Hunter DR, Young D (2009). mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software, 32(6), 1–29.

Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J. A., & Jakobsen, K. S. (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. Heredity, 119(6), 418-428.

Bergstedt, R. A., Argyle, R. L., Seelye, J. G., Scribner, K. T., & Curtis, G. L. (2003). In situ determination of the annual thermal habitat use by lake trout (Salvelinus namaycush) in Lake Huron. Journal of Great Lakes Research, 29, 347-361.

Bernatchez, S., M. Laporte, C. Perrier, P. Sirois, and L. J. M. E. Bernatchez. "Investigating genomic and phenotypic parallelism between piscivorous and planktivorous lake trout (Salvelinus namaycush) ecotypes by means of RAD seq and morphometrics analyses." Molecular Ecology 25, no. 19 (2016): 4773-4792.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., ... Guiguen, Y. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nature Communications, 5(1), 1-10.

Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., ... Macqueen, D. J. (2020). The structural variation landscape in 492 Atlantic salmon genomes. Nature Communications, 11(1), 1-16.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.

Bond, C. E. "Biology of fishes." WB Sounders Company, Sounders (1979).

Bootsma, M. L., Gruenthal, K. M., McKinney, G. J., Simmons, L., Miller, L., Sass, G. G., & Larson, W. A. (2020). A GT-seq panel for walleye (Sander vitreus) provides important insights for efficient development and implementation of amplicon panels in non-model organisms. Molecular Ecology Resources, 20(6), 1706-1722.

Bortoluzzi, C., Bosse, M., Derks, M. F., Crooijmans, R. P., Groenen, M. A., & Megens, H. J. (2020). The type of bottleneck matters: Insights into the deleterious variation landscape of small managed populations. Evolutionary applications, 13(2), 330-341

Bourgey, M., Dali, R., Eveleigh, R., Chen, K. C., Letourneau, L., Fillon, J., ... Bourque, G. (2019). GenPipes: an open-source framework for distributed and scalable genomic analyses. GigaScience, 8(6), giz037.

Blackie, C.T., Weese, D.J., & Noakes, D.L.G. (2003). Evidence for resource polymorphism in the lake charr (Salvelinus namaycush) population of Great Bear Lake, Northwest Territories, Canada. Ecoscience, 10(4), 509-514.

Bradbury, Ian R., Sophie Hubert, Brent Higgins, Sharen Bowman, Tudor Borza, Ian G. Paterson, Paul VR Snelgrove et al. "Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish." Evolutionary Applications 6, no. 3 (2013): 450-461.

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., & Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. Nature, 513(7518), 375-381.

Brenden, T. O., R. W. Brown, M. P. Ebener, K. Reid, and T. J. Newcomb. 2013. Great Lakes commercial fisheries: historical overview and prognoses for the future. Pages 339–397 in W. W. Taylor, A. J. Lynch, and N. J. Leonard, editors. Great Lakes fisheries policy and management, 2nd edition. Michigan State University Press, East Lansing.

Brenna-Hansen, Silje, Jieying Li, Matthew P. Kent, Elizabeth G. Boulding, Sonja Dominik, William S. Davidson, and Sigbjørn Lien. "Chromosomal differences between European and North American Atlantic salmon discovered by linkage mapping and supported by fluorescence in situ hybridization analysis." BMC Genomics 13, no. 1 (2012): 432.

Brieuc, Marine SO, Charles D. Waters, James E. Seeb, and Kerry A. Naish. "A dense linkage map for Chinook salmon (Oncorhynchus tshawytscha) reveals variable chromosomal divergence after an ancestral whole genome duplication event." G3: Genes, Genomes, Genetics 4, no. 3 (2014): 447-460.

Broman, Karl W., and Saunak Sen. A Guide to QTL Mapping with R/qtl. Vol. 46. New York: Springer, 2009.

Broman, Karl W., Daniel M. Gatti, Petr Simecek, Nicholas A. Furlotte, Pjotr Prins, Śaunak Sen, Brian S. Yandell, and Gary A. Churchill. "R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations." Genetics 211, no. 2 (2019): 495-502.

Bronte, Charles R. "Evidence of spring spawning lake trout in Lake Superior." Journal of Great Lakes Research 19, no. 3 (1993): 625-629.

Bronte, Charles R., and Seth A. Moore. "Morphological variation of siscowet lake trout in Lake Superior." Transactions of the American Fisheries Society 136, no. 2 (2007): 509-517.

Bronte, C. R., Holey, M. E., Madenjian, C. P., Jonas, J. L., Claramunt, R. M., McKee, P. C., ... & Olsen, E. J. (2007). Relative abundance, site fidelity, and survival of adult lake trout in Lake Michigan from 1999 to 2001: implications for future restoration strategies. North American Journal of Fisheries Management, 27(1), 137-155.

Brown Jr, E. H., Eck, G. W., Foster, N. R., Horrall, R. M., & Coberly, C. E. (1981). Historical evidence for discrete stocks of lake trout (Salvelinus namaycush) in Lake Michigan. Canadian Journal of Fisheries and Aquatic Sciences, 38(12), 1747-1758.

Browning, B. L., & Browning, S. R. (2016). Genotype imputation with millions of reference samples. The American Journal of Human Genetics, 98(1), 116-126.

Burgos-Paz, W., Ramos-Onsins, S. E., Pérez-Enciso, M., & Ferretti, L. (2014). Correcting for unequal sampling in principal component analysis of genetic data. In Proceedings of the 10 th World Congress of Genetics Applied to Livestock Production.

Burnham-Curtis, M. K. (1993). Intralacustrine speciation of Salvelinus namaycush in Lake Superior: an investigation of genetic and morphological variation and evolution of lake trout in the Laurentian Great Lakes (Doctoral dissertation, University of Michigan).

Burnham-Curtis, M. K., & Smith, G. R. (1994). Osteological evidence of genetic divergence of lake trout (Salvelinus namaycush) in Lake Superior. Copeia, 843-850.

Burnham-Curtis, M. K., & Bronte, C. R. (1996). Otoliths reveal a diverse age structure for humper lake trout in Lake Superior. Transactions of the American Fisheries Society, 125(6), 844-851.

Butler, M. A., Sawyer, S. A., & Losos, J. B. (2007). Sexual dimorphism and adaptive radiation in Anolis lizards. Nature, 447(7141), 202-205.

Cáceres, A., Sindi, S. S., Raphael, B. J., Cáceres, M., & González, J. R. (2012). Identification of polymorphic inversions from genotypes. BMC Bioinformatics, 13(1), 1-16.

Campbell, N. R., Amish, S. J., Pritchard, V. L., McKelvey, K. S., Young, M. K., Schwartz, M. K., ... & Narum, S. R. (2012). Development and evaluation of 200 novel SNP assays for population genetic studies of westslope cutthroat trout and genetic identification of related taxa. Molecular Ecology Resources, 12(5), 942-949.

Campbell, M., Narum, S. (2008). Quantitative PCR assessment of microsatellite and SNP genotyping with variable quality DNA extracts. Cons Gen. DOI 10.1007/s10592-008-9661-7.

Campbell, N. R., Harmon, S. A., & Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Molecular Ecology Resources, 15(4), 855-867.

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., ... & Kinzig, A. P. (2012). Biodiversity loss and its impact on humanity. Nature, 486(7401), 59.

Catchen, Julian, Paul A. Hohenlohe, Susan Bassham, Angel Amores, and William A. Cresko. "Stacks: an analysis tool set for population genomics." Molecular Ecology 22, no. 11 (2013): 3124-3140.

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. Molecular Ecology Resources, 17(3), 362-365.

Catchen, J., Amores, A., & Bassham, S. (2020). Chromonomer: a tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny. G3: Genes, Genomes, Genetics, 10(11), 4115-4128.

Chaisson, M. J., & Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics, 13(1), 1-18.

Chavarie, L., Howland, K. L., & Tonn, W. M. (2013). Sympatric polymorphism in lake trout: the coexistence of multiple shallow-water morphotypes in Great Bear Lake. Transactions of the American Fisheries Society, 142(3), 814-823.

Chavarie, L., Howland, K., Harris, L., & Tonn, W. (2015). Polymorphism in lake trout in Great Bear Lake: intra-lake morphological diversification at two spatial scales. Biological Journal of the Linnean Society, 114(1), 109-125.

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., ... Turner, S. W. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods, 10(6), 563.

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., ... Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods, 13(12), 1050-1054.

Chomczynski, P., & Sacchi, N. (2006). The single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction: twenty-something years on. Nature Protocols, 1(2), 581-585.

Christensen, K. A., Rondeau, E. B., Minkley, D. R., Leong, J. S., Nugent, C. M., Danzmann, R. G., ... Koop, B. F. (2018a). The Arctic Char (Salvelinus alpinus) genome and transcriptome assembly. PloS One, 13(9), e0204076.

Christensen, K.A., Leong, J.S., Sakhrani, D., Biagi, C.A., Minkley, D.R., Withler, R.E., … Devlin, R.H. (2018b). Chinook salmon (Oncorhynchus tshawytscha) genome and transcriptome. PloS One, 13(4), e0195461.

Christensen, K. A., Rondeau, E. B., Minkley, D. R., Leong, J. S., Nugent, C. M., Danzmann, R. G., ... Koop, B. F. (2021). Retraction: The Arctic charr (Salvelinus alpinus) genome and transcriptome assembly.

Copeland, N. G., Jenkins, N. A., Gilbert, D. J., Eppig, J. T., Maltais, L. J., Miller, J. C., ... & Steen, R. G. (1993). A genetic linkage map of the mouse: current applications and future prospects. Science, 262(5130), 57-66.

Cosart, T., Beja-Pereira, A., Chen, S., Ng, S. B., Shendure, J., & Luikart, G. (2011). Exome-wide DNA capture and next generation sequencing in domestic and wild species. BMC Genomics, 12(1), 1-8.

Cotto, O., & Servedio, M. R. (2017). The roles of sexual and viability selection in the evolution of incomplete reproductive isolation: from allopatry to sympatry. The American Naturalist, 190(5), 680-693.

Crête-Lafrenière, Alexis, Laura K. Weir, and Louis Bernatchez. "Framing the Salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling." PloS one 7, no. 10 (2012): e46662.

Crossman, E. J. (1995). Introduction of the lake trout (Salvelinus namaycush) in areas outside its native distribution: a review. Journal of Great Lakes Research, 21, 17-29.

