Using genetic tools to inform management and study local adaptation in Pacific salmon

Wesley Alan Larson

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

Doctor of Philosophy

University of Washington
2015

Reading Committee:
Lisa Seeb, chair
Jim Seeb
Lorenz Hauser

Program authorized to offer degree:
School of Aquatic and Fishery Sciences

University of Washington

Abstract

Using Genetic Tools to Inform Management and Study Local Adaptation in Pacific Salmon

Wesley A. Larson

Chair of the supervisory committee:
Research Professor Lisa Seeb
School of Aquatic and Fishery Sciences

Genetic analysis represents a powerful tool for informing management and studying adaptation in wild populations. For example, genetic tools can be used to delineate conservation units, assign individuals of unknown ancestry back to their populations of origin, and identify genes that are important for local adaptation. The overall goal of my thesis was to apply genetic tools to improve population-specific management and identify the genetic basis of local adaptation in Pacific salmon. Pacific salmon (*Oncorhynchus sp.*) return to their natal spawning habitats with high fidelity, promoting the formation of distinct populations that are highly adapted to their local environment. Pacific salmon are also an extremely important economic, cultural, and subsistence resource. These characteristics make Pacific salmon ideal candidates for population-specific management and facilitate the study of local adaptation.

My dissertation consists of six chapters divided into two major themes. The first three chapters focus on applied research questions aimed at developing and utilizing genetic tools to improve management of Chinook salmon (*Oncorhynchus tshawytscha*), and the last three chapters focus on understanding the genetic basis of local adaptation in sockeye salmon (*Oncorhynchus nerka*). In chapter one, we used an existing genetic baseline to elucidate the migration patterns of Chinook salmon in the marine environment. Chapters two and three explored the use of genomics in a management

context, applying data from thousands of genetic markers to develop novel resources that will aid in the conservation of Chinook salmon from western Alaska. For chapter four, we investigated patterns of selection at the major histocompatibility complex (MHC) in populations of sockeye salmon from the Wood River basin in southwestern Alaska. In chapter five, we constructed a genetic linkage map and conducted QTL analysis in five families of sockeye salmon. Finally, in chapter six we merged the linkage map with population data to study the genomic basis of adaptive divergence among three ecotypes of sockeye salmon from the Wood River basin. Taken together, these studies highlight the utility of genetic tools, especially genomics, for improving management and studying local adaptation in Pacific salmon.

Acknowledgements

I could not have completed this dissertation without help from a number of people. My advisor Lisa Seeb and unofficial co-advisor Jim Seeb have been instrumental in my growth as a scientist. They have provided excellent feedback on my work and have always encouraged new research questions. I would also like to thank Lorenz Hauser, Daniel Schindler, and Willie Swanson for serving on my committee and providing excellent feedback on my research.

I feel privileged to be part of the Seeb lab and would like to thank my amazing colleagues for their academic and personal support. Thanks to Eleni Petrou, Caroline Storer, Meredith Everett, Daniel Gomez-Uchida, Marissa Jones, Sewall Young, Morten Limborg, Ryan Waples, Tyler Dann, Carolyn Tarpey, and Garrett McKinney. I would also like to give a special thanks to Carita Pascal who has been an enormous resource in the lab and a great sounding board for personal and academic topics. Additionally, I would like to thank Fred Utter for providing excellent feedback on my work and for helping me to grow as a scientist.

I would also like to thank everyone who helped me conduct research for five summers at the Alaska Salmon Program field camps, especially Daniel Schindler, Jackie Carter, Chris Boatright, Ray Hilborn, Tom Quinn, Peter Lisi, Kale Bentley, Tim Cline, Jonny Armstrong, Rachel Hovel, and KathiJo Jankowski. Spending time at these field camps has been an amazing experience.

I could not have completed my dissertation without help from the Alaska Department of Fish and Game. They provided a huge portion of the samples I used in my dissertation and contributed extremely important local and technical knowledge to my research. I would especially like to thank Bill Templin, Chris Habicht, Tyler Dann, Nick DeCovich, Andy Barclay, Judy Burger, and Heather Hoyt.

Funding for my dissertation was provided by the National Science Foundation, the Alaska Sustainable Salmon Fund, and SAFS. I would also like to thank the administrative staff at SAFS, especially Scott Schafer, Amy Fox, Robin Weigel, Rachel Faircloth, and Kathryn Stout, for helping me navigate grad school.

Finally, I would like to thank my friends and family for providing extensive emotional support for this endeavor. My parents Keith Larson and Karen Lindvall-Larson have always supported me unconditionally. I would especially like to thank my wife Kristen Gruenthal for being an amazing life partner in every way, and my son Mattias for a smile that always fills me with joy.