Crow, J. F. (1948). Alternative hypotheses of hybrid vigor. Genetics, 33(5), 477.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156-2158.

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013). Special features of RAD Sequencing data: implications for genotyping. Molecular Ecology, 22(11), 3151-3164.

De-Kayne, R., Zoller, S., & Feulner, P. G. (2020). A de novo chromosome-level genome assembly of Coregonus sp."Balchen": One representative of the Swiss Alpine whitefish radiation. Molecular Ecology Resources, 20(4), 1093-1109.

Döring, A., Weese, D., Rausch, T., & Reinert, K. (2008). SeqAn an efficient, generic C++ library for sequence analysis. BMC Bioinformatics, 9(1), 1-9.

Du, K., Stöck, M., Kneitz, S., Klopp, C., Woltering, J. M., Adolfi, M. C., ... Schartl, M. (2020). The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. Nature Ecology & Evolution, 4(6), 841-852.

Ducrest, Anne-Lyse, Laurent Keller, and Alexandre Roulin. "Pleiotropy in the melanocortin system, coloration and behavioural syndromes." Trends in Ecology & Evolution 23, no. 9 (2008): 502-510.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z. I., Knowler, D. J., Lévêque, C., ... & Sullivan, C. A. (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. Biological reviews, 81(2), 163-182.

Ebener, M. P. (1998). A lake trout rehabilitation guide for Lake Huron. Miscellaneous Publications, Great Lakes Fishery Commission, (1998), 1-44.

Edsall, T. A., & Kennedy, G. W. (1995). Availability of lake trout reproductive habitat in the Great Lakes. Journal of Great Lakes Research, 21, 290-301.

Ellegren, Hans, Linnea Smeds, Reto Burri, Pall I. Olason, Niclas Backström, Takeshi Kawakami, Axel Künstner et al. "The genomic landscape of species divergence in Ficedula flycatchers." Nature 491, no. 7426 (2012): 756.

Elliott, L., & Russello, M. A. (2018). SNP panels for differentiating advanced-generation hybrid classes in recently diverged stocks: A sensitivity analysis to inform monitoring of sockeye salmon re-stocking programs. Fisheries Research, 208, 339-345.

Elrod, J. H., O'Gorman, R., & Schneider, C. P. (1996). Bathythermal distribution, maturity, and growth of lake trout strains stocked in US waters of Lake Ontario, 1978–1993. Journal of Great Lakes Research, 22(3), 722-743.

Enberg, Katja, Christian Jørgensen, Erin S. Dunlop, Øystein Varpe, David S. Boukal, Loïc Baulier, Sigrunn Eliassen, and Mikko Heino. "Fishing-induced evolution of growth: concepts, mechanisms and the empirical evidence." Marine Ecology 33, no. 1 (2012): 1-25.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., ... Gibbs, R. A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PloS One, 7(11), e47768.

Eshenroder, R. L., Payne, N. R., Johnson, J. E., Bowen II, C., & Ebener, M. P. (1995). Lake trout rehabilitation in Lake Huron. Journal of Great Lakes Research, 21, 108-127.

Eshenroder, Randy L. "Differentiation of deep-water lake charr Salvelinus namaycush in North American lakes." Environmental Biology of Fishes 83, no. 1 (2008): 77-90.

Eschmeyer, Paul H., and Arthur M. Phillips Jr. "Fat content of the flesh of siscowets and lake trout from Lake Superior." Transactions of the American Fisheries Society 94, no. 1 (1965): 62-74.

Euclide, P. T., MacDougall, T., Robinson, J. M., Faust, M. D., Wilson, C. C., Chen, K. Y., ... & Ludsin, S. (2021). Mixed-stock analysis using Rapture genotyping to evaluate stock-specific exploitation of a walleye population despite weak genetic structure. Evolutionary Applications, 14(5), 1403-1420.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology, 14(8), 2611-2620.

Everett, Meredith V., Michael R. Miller, and James E. Seeb. "Meiotic maps of sockeye salmon derived from massively parallel DNA sequencing." BMC Genomics 13, no. 1 (2012): 521.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics, 131(2), 479-491.

Fernández, J., & Caballero, A. (2001). Accumulation of deleterious mutations and equalization of parental contributions in the conservation of genetic resources. Heredity, 86(4), 480-488.

Fisheries and Oceans Canada (DFO). 2012. Survey of recreational fishing in Canada 2010. Ottawa, Ontario, Canada

Fitzsimons, J. D., Brown, S., Brown, L., Honeyfield, D., He, J., & Johnson, J. E. (2010). Increase in lake trout reproduction in Lake Huron following the collapse of alewife: relief from thiamine deficiency or larval predation?. Aquatic Ecosystem Health & Management, 13(1), 73-84.

Fitzsimons, J. D., Brown, S. B., Williston, B., Williston, G., Brown, L. R., Moore, K., ... Tillitt, D. E. (2009). Influence of thiamine deficiency on lake trout larval growth, foraging, and predator avoidance. Journal of Aquatic Animal Health, 21(4), 302-314.

Flanagan, S. P., Forester, B. R., Latch, E. K., Aitken, S. N., & Hoban, S. (2018). Guidelines for planning genomic assessment and monitoring of locally adaptive variation to inform species conservation. Evolutionary Applications, 11(7), 1035-1052.

Fleming, I. A., & Petersson, E. (2001). The ability of released, hatchery salmonids to breed and contribute to the natural productivity of wild populations. Nordic Journal of Freshwater Research, 71-98.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences, 117(17), 9451-9457.

Froese, Rainer. "Cube law, condition factor and weight–length relationships: history, meta-analysis and recommendations." Journal of applied ichthyology 22, no. 4 (2006): 241-253.

Fujiwara, Atushi, Syuiti Abe, Etsuro Yamaha, Fumio Yamazaki, and Michihiro C. Yoshida. "Uniparental chromosome elimination in the early embryogenesis of the inviable salmonid hybrids between masu salmon female and rainbow trout male." Chromosoma 106, no. 1 (1997): 44-52.

Gagnaire, P. A., Normandeau, E., Pavey, S. A., & Bernatchez, L. (2013). Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (Coregonus clupeaformis). Molecular Ecology, 22(11), 3036-3048.

Gagnaire, Pierre-Alexandre, Scott A. Pavey, Eric Normandeau, and Louis Bernatchez. "The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing." Evolution 67, no. 9 (2013): 2483-2497.

Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., ... & Luikart, G. (2016). Genomics in conservation: case studies and bridging the gap between data and application. Trends in Ecology & Evolution, 31(2), 81-83.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv. preprint arXiv:1207.3907.

Gatch, A., Gorsky, D., Biesinger, Z., Bruestle, E., Lee, K., Karboski, C., ... & Wagner, T. (2021). Evidence of successful river spawning by lake trout (Salvelinus namaycush) in the Lower Niagara River, Lake Ontario. Journal of Great Lakes Research, 47(2), 486-493.

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. Genetics, 201(4), 1555-1579.

Gedir, J. V., Everest, T. I. A. N., & Moehrenschlager, A. X. E. L. (2004). Evaluating the potential for species reintroductions in Canada. In Proceedings of the Species at Risk 2004 Pathways to Recovery Conference (pp. 2-6).

Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C. S. (2017). Scaffolding of long read assemblies using long range contact information. BMC Genomics, 18(1), 1-11.

Gillard, G. B., Grønvold, L., Røsæg, L. L., Holen, M. M., Monsen, Ø., Koop, B. F., ... Hvidsten, T. R. (2021). Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. Genome Biology, 22(1), 1-18.

Goetz, F., Rosauer, D., Sitar, S., Goetz, G., Simchick, C., Roberts, S., ... Mackenzie, S. (2010). A genetic basis for the phenotypic differentiation between siscowet and lean lake trout (Salvelinus namaycush). Molecular Ecology, 19, 176-196.

Goetz, F., Smith, S. E., Goetz, G., & Murphy, C. A. (2016). Sea lampreys elicit strong transcriptomic responses in the lake trout liver during parasitism. BMC Genomics, 17(1), 1-16.

Goetz, F., Jasonowicz, A., Johnson, R., Biga, P., Fischer, G., & Sitar, S. (2014). Physiological differences between lean and siscowet lake trout morphotypes: Are these metabolotypes?. Canadian Journal of Fisheries and Aquatic Sciences, 71(3), 427-435.

Gogarten, S. M., Zheng, X., Gogarten, M. S. M., SeqArray, D., BiocGenerics, S., biocViews, S. N. P., & GeneticVariability, S. (2014). Package 'SeqVarTools'.

Gomez-Uchida D., Seeb J.E., Smith M.J, Habicht C, Quinn TP, Seeb LW (2011) Single nucleotide polymorphisms unravel hierarchical divergence and signatures of selection among Alaskan sockeye salmon (Oncorhynchus nerka) populations. BMC Evolutionary Biology, 11, 48. doi.org/10.1186/1471-2148-11-48.

Gomez-Uchida, D., K. P. Dunphy, M. F. O'connell, and Daniel Eduardo Ruzzante. "Genetic divergence between sympatric Arctic charr Salvelinus alpinus morphs in Gander Lake, Newfoundland: roles of migration, mutation and unequal effective population sizes." Journal of Fish Biology

Gonen, Serap, Natalie R. Lowe, Timothé Cezard, Karim Gharbi, Stephen C. Bishop, and Ross D. Houston. "Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing." BMC genomics 15, no. 1 (2014): 166. 73, no. 8 (2008): 2040-2057.

Gonzalez-Pena, Dianelys, Guangtu Gao, Matthew Baranski, Thomas Moen, Beth M. Cleveland, P. Brett Kenney, Roger L. Vallejo, Yniv Palti, and Timothy D. Leeds. "Genome-wide association study for identifying loci that affect fillet yield, carcass, and body weight traits in rainbow trout (Oncorhynchus mykiss)." Frontiers in genetics 7 (2016): 203.

Goodier, J. L. (1981). Native lake trout (Salvelinus namaycush) stocks in the Canadian waters of Lake Superior prior to 1955. Canadian Journal of Fisheries and Aquatic Sciences, 38(12), 1724-1737

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes, 5(1), 184-186.

Grant, P. R., & Grant, B. R. (2020). How and why species multiply. Princeton University Press.

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize implements and enhances circular visualization in R. Bioinformatics, 30(19), 2811-2812.

Guinand, B., Scribner, K. T., Page, K. S., & Burnham-Curtis, M. K. (2003). Genetic variation over space and time: analyses of extinct and remnant lake trout populations in the Upper Great Lakes. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270(1513), 425-433.

Guinand, B., Page, K. S., Burnham-Curtis, M. K., & Scribner, K. T. (2012). Genetic signatures of historical bottlenecks in sympatric lake trout (Salvelinus namaycush) morphotypes in Lake Superior. Environmental biology of fishes, 95(3), 323-334.

Gundappa, M. K., To, T. H., Grønvold, L., Martin, S. A., Lien, S., Geist, J., ... Macqueen, D. J. (2021). Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution. bioRxiv.

Haas, B. J., Delcher, A. L., Wortman, J. R., & Salzberg, S. L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics, 20(18), 3643-3646.

Hale, Matthew C., Garrett J. McKinney, Courtney L. Bell, and Krista M. Nichols. "Using linkage maps as a tool to determine patterns of chromosome synteny in the genus Salvelinus." G3: Genes, Genomes, Genetics 7, no. 11 (2017): 3821-3830.

Hansen, T. J., Penman, D., Glover, K. A., Fraser, T. W. K., Vågseth, T., Thorsen, A., ... Fjelldal, P. G. (2020). Production and verification of the first Atlantic salmon (Salmo salar L.) clonal lines. BMC Genetics, 21(1), 1-10.

Hansen, M. J. (1999). Lake trout in the Great Lakes: basin-wide stock collapse and binational restoration. In Taylor, W.W., Ferreri, C. P. (Eds.), Great Lakes Fishery Policy and Management: A Binational Perspective. (pp. 417-453). East Lansing, MI: Michigan State University Press.

Hansen, M. J., Nate, N. A., Krueger, C. C., Zimmerman, M. S., Kruckman, H. G., & Taylor, W. W. (2012). Age, growth, survival, and maturity of lake trout morphotypes in Lake Mistassini, Quebec. Transactions of the American Fisheries Society, 141(6), 1492-1503.

Hanson, S. D., Holey, M. E., Treska, T. J., Bronte, C. R., & Eggebraaten, T. H. (2013). Evidence of wild juvenile lake trout recruitment in western Lake Michigan. North American Journal of Fisheries Management, 33(1), 186-191.

Hansen, M. J., Nate, N. A., Muir, A. M., Bronte, C. R., Zimmerman, M. S., & Krueger, C. C. (2016a). Life history variation among four lake trout morphs at Isle Royale, Lake Superior. Journal of Great Lakes Research, 42(2), 421-432.

Hansen, M. J., Nate, N. A., Chavarie, L., Muir, A. M., Zimmerman, M. S., & Krueger, C. C. (2016b). Life history differences between fat and lean morphs of lake charr (Salvelinus namaycush) in Great Slave Lake, Northwest Territories, Canada. Hydrobiologia, 783(1), 21-35.

Hargrove, J. S., Camacho, C. A., Schrader, W. C., Powell, J. H., Delomas, T. A., Hess, J. E., ... & Campbell, M. R. (2021). Parentage-based tagging improves escapement estimates for ESA-listed adult Chinook salmon and steelhead in the Snake River basin. Canadian Journal of Fisheries and Aquatic Sciences, 78(4), 349-360.

Harrington, K. A., Hrabik, T. R., & Mensinger, A. F. (2015). Visual sensitivity of deepwater fishes in Lake Superior. PloS one, 10(2), e0116173.

Harris, L. N., Chavarie, L., Bajno, R., Howland, K. L., Wiley, S. H., Tonn, W. M., & Taylor, E. B. (2015). Evolution and origin of sympatric shallow-water morphotypes of Lake Trout, Salvelinus namaycush, in Canada's Great Bear Lake. Heredity, 114(1), 94-106.

Harvey, Chris J., Stephen T. Schram, and James F. Kitchell. "Trophic relationships among lean and siscowet lake trout in Lake Superior." Transactions of the American Fisheries Society 132, no. 2 (2003): 219-228.

Haug-Baltzell, A., Stephens, S. A., Davey, S., Scheidegger, C. E., & Lyons, E. (2017). SynMap2 and SynMap3D: web-based whole-genome synteny browsers. Bioinformatics, 33(14), 2197-2198.

Hayward, M. W., & Slotow, R. (2016). 13. Management of Reintroduced Wildlife Populations. In Reintroduction of Fish and Wildlife Populations (pp. 319-340). University of California Press.

He, J. X., Ebener, M. P., Riley, S. C., Cottrill, A., Kowalski, A., Koproski, S., ... & Johnson, J. E. (2012). Lake trout status in the main basin of Lake Huron, 1973–2010. North American Journal of Fisheries Management, 32(2), 402-412.

He, X., Johansson, M. L., & Heath, D. D. (2016). Role of genomics and transcriptomics in selection of reintroduction source populations. Conservation Biology, 30(5), 1010-1018.

Hendricks, S., Anderson, E., Antao, T., Bernatchez, L., Forester, B., Garner, B.A., Hand, B., Hohenlohe, P., Kardos, M., Koop, L.B., Waples, R.S., Luikart, G. (2018) Recent advances in population genomics data analysis: Improving bioinformatics and computational approaches. Evolutionary Applications, 11, 1197–1211.

Henikoff, Steven, Kami Ahmad, and Harmit S. Malik. "The centromere paradox: stable inheritance with rapidly evolving DNA." Science 293, no. 5532 (2001): 1098-1102.

Hindar, Kjetil, and Bror Jonsson. "Ecological polymorphism in Arctic charr." Biological Journal of the Linnean Society 48, no. 1 (1993): 63-74.

Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G., & McKeigue, P. M. (2004). Design and analysis of admixture mapping studies. The American Journal of Human Genetics, 74(5), 965-978.

Hoglund, E., P. H. Balm, and Svante Winberg. "Skin darkening, a potential social signal in subordinate Arctic charr (Salvelinus alpinus): the regulatory role of brain monoamines and pro-opiomelanocortin-derived peptides." Journal of Experimental Biology 203, no. 11 (2000): 1711-1721.

Hollenbeck, C. M., D. S. Portnoy, and J. R. Gold. "A method for detecting recent changes in contemporary effective population size from linkage disequilibrium at linked and unlinked loci." Heredity 117, no. 4 (2016): 207.

Hotaling, S., & Kelley, J. L. (2020). The rising tide of high-quality genomic resources. Molecular Ecology Resources, 19, 567–569.

Houston, Ross D., Chris S. Haley, Alastair Hamilton, Derrick R. Guy, Alan E. Tinch, John B. Taggart, Brendan J. McAndrew, and Stephen C. Bishop. "Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar)." Genetics 178, no. 2 (2008): 1109-1115.

Hunter, M. E., Hoban, S. M., Bruford, M. W., Segelbacher, G., & Bernatchez, L. (2018). Next-generation conservation genetics and biodiversity monitoring. Evolutionary Applications, 11(7), 1029-1034.

IUCN, S. (2013). Guidelines for reintroductions and other conservation translocations. Gland Switz Camb UK IUCNSSC Re-Introd Spec Group.

Jeffreys, H. (1998). The theory of probability. OUP Oxford.

Johnson, K. R., J. E. Wright, and B. May. "Linkage relationships reflecting ancestral tetraploidy in salmonid fish." Genetics 116, no. 4 (1987): 579-591.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics, 24(11), 1403-1405.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics, 11(1), 1-15.

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. Molecular Ecology, 25(1), 185-202.

Jones, M. R., Mills, L. S., Alves, P. C., Callahan, C. M., Alves, J. M., Lafferty, D. J., ... & Good, J. M. (2018). Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. Science, 360(6395), 1355-1358.

Jonsson, Bror, S. Skúlason, Sigur-dur S. Snorrason, Odd T. Sandlund, Hilmar J. Malmquist, P. M. Jónasson, R. Cydemo, and T. Lindem. "Life history variation of polymorphic Arctic charr (Salvelinus alpinus) in Thingvallavatn, Iceland." Canadian Journal of Fisheries and Aquatic Sciences 45, no. 9 (1988): 1537-1547.

Jordan, D. S., & Evermann, B. W. (1923). American Food and Game Fishes: A Popular Account of All the Species Found in America North of the Equator. Doubleday.

Kalitsis, Paul, and KH Andy Choo. "The evolutionary life cycle of the resilient centromere." Chromosoma 121, no. 4 (2012): 327-340.

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ, 2, e281.

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., & Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. Evolutionary Applications, 9(10), 1205-1218.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., ... Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Research, 24(8), 1384-1395.

Keenan, K., McGinnity, P., Cross, T. F., Crozier, W. W., & Prodöhl, P. A. (2013). diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. Methods in Ecology and Evolution, 4(8), 782-788.

Keller, L. F., Biebach, I., Ewing, S. R., & Hoeck, P. E. (2012). The genetics of reintroductions: inbreeding and genetic drift. Reintroduction biology: integrating science and management, 9, 360.

Kess, T., Dempson, J. B., Lehnert, S. J., Layton, K., Einfeldt, A., Bentzen, P., & Bradbury, I. R. (2021). Genomic basis of deep-water adaptation in Arctic Charr (Salvelinus alpinus) morphs. Molecular Ecology.

Khan, N. Y., & Qadri, S. U. (1970). Morphological differences in Lake Superior lake char. Journal of the Fisheries Board of Canada, 27(1), 161-167.

Kijas, James, Sean McWilliam, Marina Naval Sanchez, Peter Kube, Harry King, Bradley Evans, Torfinn Nome, Sigbjørn Lien, and Klara Verbyla. "Evolution of sex determination loci in Atlantic salmon." Scientific reports 8, no. 1 (2018): 1-11.

Kikuchi, A., Yamamoto, H., Sato, A., & Matsumoto, S. (2012). Wnt5a: its signalling, functions and implication in diseases. Acta Physiologica, 204(1), 17-33.

Kincaid, H. L., Krueger, C. C., & May, B. (1993). Preservation of genetic variation in the Green Lake strain lake trout derived from remnant domestic and feral populations. North American Journal of Fisheries Management, 13(2), 318-325.

Kim, B. Y., Huber, C. D., & Lohmueller, K. E. (2018). Deleterious variation shapes the genomic landscape of introgression. PLoS Genetics, 14(10), e1007741.