Table of Contents

Introduction

Genetic tools have been used to inform management of Pacific salmon *(Oncorhynchus sp.)* for over four decades (Utter 2004). In particular, a genetic tool termed genetic stock identification (GSI) is often employed to assign fish of unknown ancestry back to their population of origin (Shaklee *et al.* 1999). This tool is particularly valuable for salmon because the catch in many salmon fisheries is composed of multiple stocks. GSI can be used to determine the proportion of each stock caught in the fishery to ensure that certain stocks are not overharvested (e.g., Dann *et al.* 2013). Additionally, GSI can help to elucidate stock-specific migration and distribution patterns (Habicht *et al.* 2010).

Historically, GSI has been conducted with panels of between 10 and 100 neutral genetic markers. These panels work well when populations are moderately or highly differentiated, but when populations show little divergence at neutral markers, GSI is generally not feasible. Genomics provides a potential solution to this problem. Genomic techniques now make it possible to screen thousands of markers in hundreds of individuals (Tonsor 2012). Additionally, genomic data facilitate the creation of high-density linkage maps that assign markers to specific genomic locations (Stinchcombe & Hoekstra 2008). These tools allow researchers to scan large portions of the genome to look for loci displaying high levels of divergence between populations (Funk *et al.* 2012). For example Bradbury *et al.* (2013) discovered several genomic regions that were highly diverged in Atlantic cod (*Gadus morhua*) that can be used to define new management units, and Miller *et al.* (2012) identified a highly diverged genomic region related to growth rate in rainbow trout (*Oncorhynchus mykiss*). Markers from this genomic region also show high levels of divergence between resident and migratory forms of rainbow trout (Pearse *et al.* 2014) and will likely be incorporated in to future management programs.

Genomic tools also facilitate studies investigating the genetic basis of local adaptation (Allendorf *et al.* 2010). These studies have begun to reveal the genomic architecture of local adaptation during early speciation. In particular, multiple studies have found that divergence during early speciation is localized in relatively few genomic regions known as genomic islands (Via & West 2008; Nosil *et al.* 2009). The

mechanisms that create these islands are still unclear. Nevertheless, these island represent excellent candidates for future research and likely harbor many important genes that are involved in local adaptation in wild populations.

This dissertation represents a transition from genetic datasets (< 100 markers), to genomic datasets (over 10,000 markers). In the first chapter we used an existing baseline containing 43 SNPs to investigate distribution and migration patterns of Chinook salmon on the high seas. Chapters two and three employed restriction-site-associated DNA (RAD) sequencing to investigate population structure and improve GSI in Chinook salmon from western Alaska. Chapter four involved sequencing a 350 base pair region of the major histocompatibility complex (MHC) in sockeye salmon to investigate the role of pathogen mediated selection in the Wood River basin. In chapter five, we constructed a genetic linkage map for sockeye salmon as well as other important genomic resource for future studies and conducted QTL analysis. Finally, in chapter six we used the linkage map along with RAD data and data from high-throughput assays to investigate the genomic basis of adaptive divergence in 14 populations of sockeye salmon. Taken together, this research illustrates the utility of genomic data for informing conservation and studying local adaptation in salmon and will serve as an excellent guide for future research in these areas.

Chapter 1

Single-nucleotide polymorphisms reveal distribution and migration of Chinook salmon (*Oncorhynchus tshawytscha*) in the Bering Sea and North Pacific Ocean[1]

Abstract

We genotyped Chinook salmon (*Oncorhynchus tshawytscha*) from the Bering Sea and North Pacific Ocean for 43 single-nucleotide polymorphisms (SNPs) to investigate seasonal distribution and migration patterns. We analyzed 3 573 immature fish from 22 spatiotemporal strata; composition analyses were performed using genotype data from spawning stocks spanning the species range. Substantial variation in stock composition existed among spatial and seasonal strata. We inferred patterns of seasonal migration based upon these data along with data from previous tag, scale, and parasite studies. We found that stocks from western Alaska and Yukon River overwinter on the Alaska continental shelf then travel to the middle and western Bering Sea during spring-fall. Stocks from California to Southeast Alaska are distributed in Gulf of Alaska year-round with a substantial portion of this group migrating northward to the eastern Bering Sea during spring-fall. Proportions of Russian stocks increase when moving east to west in both the Bering Sea and North Pacific Ocean. These data can be used to better understand the impacts of fisheries and climate change on this valuable resource.