Kirkpatrick, Mark. "How and why chromosome inversions evolve." PLoS biology 8, no. 9 (2010): e1000501.

Kittilsen, Silje, Joachim Schjolden, I. Beitnes-Johansen, J. C. Shaw, Tom G. Pottinger, Christina Sørensen, Bjarne Olai Braastad, Morten Bakken, and Ø. Øverli. "Melanin-based skin spots reflect stress responsiveness in salmonid fish." Hormones and behavior 56, no. 3 (2009): 292-298.

Kodama, M., Brieuc, M. S., Devlin, R. H., Hard, J. J., & Naish, K. A. (2014). Comparative mapping between Coho Salmon (Oncorhynchus kisutch) and three other salmonids suggests a role for chromosomal rearrangements in the retention of duplicated regions following a whole genome duplication event. G3: Genes, Genomes, Genetics, 4(9), 1717-1730.

Koel, T. M., Bigelow, P. E., Doepke, P. D., Ertel, B. D., & Mahony, D. L. (2005). Nonnative lake trout result in Yellowstone cutthroat trout decline and impacts to bears and anglers. Fisheries, 30(11), 10-19.

Komen, H., & Thorgaard, G. H. (2007). Androgenesis, gynogenesis and the production of clones in fishes: a review. Aquaculture, 269(1-4), 150-173.

Komoroske, L. M., Miller, M. R., O'Rourke, S. M., Stewart, K. R., Jensen, M. P., & Dutton, P. H. (2019). A versatile Rapture (RAD-Capture) platform for genotyping marine turtles. Molecular Ecology Resources, 19(2), 497-511.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research, 27(5), 722-736.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. BMC Bioinformatics, 15(1), 1-13.

Kornis M.S., Bronte C.R., Holey M.E., Hanson S.D., Treska T.J., Jonas J.L., et al. 2019b. Factors affecting postrelease survival of coded-wire-tagged lake trout in Lake Michigan at four historical spawning locations. N. Am. J. Fish. Manage. 39(5): 868–895.

Kovach, R. P., Hand, B. K., Hohenlohe, P. A., Cosart, T. F., Boyer, M. C., Neville, H. H., ... & Luikart, G. (2016). Vive la résistance: genome-wide selection against introduced alleles in invasive hybrid zones. Proceedings of the Royal Society B: Biological Sciences, 283(1843), 20161380.

Krueger, C. C., Horrall, R. M., & Gruenthal, H. (1983). Strategy for the use of lake trout strains in Lake Michigan. Wisconsin Department of Natural Resources, Administrative Report, 17.

Krueger, C. C., Jones, M. L., & Taylor, W. W. (1995). Restoration of lake trout in the Great Lakes: challenges and strategies for future management. Journal of Great Lakes Research, 21, 547-558.

Krueger, C. C., & Ihssen, P. E. (1995). Review of genetics of lake trout in the Great Lakes: history, molecular genetics, physiology, strain comparisons, and restoration management. Journal of Great Lakes Research, 21, 348-363.

Küpper, Clemens, Michael Stocks, Judith E. Risse, Natalie dos Remedios, Lindsay L. Farrell, Susan B. McRae, Tawna C. Morgan et al. "A supergene determines highly divergent male reproductive morphs in the ruff." Nature genetics 48, no. 1 (2016): 79.

Lantry, J.R. (2015). Eastern basin of Lake Ontario warmwater fisheries assessment, 1976–2014. In 2014 annual report, Bureau of Fisheries, Lake Ontario Unit and St Lawrence River Unit to the Great Lakes Fishery Commission's Lake Ontario Committee. (pp. 1–35).

Larson WA, Seeb JE, Dann TH, Schindler DE, Seeb LW (2014) Signals of heterogeneous selection at an MHC locus in geographically proximate ecotypes of sockeye salmon. Molecular Ecology, 23, 5448–5461.

Larson, Wesley A., Garrett J. McKinney, Morten T. Limborg, Meredith V. Everett, Lisa W. Seeb, and James E. Seeb. "Identification of multiple QTL hotspots in Sockeye Salmon (Oncorhynchus nerka) using genotyping-by-sequencing and a dense linkage map." Journal of Heredity 107, no. 2 (2015): 122-133.

Larson, W. A., Limborg, M. T., McKinney, G. J., Schindler, D. E., Seeb, J. E., & Seeb, L. W. (2017). Genomic islands of divergence linked to ecotypic variation in sockeye salmon. Molecular Ecology, 26(2), 554-570.

Larson, W. A., Kornis, M. S., Turnquist, K. N., Bronte, C. R., Holey, M. E., Hanson, S. D., ... & Sloss, B. L. (2021). The genetic composition of wild recruits in a recovering lake trout population in Lake Michigan. Canadian Journal of Fisheries and Aquatic Sciences, 78(3), 286-300.

Leary, Robb F., Fred W. Allendorf, and Stephen H. Forbes. "Conservation genetics of bull trout in the Columbia and Klamath River drainages." Conservation Biology 7, no. 4 (1993): 856-865.

Li, J., Takaishi, K., Cook, W., McCorkle, S. K., & Unger, R. H. (2003). Insig-1 "brakes" lipogenesis in adipocytes and inhibits differentiation of preadipocytes. Proceedings of the National Academy of Sciences, 100(16), 9476-9481.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094-3100.

Leitwein, Maeva, John Carlos Garza, and Devon E. Pearse. "Ancestry and adaptive evolution of anadromous, resident, and adfluvial rainbow trout (Oncorhynchus mykiss) in the San Francisco bay area: application of adaptive genomic variation to conservation in a highly impacted landscape." Evolutionary applications 10, no. 1 (2017): 56-67.

Leitwein, Maeva, Pierre-Alexandre Gagnaire, Erick Desmarais, Patrick Berrebi, and Bruno Guinand. "Genomic consequences of a recent three-way admixture in supplemented wild brown trout populations revealed by local ancestry tracts." Molecular ecology 27, no. 17 (2018): 3466-3483.

Leitwein, M., Duranton, M., Rougemont, Q., Gagnaire, P. A., & Bernatchez, L. (2020). Using haplotype information for conservation genomics. Trends in ecology & evolution, 35(3), 245-258.

Lien, Sigbjørn, Lars Gidskehaug, Thomas Moen, Ben J. Hayes, Paul R. Berg, William S. Davidson, Stig W. Omholt, and Matthew P. Kent. "A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns." BMC genomics 12, no. 1 (2011): 615.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., ... Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. Nature, 533(7602), 200-205.

Limborg, M. T., Seeb, L. W., & Seeb, J. E. (2016). Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. Molecular Ecology, 25(10), 2117-2129.

Limborg, Morten T., Ryan K. Waples, Fred W. Allendorf, and James E. Seeb. "Linkage mapping reveals strong chiasma interference in sockeye salmon: implications for interpreting genomic data." G3: Genes, Genomes, Genetics 5, no. 11 (2015): 2463-2473.

Limborg, Morten T., Garrett J. McKinney, Lisa W. Seeb, and James E. Seeb. "Recombination patterns reveal information about centromere location on linkage maps." Molecular ecology resources 16, no. 3 (2016): 655-661.

Lindner, K. R., J. E. Seeb, C. Habicht, K. L. Knudsen, E. Kretschmer, D. J. Reedy, P. Spruell, and F. W. Allendorf. "Gene-centromere mapping of 312 loci in pink salmon by half-tetrad analysis." Genome 43, no. 3 (2000): 538-549.

Liu, S., Palti, Y., Gao, G., & Rexroad III, C. E. (2016). Development and validation of a SNP panel for parentage assignment in rainbow trout. Aquaculture, 452, 178-182.

Lowry, D. B. (2012). Ecotypes and the controversy over stages in the formation of new species. Biological Journal of the Linnean Society, 106(2), 241-257.

Lowry, David B., and John H. Willis. "A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation." PLoS biology 8, no. 9 (2010): e1000500.

Luikart, G., Antao, T., Hand, B. K., Muhlfeld, C. C., Boyer, M. C., Cosart, T., ... & Waples, R. S. (2021). Detecting population declines via monitoring the effective number of breeders (Nb). Molecular Ecology Resources, 21(2), 379-393.

Lu, Guoqing, and Louis Bernatchez. "Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (Coregonus clupeaformis): support for the ecological speciation hypothesis." Evolution 53, no. 5 (1999): 1491-1505.

Lubieniecki, Krzysztof P., Song Lin, Emily I. Cabana, Jieying Li, Yvonne YY Lai, and William S. Davidson. "Genomic instability of the sex-determining locus in Atlantic salmon (Salmo salar)." G3: Genes, Genomes, Genetics 5, no. 11 (2015): 2513-2522.

Luo, L., & Xu, S. (2003). Mapping viability loci using molecular markers. Heredity, 90(6), 459-467.

Luu, K., Bazin, E., & Blum, M. G. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. Molecular ecology resources, 17(1), 67-77.

Lynch, M., & Force, A. G. (2000). The origin of interspecific genomic incompatibility via gene duplication. The American Naturalist, 156(6), 590-605.

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics, 93(2), 278-288.

Ma, J., & Amos, C. I. (2012). Principal components analysis of population admixture. PloS one, 7(7), e40115.

Mac, M. J., & Edsall, C. C. (1991). Environmental contaminants and the reproductive success of lake trout in the Great Lakes: an epidemiological approach. Journal of Toxicology and Environmental Health, Part A Current Issues, 33(4), 375-394.

Macqueen, D. J., & Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proceedings of the Royal Society B: Biological Sciences, 281(1778), 20132881.

Macqueen, D. J., Primmer, C. R., Houston, R. D., Nowak, B. F., Bernatchez, L., Bergseth, S., ... Yáñez, J. M. (2017). Functional Annotation of All Salmonid Genomes (FAASG): an international initiative supporting future salmonid research, conservation and aquaculture. 1-9.

Madenjian, C. P., DeSorcie, T. J., & Stedman, R. M. (1998). Maturity schedules of lake trout in Lake Michigan. Journal of Great Lakes Research, 24(2), 404-410.

Madenjian, C. P., O'Gorman, R., Bunnell, D. B., Argyle, R. L., Roseman, E. F., Warner, D. M., ... Stapanian, M. A. (2008). Adverse effects of alewives on Laurentian Great Lakes fish communities. North American Journal of Fisheries Management, 28(1), 263-282.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., ... & Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. Nature Methods, 7(2), 111-118.

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics, 93(2), 278-288.

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

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLoS Computational Biology, 14(1), e1005944.

Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. Nature Protocols, 12(2), 213-218.

Marin, K., Coon, A., Carson, R., Debes, P. V., & Fraser, D. J. (2016). Striking phenotypic variation yet low genetic differentiation in sympatric lake trout (Salvelinus namaycush). PloS One, 11(9), e0162325.

Marin, Kia, Andrew Coon, and Dylan Fraser. "Traditional ecological knowledge reveals the extent of sympatric lake trout diversity and habitat preferences." Ecology and Society 22, no. 2 (2017).

Marsden, J. Ellen, Charles C. Krueger, Peter M. Grewe, Harold L. Kincaid, and Bernie May. "Genetic comparison of naturally spawned and artificially propagated Lake Ontario lake trout fry: evaluation of a stocking strategy for species rehabilitation." North American Journal of Fisheries Management 13, no. 2 (1993): 304-317.

Marsden, C. D., Ortega-Del Vecchyo, D., O'Brien, D. P., Taylor, J. F., Ramirez, O., Vilà, C., ... & Lohmueller, K. E. (2016). Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proceedings of the National Academy of Sciences, 113(1), 152-157.

Marsden, J. E., Noakes, D. L., Krueger, C. C. (2021). Terminology Issues in Lake Charr Early Development. In Muir, A. M. (Ed.), The Lake Charr Salvelinus namaycush: Biology, Ecology, Distribution, and Management. (pp. 487-497). Cham, Switzerland: Springer International Publishing.

Martin, N. V., and C. H. Olver. 1980. The Lake Charr, Salvelinus namaycush. Pages 205–277 in E. K. Balon, editor. Charrs: salmonid fishes of the genus Salvelinus. Dr W. Junk, The Hague, The Netherlands.

Maruki, T., & Lynch, M. (2017). Genotype calling from population-genomic sequencing data. G3: Genes, Genomes, Genetics, 7(5), 1393-1404.

May, S. A., McKinney, G. J., Hilborn, R., Hauser, L., & Naish, K. A. (2020). Power of a dual-use SNP panel for pedigree reconstruction and population assignment. Ecology and Evolution, 10(17), 9522-9531.

Margres, M. J., Jones, M. E., Epstein, B., Kerlin, D. H., Comte, S., Fox, S., ... & Storfer, A. (2018). Large-effect loci affect survival in Tasmanian devils (Sarcophilus harrisii) infected with a transmissible cancer. Molecular Ecology, 27(21), 4189-4199.

May, Bernie, James E. Wright, and Mark Stoneking. "Joint segregation of biochemical loci in Salmonidae: results from experiments with Salvelinus and review of the literature on other species." Journal of the Fisheries Board of Canada 36, no. 9 (1979): 1114-1128.

McGlauflin MT, Schindler DE, Seeb LW, Smith CT, Habicht C, Seeb JE (2011) Spawning habitat and geography influence population structure and juvenile migration timing of sockeye salmon in the Wood River Lakes, Alaska. Transactions of the American Fisheries Society, 140, 763–782.

McKinney, G. J., L. W. Seeb, W. A. Larson, D. Gomez-Uchida, Morten Tønsberg Limborg, M. S. O. Brieuc, M. V. Everett, K. A. Naish, R. K. Waples, and J. E. Seeb. "An integrated linkage map reveals candidate genes underlying adaptive variation in Chinook salmon (Oncorhynchus tshawytscha)." Molecular ecology resources 16, no. 3 (2016): 769-783.

McKinney, G. J., Seeb, J. E., & Seeb, L. W. (2017). Managing mixed-stock fisheries: Genotyping multi-SNP haplotypes increases power for genetic stock identification. Canadian Journal of Fisheries and Aquatic Sciences, 74, 429–434.

McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Molecular Ecology Resources, 17(4), 656-669.

McKinney, Garrett J., Carita E. Pascal, William D. Templin, Sara E. Gilk-Baumer, Tyler H. Dann, Lisa W. Seeb, and James E. Seeb. "Dense SNP panels resolve closely related Chinook salmon populations." Canadian Journal of Fisheries and Aquatic Sciences 999 (2019): 1-11.

McDermid, J. L., Walker, J., Al-Shamlih, M., & Wilson, C. C. (2020). Genetic Integrity of Lake Trout in Cold Lake, Alberta, Despite Decades of Supplemental Stocking. North American Journal of Fisheries Management, 40(2), 459-474.

Meek, M. H., & Larson, W. A. (2019). The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era. Molecular Ecology Resources, 19, 795-803.

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. Genetics, 210(2), 719-731.

Meisner, J., Albrechtsen, A., & Hanghøj, K. (2021). Detecting Selection in Low-Coverage High-Throughput Sequencing Data using Principal Component Analysis. bioRxiv.

Mérot, C. (2020). Making the most of population genomic data to understand the importance of chromosomal inversions for adaptation and speciation. Molecular Ecology, 29(14), 2513-2516.

Miller, M. R., Brunelli, J. P., Wheeler, P. A., Liu, S., REXROAD III, C. E., Palti, Y., ... & Thorgaard, G. H. (2012). A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. Molecular Ecology, 21(2), 237-249.

Moen, T., B. Hoyheim, H. Munck, and L. Gomez-Raya. "A linkage map of Atlantic salmon (Salmo salar) reveals an uncommonly large difference in recombination rate between the sexes." Animal genetics 35, no. 2 (2004): 81-92.

Moen, Thomas, Ben Hayes, Matthew Baranski, Paul R. Berg, Sissel Kjøglum, Ben F. Koop, Willie S. Davidson, Stig W. Omholt, and Sigbjørn Lien. "A linkage map of the Atlantic salmon (Salmo salar) based on EST-derived SNP markers." BMC genomics 9, no. 1 (2008): 223.

Moen, Thomas, Matthew Baranski, Anna K. Sonesson, and Sissel Kjøglum. "Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar): population-level associations between markers and trait." BMC genomics 10, no. 1 (2009): 368.

Moore, S. A., & Bronte, C. R. (2001). Delineation of sympatric morphotypes of lake trout in Lake Superior. Transactions of the American Fisheries Society, 130(6), 1233-1240.

Muir, A. M., Krueger, C. C., & Hansen, M. J. (2012). Re-establishing lake trout in the Laurentian Great Lakes: past, present, and future. Great Lakes fishery policy and management: a binational perspective, 2nd edition. Michigan State University Press, East Lansing, 533-588.

Muir, A. M., Bronte, C. R., Zimmerman, M. S., Quinlan, H. R., Glase, J. D., & Krueger, C. C. (2014). Ecomorphological diversity of lake trout at Isle Royale, Lake Superior. Transactions of the American Fisheries Society, 143(4), 972-987.

Muir, A. M., Hansen, M. J., Bronte, C. R., & Krueger, C. C. (2016). If Arctic charr Salvelinus alpinus is 'the most diverse vertebrate', what is the lake charr Salvelinus namaycush?. Fish and Fisheries, 17(4), 1194-1207.

Muir, A. M., Krueger, C. C., Hansen, M. J., & Riley, S. C. (2021) The Lake Charr Salvelinus namaycush: Biology, Ecology, Distribution, and Management(Vol. 39). Springer Nature.

Mukhopadhyay, S., & Jackson, P. K. (2011). The tubby family proteins. Genome Biology, 12(6), 1-9.

Muhlfeld, C. C., Kalinowski, S. T., McMahon, T. E., Taper, M. L., Painter, S., Leary, R. F., & Allendorf, F. W. (2009). Hybridization rapidly reduces fitness of a native trout in the wild. Biology letters, 5(3), 328-331.

Nicholson, G., Smith, A. V., Jónsson, F., Gústafsson, Ó., Stefánsson, K., & Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(4), 695-715.

Nugent, Cameron M., Anne A. Easton, Joseph D. Norman, Moira M. Ferguson, and Roy G. Danzmann. "A SNP based linkage map of the Arctic Charr (Salvelinus alpinus) genome provides insights into the diploidization process after whole genome duplication." G3: Genes, Genomes, Genetics 7, no. 2 (2017): 543-556.

Murray, J. C., Buetow, K. H., Weber, J. L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., ... & Duyk, G. M. (1994). A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). Science, 265(5181), 2049-2054.

Ødegård, J., Moen, T., Santi, N., Korsvoll, S. A., Kjøglum, S., & Meuwissen, T. H. (2014). Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). Frontiers in Genetics, 5, 402.

Ohno, S. (1970). Evolution by gene duplication. New York, NY: Springer-Verlag.

Olazcuaga, L., Loiseau, A., Parrinello, H., Paris, M., Fraimout, A., Guedot, C., ... & Gautier, M. (2020). A whole-genome scan for association with invasion success in the fruit fly Drosophila suzukii using contrasts of allele frequencies corrected for population structure. Molecular biology and evolution, 37(8), 2369-2385.

Oleksyk, T. K., Smith, M. W., & O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1537), 185-205.

Oosting, J., Eilers, P., Menezes, R., Oosting, M. J., & biocViews Visualization, C. (2005). Package 'quantsmooth'. Bioinformatics, 21(7), 1146-53.

O'Sullivan, T. Norene, Xufeng S. Wu, Rivka A. Rachel, Jiang-Dong Huang, Deborah A. Swing, Lydia E. Matesic, John A. Hammer, Neal G. Copeland, and Nancy A. Jenkins. "dsu functions in a MYO5A-independent pathway to suppress the coat color of dilute mice." Proceedings of the National Academy of Sciences 101, no. 48 (2004): 16831-16836.

Oziolor, E. M., Reid, N. M., Yair, S., Lee, K. M., VerPloeg, S. G., Bruns, P. C., ... & Matson, C. W. (2019). Adaptive introgression enables evolutionary rescue from extreme environmental pollution. Science, 364(6439), 455-457.

Page, K. S. (2001). Genetic diversity and interrelationships of wild and hatchery lake trout in the upper Great Lakes: inferences for broodstock management and development of restoration strategies. Michigan State University.

Page, K. S., Scribner, K. T., Bennett, K. R., Garzel, L. M., & Burnham-Curtis, M. K. (2003). Genetic assessment of strain-specific sources of lake trout recruitment in the Great Lakes. Transactions of the American Fisheries Society, 132(5), 877-894.

Page, K. S., Scribner, K. T., & Burnham-Curtis, M. (2004). Genetic diversity of wild and hatchery lake trout populations: relevance for management and restoration in the Great Lakes. Transactions of the American Fisheries Society, 133(3), 674-691.

Page, K. S., Scribner, K. T., Bast, D., Holey, M. E., & Burnham-Curtis, M. K. (2005). Genetic evaluation of a Great Lakes lake trout hatchery program. Transactions of the American Fisheries Society, 134(4), 872-891.

Palti, Yniv, Carine Genet, Guangtu Gao, Yuqin Hu, Frank M. You, Mekki Boussaha, Caird E. Rexroad, and Ming-Cheng Luo. "A second generation integrated map of the rainbow trout (Oncorhynchus mykiss) genome: analysis of conserved synteny with model fish genomes." Marine biotechnology 14, no. 3 (2012): 343-357.

Pan, W., Jiang, T., & Lonardi, S. (2020). OMGS: optical map-based genome scaffolding. Journal of Computational Biology, 27(4), 519-533.

Parejo, M., Wragg, D., Henriques, D., Charrière, J. D., & Estonba, A. (2020). Digging into the genomic past of Swiss honey bees by whole-genome sequencing museum specimens. Genome Biology and Evolution, 12(12), 2535-2551.

Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadía-Cardoso, A., ... & Lien, S. (2019). Sex-dependent dominance maintains migration supergene in rainbow trout. Nature Ecology & Evolution, 3(12), 1731-1742.

Peichel, C. L., Nereng, K. S., Ohgi, K. A., Cole, B. L., Colosimo, P. F., Buerkle, C. A., ... & Kingsley, D. M. (2001). The genetic architecture of divergence between threespine stickleback species. Nature, 414(6866), 901-905.

Perkins, D. L., Fitzsimons, J. D., Marsden, J. E., Krueger, C. C., & May, B. (1995). Differences in reproduction among hatchery strains of lake trout at eight spawning areas in Lake Ontario: genetic evidence from mixed-stock analysis. Journal of Great Lakes Research, 21, 364-374.

Perreault-Payette, A., Muir, A. M., Goetz, F., Perrier, C., Normandeau, E., Sirois, P., & Bernatchez, L. (2017). Investigating the extent of parallelism in morphological and genomic divergence among lake trout ecotypes in Lake Superior. Molecular Ecology, 26(6), 1477-1497.

Pfennig, David W., and Karin S. Pfennig. Evolution's wedge: competition and the origins of diversity. No. 12. Univ of California Press, 2012.

Pflug, J. M., Holmes, V. R., Burrus, C., Johnston, J. S., & Maddison, D. R. (2020). Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). G3: Genes, Genomes, Genetics, 10(9), 3047-3060.

Phillips, Ruth B., and Kerry D. Zajicek. "Q band chromosomal polymorphisms in lake trout (Salvelinus namaycush)." Genetics 101, no. 2 (1982): 227-234.

Phillips, R. B., and P. E. Ihssen. "Identification of sex chromosomes in lake trout (Salvelinus namaycush)." Cytogenetic and Genome Research 39, no. 1 (1985): 14-18.

Pierron, D., Heiske, M., Razafindrazaka, H., Pereda-Loth, V., Sanchez, J., Alva, O., ... & Letellier, T. (2018). Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. Nature communications, 9(1), 1-9.

Pool, J. E. (2015). The mosaic ancestry of the Drosophila genetic reference panel and the D. melanogaster reference genome reveals a network of epistatic fitness interactions. Molecular biology and evolution, 32(12), 3236-3251.

Postlethwait, J. H., Johnson, S. L., Midson, C. N., Talbot, W. S., Gates, M., Ballinger, E. W., ... & Eisen, J. S. (1994). A genetic linkage map for the zebrafish. Science, 264(5159), 699-703.

Primmer, C. R. (2009). From conservation genetics to conservation genomics. Annals of the New York Academy of Sciences, 1162(1), 357-368.

Primmer, C., Koskinen, M T., Piironen, J. (2000). The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. Proc. R. Soc. Lond. B, 267, 1699-1704

Prince, D. J., O'Rourke, S. M., Thompson, T. Q., Ali, O. A., Lyman, H. S., Saglam, I. K., & Miller, M. R. (2017). The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation. Science Advances, 3(8), e1603198.

Protas, Meredith E., and Nipam H. Patel. "Evolution of coloration patterns." Annual review of cell and developmental biology 24 (2008): 425-446.

Puckett, E. E. (2017). Variability in total project and per sample genotyping costs under varying study designs including with microsatellites or SNPs to answer conservation genetic questions. Conservation Genetics Resources, 9(2), 289-304.

Pycha, R. L. (1980). Changes in mortality of lake trout (Salvelinus namaycush) in Michigan waters of Lake Superior in relation to sea lamprey (Petromyzon marinus) predation, 1968–78. Canadian Journal of Fisheries and Aquatic Sciences, 37(11), 2063-2073.

Qiu, Changliang, Zhaofang Han, Wanbo Li, Kun Ye, Yangjie Xie, and Zhiyong Wang. "A high-density genetic linkage map and QTL mapping for growth and sex of yellow drum (Nibea albiflora)." Scientific reports 8, no. 1 (2018): 17271.

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

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rahrer, J. F. (1965). Age, growth, maturity, and fecundity of humper lake trout, Isle Royale, Lake Superior. Transactions of the American Fisheries Society, 94(1), 75-83.

Rakestraw, L. 1968. Commercial fishing on Isle Royale. Isle Royale National History Association, Houghton, Michigan

Rastas, Pasi, Federico CF Calboli, Baocheng Guo, Takahito Shikano, and Juha Merilä. "Construction of ultradense linkage maps with Lep-MAP2: stickleback F 2 recombinant crosses as an example." Genome biology and evolution 8, no. 1 (2015): 78-93.

Rastas, Pasi. "Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data." Bioinformatics 33, no. 23 (2017): 3726-3732.

Reading, R. P., Miller, B., & Shepherdson, D. (2013). The value of enrichment to reintroduction success. Zoo biology, 32(3), 332-341.

Reed, Kent M., and Ruth B. Phillips. "Molecular characterization and cytogenetic analysis of highly repeated DNAs of lake trout, Salvelinus namaycush." Chromosoma 104, no. 4 (1995): 242-251.

Reid, B. N., Moran, R. L., Kopack, C. J., & Fitzpatrick, S. W. (2021). Rapture-ready darters: Choice of reference genome and genotyping method (whole-genome or sequence capture) influence population genomic inference in Etheostoma. Molecular Ecology Resources, 21(2), 404-420.

Reinecke, M. "Influences of the environment on the endocrine and paracrine fish growth hormone–insulin-like growth factor-I system." Journal of fish biology 76, no. 6 (2010): 1233-1254.

Renaut, S., and L. Bernatchez. "Transcriptome-wide signature of hybrid breakdown associated with intrinsic reproductive isolation in lake whitefish species pairs (Coregonus spp. Salmonidae)." Heredity 106, no. 6 (2011): 1003.

Rezvoy, C., Charif, D., Guéguen, L., & Marais, G. A. (2007). MareyMap: an R-based tool with graphical interface for estimating recombination rates. Bioinformatics, 23(16), 2188-2189.

Ricciardi, A., & Rasmussen, J. B. (1999). Extinction rates of North American freshwater fauna. Conservation biology, 13(5), 1220-1222.

Riley, S. C., He, J. X., Johnson, J. E., O'Brien, T. P., & Schaeffer, J. S. (2007). Evidence of widespread natural reproduction by lake trout Salvelinus namaycush in the Michigan waters of Lake Huron. Journal of Great Lakes Research, 33(4), 917-921.

Rius, M., & Darling, J. A. (2014). How important is intraspecific genetic admixture to the success of colonising populations?. Trends in ecology & evolution, 29(4), 233-242.

Robertson, F. M., Gundappa, M. K., Grammes, F., Hvidsten, T. R., Redmond, A. K., Lien, S., ... Macqueen, D. J. (2017). Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biology, 18(1), 1-14.

Robinson, B. W., and D. Schluter. "Natural selection and the evolution of adaptive genetic variation in northern freshwater fishes." Pages 65-94 in W.W. Taylor, B. Sinervo, and J. Endler (eds) Adaptive Genetic Variation in the Wild (2000). Oxford University Press.

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Molecular Ecology, 28(21), 4737-4754.

Roffler, G. H., Amish, S. J., Smith, S., Cosart, T., Kardos, M., Schwartz, M. K., & Luikart, G. (2016). SNP discovery in candidate adaptive genes using exon capture in a free-ranging alpine ungulate. Molecular Ecology Resources, 16(5), 1147-1164.

Rogers, S. M., and L. Bernatchez. "The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (Coregonus sp. Salmonidae) species pairs." Molecular biology and evolution 24, no. 6 (2007): 1423-1438.

Rohlf, F. James. "tpsDig, digitize landmarks and outlines, version 2.05." Department of Ecology and Evolution, State University of New York at Stony Brook (2005).

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Research, 22(5), 939-946.

Rondeau, E. B., Minkley, D. R., Leong, J. S., Messmer, A. M., Jantzen, J. R., von Schalburg, K. R., ... Koop, B. F. (2014). The genome and linkage map of the northern pike (Esox lucius): conserved synteny revealed between the salmonid sister group and the Neoteleostei. PloS One, 9(7), e102089.

Rothbächer, U., Laurent, M. N., Blitz, I. L., Watabe, T., Marsh, J. L., & Cho, K. W. (1995). Functional conservation of the Wnt signaling pathway revealed by ectopic expression of Drosophila dishevelled in Xenopus. Developmental Biology, 170(2), 717-721.

Rougeux, C., Gagnaire, P. A., Praebel, K., Seehausen, O., & Bernatchez, L. (2019). Polygenic selection drives the evolution of convergent transcriptomic landscapes across continents within a Nearctic sister species complex. Molecular Ecology, 28(19), 4388-4403.

Rowe, K. C., Singhal, S., Macmanes, M. D., Ayroles, J. F., Morelli, T. L., Rubidge, E. M., & Moritz, C. C. (2011). Museum genomics: low-cost and high-accuracy genetic data from historical specimens. Molecular Ecology Resources, 11(6), 1082-1092.

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nature Methods, 17(2), 155-158.

Rüber, Lukas, Erik Verheyen, and Axel Meyer. "Replicated evolution of trophic specializations in an endemic cichlid fish lineage from Lake Tanganyika." Proceedings of the National Academy of Sciences 96, no. 18 (1999): 10230-10235.

Ruzycki, J. R., Beauchamp, D. A., & Yule, D. L. (2003). Effects of introduced lake trout on native cutthroat trout in Yellowstone Lake. Ecological Applications, 13(1), 23-37.

Ryder, R. A., S. R. Kerr, W. W. Taylor, and P. A. Larkin. "Community consequences of fish stock diversity." Canadian Journal of Fisheries and Aquatic Sciences 38, no. 12 (1981): 1856-1866.

Ryman, N., & Laikre, L. (1991). Effects of supportive breeding on the genetically effective population size. Conservation Biology, 5(3), 325-329.

Salvesen, K. E. (2015). Lake Trout Restoration in the Great Lakes: From Hatchery to Natural Reproduction. M.S. Thesis, Pennsylvania State University.

Salzberg, S. L. (2019). Next-generation genome annotation: we still struggle to get it right. Genome Biology, 20, 92.

Sambrook, J., & Russell, D. W. (2001). Molecular Cloning -Vol. 1, 2, 3. Cold Spring Harbor, NY: Cold Springs Harbour Laboratory Press.

Sankararaman, S., Sridhar, S., Kimmel, G., & Halperin, E. (2008). Estimating local ancestry in admixed populations. The American Journal of Human Genetics, 82(2), 290-303.

Santure, Anna W., Jocelyn Poissant, Isabelle De Cauwer, Kees Van Oers, Matthew R. Robinson, John L. Quinn, Martien AM Groenen, Marcel E. Visser, Ben C. Sheldon, and Jon Slate. "Replicated analysis of the genetic architecture of quantitative traits in two wild great tit populations." Molecular Ecology 24, no. 24 (2015): 6148-6162.

Sard, N. M., Smith, S. R., Homola, J. J., Kanefsky, J., Bravener, G., Adams, J. V., ... & Scribner, K. T. (2020). RAPTURE (RAD capture) panel facilitates analyses characterizing sea lamprey reproductive ecology and movement dynamics. Ecology and Evolution, 10(3), 1469-1488.

Sauvage, Christopher, Marie Vagner, Nicolas Derôme, Céline Audet, and Louis Bernatchez. "Coding gene single nucleotide polymorphism mapping and quantitative trait loci detection for physiological reproductive traits in brook charr, Salvelinus fontinalis." G3: Genes, Genomes, Genetics 2, no. 3 (2012): 379-392.

Sävilammi, Tiina, Craig R. Primmer, Srinidhi Varadharajan, René Guyomard, Yann Guiguen, Simen R. Sandve, L. Asbjørn Vøllestad, Spiros Papakostas, and Sigbjørn Lien. "The chromosome-level genome assembly of European grayling reveals aspects of a unique genome evolution process within salmonids." G3: Genes, Genomes, Genetics 9, no. 5 (2019): 1283-1294.

Schindler, D. E., Hilborn, R., Chasco, B., Boatright, C. P., Quinn, T. P., Rogers, L. A., & Webster, M. S. (2010). Population diversity and the portfolio effect in an exploited species. Nature, 465(7298), 609-612.

Schluter, D. (1996). Ecological causes of adaptive radiation. The American Naturalist, 148, S40-S64.

Schluter, D. (2000). The ecology of adaptive radiation. OUP Oxford.

Schneider, C. P., Owens, R. W., Bergstedt, R. A., & O'Gorman, R. (1996). Predation by sea lamprey (Petromyzon marinus) on lake trout (Salvelinus namaycush) in southern Lake Ontario, 1982-1992. Canadian Journal of Fisheries and Aquatic Sciences, 53(9), 1921-1932.

Schroeter, J. C., Maloy, A. P., Rees, C. B., & Bartron, M. L. (2020). Fish mitochondrial genome sequencing: expanding genetic resources to support species detection and biodiversity monitoring using environmental DNA. Conservation Genetics Resources, 12(3), 433-446.

Scribner, K., Tsehaye, I., Brenden, T., Stott, W., Kanefsky, J., & Bence, J. (2018). Hatchery strain contributions to emerging wild lake trout populations in Lake Huron. Journal of Heredity, 109(6), 675-688.

Secolin, R., Mas-Sandoval, A., Arauna, L. R., Torres, F. R., de Araujo, T. K., Santos, M. L., ... & Comas, D. (2019). Distribution of local ancestry and evidence of adaptation in admixed populations. Scientific reports, 9(1), 1-12.

Shedko, S. V. (2019). Assembly ASM291031v2 (Genbank: GCA_002910315. 2) identified as assembly of the Northern Dolly Varden (Salvelinus malma malma) genome, and not the Arctic char (S. alpinus) genome. arXiv preprint arXiv:1912.02474.

Siberchicot, A., Bessy, A., Guéguen, L., & Marais, G. A. (2017). MareyMap online: a user-friendly web application and database service for estimating recombination rates using physical and genetic maps. Genome Biology and Evolution, 9(10), 2506-2509.

Sinclair-Waters, M., Ødegård, J., Korsvoll, S. A., Moen, T., Lien, S., Primmer, C. R., & Barson, N. J. (2020). Beyond large-effect loci: large-scale GWAS reveals a mixed large-effect and polygenic architecture for age at maturity of Atlantic salmon. Genetics Selection Evolution, 52(1), 1-11.

Sitar, S. P., Bence, J. R., Johnson, J. E., Ebener, M. P., & Taylor, W. W. (1999). Lake trout mortality and abundance in southern Lake Huron. North American Journal of Fisheries Management, 19(4), 881-900.

Sitar, S. P., Jasonowicz, A. J., Murphy, C. A., & Goetz, F. W. (2014). Estimates of skipped spawning in lean and siscowet lake trout in southern Lake Superior: implications for stock assessment. Transactions of the American Fisheries Society, 143(3), 660-672.

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. Genetics, 195(3), 693-702.

Skulason, Skúli, and Thomas B. Smith. "Resource polymorphisms in vertebrates." Trends in ecology & evolution 10, no. 9 (1995): 366-370.

Smit, A. F. A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0. <http://www.repeatmasker.org>.

Smith, S. H. (1968). Species succession and fishery exploitation in the Great Lakes. Journal of the Fisheries Research Board of Canada, 25: 667-693.

Smith, M. W., & O'Brien, S. J. (2005). Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nature Reviews Genetics, 6(8), 623-632.

Smith, S. R., Amish, S. J., Bernatchez, L., Le Luyer, J., C. Wilson, C., Boeberitz, O., ... & Scribner, K. T. (2020). Mapping of Adaptive Traits Enabled by a High-Density Linkage Map for Lake Trout. G3: Genes, Genomes, Genetics, 10(6), 1929-1947.

Smith, S., Normandeau, E., Djambazian, H., Nawarathna, P., Berube, P., Muir, A., ... & Bernatchez, L. (2021). A chromosome-anchored genome assembly for Lake Trout (Salvelinus namaycush). Molecular Ecology Resources. 00, 1– 16. https://doi.org/10.1111/1755-0998.13483

Snorrason, Sigurdur S., Skúli Skúlason, Bror Jonsson, Hilmar J. Malmquist, Pétur M. Jónasson, Odd Terje Sandlund, and Torfinn Lindem. "Trophic specialization in Arctic charr Salvelinus alpinus (Pisces; Salmonidae): morphological divergence and ontogenetic niche shifts." Biological Journal of the Linnean society 52, no. 1 (1994): 1-18.

Soderlund, C., Nelson, W., Shoemaker, A., & Paterson, A. (2006). SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome Research, 16(9), 1159-1168.

Soderlund, C., Bomhoff, M., & Nelson, W. M. (2011). SyMAP v3. 4: a turnkey synteny system with application to plant genomes. Nucleic Acids Research, 39(10), e68-e68.

Spruell, P., K. L. Pilgrim, B. A. Greene, C. Habicht, K. L. Knudsen, K. R. Lindner, J. B. Olsen, G. K. Sage, J. E. Seeb, and F. W. Allendorf. "Inheritance of nuclear DNA markers in gynogenetic haploid pink salmon." Journal of Heredity 90, no. 2 (1999): 289-296.

Staats, M., Erkens, R. H., van de Vossenberg, B., Wieringa, J. J., Kraaijeveld, K., Stielow, B., ... & Bakker, F. T. (2013). Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. PloS One, 8(7), e69189.

Stafford, C. P., McPhee, M. V., Eby, L. A., & Allendorf, F. W. (2014). Introduced lake trout exhibit life history and morphological divergence with depth. Canadian Journal of Fisheries and Aquatic Sciences, 71(1), 10-20.

Stahlke, A., Bell, D., Dhendup, T., Kern, B., Pannoni, S., Robinson, Z., ... & Luikart, G. (2020). Population genomics training for the next generation of conservation geneticists: ConGen 2018 Workshop. Journal of Heredity, 111(2), 227-236.

Stebbins, G. Ledyard. "The inviability, weakness, and sterility of interspecific hybrids." In Advances in genetics, vol. 9, pp. 147-215. Academic Press, 1958.

Stetz, J. B., Smith, S., Sawaya, M. A., Ramsey, A. B., Amish, S. J., Schwartz, M. K., & Luikart, G. (2016). Discovery of 20,000 RAD-SNPs and development of a 52-SNP array for monitoring river otters. Conservation Genetics Resources. 8 (3): 299-302.

Strait, J. T., Eby, L. A., Kovach, R. P., Muhlfeld, C. C., Boyer, M. C., Amish, S. J., ... & Luikart, G. (2021). Hybridization alters growth and migratory life-history expression of native trout. EvolutionaryApplications, 14(3), 821-833.

Stronen, A. V., Iacolina, L., & Ruiz-Gonzalez, A. (2019). Rewilding and conservation genomics: How developments in (re) colonization ecology and genomics can offer mutual benefits for understanding contemporary evolution. Global Ecology and Conservation, 17, e00502.

Sutherland, Ben JG, Thierry Gosselin, Eric Normandeau, Manuel Lamothe, Nathalie Isabel, Celine Audet, and Louis Bernatchez. "Salmonid chromosome evolution as revealed by a novel method for comparing RADseq linkage maps." Genome biology and evolution 8, no. 12 (2016): 3600-3617.

Sutherland, Ben JG, Ciro Rico, Céline Audet, and Louis Bernatchez. "Sex chromosome evolution, heterochiasmy, and physiological QTL in the salmonid brook charr Salvelinus fontinalis." G3: Genes, Genomes, Genetics 7, no. 8 (2017): 2749-2762.

Syslo, J. M., Guy, C. S., & Cox, B. S. (2013). Comparison of harvest scenarios for the cost-effective suppression of Lake Trout in Swan Lake, Montana. North American Journal of Fisheries Management, 33(6), 1079-1090.

Tallmon, D. A., Waples, R. S., Gregovich, D., & Schwartz, M. K. (2012). Detecting population recovery using gametic disequilibrium-based effective population size estimates. Conservation Genetics Resources, 4, 987–989.

Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E. G., & Risch, N. J. (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. The American Journal of Human Genetics, 81(3), 626-633.

Tang, H., Lyons, E., Pedersen, B., Schnable, J. C., Paterson, A. H., & Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics, 12(1), 1-11.

Teixeira, J. C., & Huber, C. D. (2021). The inflated significance of neutral genetic diversity in conservation genetics. Proceedings of the National Academy of Sciences, 118(10).

Thibaud-Nissen, F., DiCuccio, M., Hlavina, W., Kimchi, A., Kitts, P. A., Murphy, T. D., ... Souvorov, A. (2016). P8008 The NCBI Eukaryotic Genome Annotation Pipeline. Journal of Animal Science, 94(suppl_4), 184-184.

Thomas, Stephen M., Martin J. Kainz, Per-Arne Amundsen, Brian Hayden, Sami J. Taipale, and Kimmo K. Kahilainen. "Resource polymorphism in European whitefish: Analysis of fatty acid profiles provides more detailed evidence than traditional methods alone." PloS one 14, no. 8 (2019): e0221338.

Thompson, N. F., Anderson, E. C., Clemento, A. J., Campbell, M. A., Pearse, D. E., Hearsey, J. W., ... & Garza, J. C. (2020). A complex phenotype in salmon controlled by a simple change in migratory timing. Science, 370(6516), 609-613.

Thorgaard, G. H., Allendorf, F. W., & Knudsen, K. L. (1983). Gene-centromere mapping in rainbow trout: high interference over long map distances. Genetics, 103(4), 771-783.

Thurston, Claude E. "Physical characteristics and chemical composition of two subspecies of lake trout." Journal of the Fisheries Board of Canada 19, no. 1 (1962): 39-44.

Törönen, P., Medlar, A., & Holm, L. (2018). PANNZER2: a rapid functional annotation web server. Nucleic acids research, 46(W1), W84-W88.

Tronstad, L. M., Hall Jr, R. O., & Koel, T. M. (2015). Introduced lake trout alter nitrogen cycling beyond Yellowstone Lake. Ecosphere, 6(11), 1-24.

Valiquette, E., Perrier, C., Thibault, I., & Bernatchez, L. (2014). Loss of genetic integrity in wild lake trout populations following stocking: insights from an exhaustive study of 72 lakes from Québec, Canada. Evolutionary Applications, 7(6), 625-644.

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. Nature Reviews Genetics, 18(7), 411.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. Journal of dairy science, 91(11), 4414-4423.

Veale, A. J., & Russello, M. A. (2017). An ancient selective sweep linked to reproductive life history evolution in sockeye salmon. Scientific Reports, 7(1), 1-10.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics, 33(14), 2202-2204.

Wada, H., Naruse, K., Shimada, A., & Shima, A. (1995). Genetic linkage map of a fish, the Japanese medaka Oryzias latipes. Molecular marine biology and biotechnology, 4(3), 269-274.

Waples, R. S., & Do, C. H. I. (2008). LDNE: a program for estimating effective population size from data on linkage disequilibrium. Molecular Ecology Resources, 8(4), 753-756.

Waples, R.K., Seeb, L.W. and Seeb, J.E., 2016. Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (Oncorhynchus keta). Molecular Ecology Resources, 16(1), pp.17-28.

Waples, R. S., Naish, K. A., & Primmer, C. R. (2020). Conservation and management of salmon in the age of genomics. Annual Review of Animal Biosciences, 8, 117-143.

Waters, Thomas F. The Superior North Shore: A Natural History of Lake Superior's Northern Lands and Waters. U of Minnesota Press, 1987.

Waters, C. D., Clemento, A., Aykanat, T., Garza, J. C., Naish, K. A., Narum, S., & Primmer, C. R. (2021). Heterogeneous genetic basis of age at maturity in salmonid fishes. Molecular Ecology, 30(6), 1435-1456.

Watson, M., & Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. Nature Biotechnology, 37(2), 124-126.

Webb, P. W. "Body form, locomotion and foraging in aquatic vertebrates." American Zoologist 24, no. 1 (1984): 107-120.

Weise, E. M., Scribner, K. T., Adams, J. V., Boeberitz, O., Jubar, A., Bravener, G., Johnson, N., Robinson, J. D. 2021. Applying effective breeding estimates and family structure to assess invasive sea lamprey populations. Evolutionary Applications. In revision.

Wellenkamp, W., He, J. X., & Vercnocke, D. (2015). Using maxillae to estimate ages of Lake Trout. North American Journal of Fisheries Management, 35(2), 296-301.

Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. Trends in Ecology & Evolution, 33(6), 427-440.

Whibley, A., Kelley, J., & Narum, S. (2020). The changing face of genome assemblies: guidance on achieving high-quality reference genomes. Molecular Ecology Resources, 21(3), 641-652.

Whiteley, Andrew R. "Trophic polymorphism in a riverine fish: morphological, dietary, and genetic analysis of mountain whitefish." Biological Journal of the Linnean Society 92, no. 2 (2007): 253-267.

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Computational Biology, 13(6), e1005595.

Wickham, H., & Wickham, M. H. (2007). The ggplot package. Google Scholar. http://ftp. uni-bayreuth. de/math/statlib/R/CRAN/doc/packages/ggplot. pdf.

Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. The American Journal of Human Genetics, 76(5), 887-893.

Williams, D., Trimble, W. L., Shilts, M., Meyer, F., & Ochman, H. (2013). Rapid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes. BMC Genomics, 14(1), 1-11.

Willoughby, J. R., Harder, A. M., Tennessen, J. A., Scribner, K. T., & Christie, M. R. (2018). Rapid genetic adaptation to a novel environment despite a genome-wide reduction in genetic diversity. Molecular Ecology, 27(20), 4041-4051.

Wilson, Chris C., and Nicholas E. Mandrak. "History and evolution of lake trout in Shield lakes: past and future challenges." Boreal Shield watersheds: lake trout ecosystems in a changing environment (2004): Pages 21-35 in J. Gunn, R. Steedman, and R. Ryder (eds). Boreal Shield Watersheds: Lake Trout Ecosystems in a Changing Environment. Lewis/CRC Press.

Wilson, Chris, and David Evans. "Algonquin's silver lake trout: highlighting the history, habitat, and concerns for a unique biodiversity element." (2010). Ontario Ministry of Natural Resources, Peterborough ON.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., & Andrews, S. (2015). HiCUP: pipeline for mapping and processing Hi-C data. F1000Research, 4.

Wodarz, A., & Nusse, R. (1998). Mechanisms of Wnt signaling in development. Annual Review of Cell and Developmental Biology, 14(1), 59-88.

Wood, C. C., Bickham, J. W., John Nelson, R., Foote, C. J., & Patton, J. C. (2008). Recurrent evolution of life history ecotypes in sockeye salmon: implications for conservation and future evolution. Evolutionary Applications, 1(2), 207-221.

Woram, Rachael A., Karim Gharbi, Takashi Sakamoto, Bjorn Hoyheim, Lars-Erik Holm, Kerry Naish, Colin McGowan et al. "Comparative genome analysis of the primary sex-determining locus in salmonid fishes." Genome research 13, no. 2 (2003): 272-280.

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., ... Timp, W. (2019). Nanopore native RNA sequencing of a human poly (A) transcriptome. Nature Methods, 16(12), 1297-1305.

Wu, Xufeng S., Andreas Masedunskas, Roberto Weigert, Neal G. Copeland, Nancy A. Jenkins, and John A. Hammer. "Melanoregulin regulates a shedding mechanism that drives melanosome transfer from melanocytes to keratinocytes." Proceedings of the National Academy of Sciences 109, no. 31 (2012): E2101-E2109.

Yan, Biao, Ban Liu, Chang-Dong Zhu, Kang-Le Li, Li-Jia Yue, Jin-Liang Zhao, Xiao-Ling Gong, and Cheng-Hui Wang. "microRNA regulation of skin pigmentation in fish." J Cell Sci 126, no. 15 (2013): 3401-3408.

Yano, Ayaka, Barbara Nicol, Elodie Jouanno, Edwige Quillet, Alexis Fostier, René Guyomard, and Yann Guiguen. "The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids." Evolutionary applications 6, no. 3 (2013): 486-496.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., & Wang, J. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science, 329(5987), 75-78.

Zimin, A. V., & Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Computational Biology, 16(6), e1007981.

Zimmerman, M. S., Krueger, C. C., & Eshenroder, R. L. (2006). Phenotypic diversity of lake trout in Great Slave Lake: differences in morphology, buoyancy, and habitat depth. Transactions of the American Fisheries Society, 135(4), 1056-1067.

Zimmerman, Mara S., Charles C. Krueger, and Randy L. Eshenroder. "Morphological and ecological differences between shallow-and deep-water lake trout in Lake Mistassini, Quebec." Journal of Great Lakes Research 33, no. 1 (2007): 156-169.

Zimmerman, M. S., Schmidt, S. N., Krueger, C. C., Vander Zanden, M. J., & Eshenroder, R. L. (2009). Ontogenetic niche shifts and resource partitioning of lake trout morphotypes. Canadian Journal of Fisheries and Aquatic Sciences, 66(6), 1007-1018.

Zimmerman, M. S., & Krueger, C. C. (2009). An ecosystem perspective on re-establishing native deepwater fishes in the Laurentian Great Lakes. North American Journal of Fisheries Management, 29(5), 1352-1371

Wada, H., Naruse, K., Shimada, A., & Shima, A. (1995). Genetic linkage map of a fish, the Japanese medaka Oryzias latipes. Molecular marine biology and biotechnology, 4(3), 269-274.