Introduction

Knowledge of the dynamics of migration and harvest of marine organisms is necessary to enable adequate conservation and management (Ruggerone & Goetz 2004; Welch *et al.* 2011). This critical need remains largely unfulfilled because of factors that include the vast and largely invisible oceanic domain, indiscriminate harvest, common ownership, and indistinct boundaries (Myers *et al.* 2007). Despite these challenges some

[1] Full citation: Larson, W. A., F. M. Utter, K. W. Myers, W. D. Templin, J. E. Seeb, C. M. Guthrie, A. V. Bugaev, and L. W. Seeb. 2013. Single-nucleotide polymorphisms reveal distribution and migration of Chinook salmon (*Oncorhynchus tshawytscha*) in the Bering Sea and North Pacific Ocean. *Canadian Journal of Fisheries and Aquatic Sciences* **70**:128-141. Table A1 available from the online version of this manuscript (http://www.nrcresearchpress.com/doi/abs/10.1139/cjfas-2012-0233).

remedial management programs have been implemented (Worm *et al.* 2006; Gunderson 2011). Inherent in such management is identifying the distribution of distinct breeding groups known as stocks in multistock mixtures (Begg *et al.* 1999; Sagarin *et al.* 2009). Discrete management of these stocks will preserve genetic diversity, thereby reducing interannual variation in fishery yield and facilitating adaptation to a variety of environmental stressors including climate change (Hilborn *et al.* 2003; Schindler *et al.* 2010). Discrete management of stocks is especially important for Pacific salmonids (*Oncorhynchus spp.*) that migrate across international borders and through multiple fisheries.

Our understanding of the distribution of Pacific salmon stocks on the high seas is facilitated by their strong homing ability, which has generated discrete and stable stock structure, commonly partitioned within species into more or less geographic hierarchies (e.g., Seeb *et al.* 2004; Neville *et al.* 2006; Habicht *et al.* 2010). This structure creates sets of locally adapted stocks which can be distinguished based on unique traits (Quinn 2005). Historically, procedures for discriminating oceanic salmonid stocks based on these unique traits have included scale analyses (Major *et al.* 1978; Davis *et al.* 1990), parasite presence (e.g., Margolis 1963; Urawa *et al.* 1998) and genetic data (Beacham *et al.* 1985; Seeb *et al.* 2000; Yoon *et al.* 2009). The temporal stability of genetic structure and resulting baselines (Gomez-Uchida *et al.* 2012; Iwamoto *et al.* 2012) contrasts with updating of standards required for non-genetic traits (Pella & Milner 1987), and has stimulated increasing application of genetic data when managing many stock-structured salmon fisheries (Habicht *et al.* 2007; Gauthier-Ouellet *et al.* 2009; Beacham *et al.* 2011).

Use of genetics to manage Pacific salmonids is based on observed allelic frequencies of baseline stocks to infer the natal origin of fish captured in mixed-stock fisheries (Utter & Ryman 1993). This technique, termed genetic stock identification, is routinely applied by management agencies throughout the Pacific Rim to gain insight into oceanic migrations of Pacific salmonids (Seeb & Crane 1999a; Beacham *et al.* 2006; Moriya *et al.* 2007). More recently, the potential for broader oceanic genetic stock identification estimations has become feasible as species-wide SNP baselines have been assembled in chum salmon (*O. keta*) (Seeb *et al.* 2011c), sockeye salmon (*O. nerka*)

(Habicht *et al.* 2010; Gomez-Uchida *et al.* 2012), and Chinook salmon (*O. tshawytscha*) (Templin *et al.* 2011).

Chinook salmon typify the biological complexity of anadromous salmonids with a diversity of largely stock-specific and adaptive life history variations (Healey 1991; Quinn 2005). Early life histories of some stocks are characterized by seaward (ocean-type) migrations following emergence from gravel. In other stocks, juveniles remain near natal environments throughout their first year (stream-type) prior to smoltification and marine migration. During their oceanic feeding migrations immature Chinook salmon are distributed in epipelagic and mesopelagic waters over the continental shelves (200-m depth contour) and deep ocean basins. Variable maturation schedules at one extreme include some males that mature within one year without reaching the sea and extend to fish of both sexes spending up to six years in the marine environment prior to returning to freshwater to spawn.

Genetic stock identification estimates have revealed somewhat fragmented patterns of Chinook salmon distribution. For example, stock-composition estimates from the west coast of Vancouver Island, British Columbia, were mostly composed of US-origin salmon, contrasting with estimates from the Queen Charlotte Islands, British Columbia, which contained mostly Canadian salmon (Beacham *et al.* 2006). Estimates of ocean age 1 fish from the North Pacific Ocean extended distribution patterns inferred from coded-wire tagging data (Trudel *et al.* 2009; Tucker *et al.* 2011), where oceanic distributions varied by region of origin and by age as older fish migrated from coastal shelves to deeper waters. Chinook salmon from throughout the species range were found in bycatch of the US groundfish fishery in the eastern Bering Sea (Templin *et al.* 2011; Guthrie *et al.* 2012). Although fish from western Alaska streams were predominant in the bycatch, large seasonal and interannual variations in stock composition were evident. These studies suggest a complex interaction of spatial, temporal, and developmental factors in determining ocean migration patterns.

This study applies the species-wide SNP baseline of Templin *et al.* (2011; Table 1.1; Fig. 1.1) to estimate distributions and migrations of Chinook salmon sampled in the Bering Sea and the North Pacific Ocean. Providing greater geographic and stock-specific accuracy and precision than earlier non-genetic estimates of stock mixtures in these

regions (Major *et al.* 1978; Myers *et al.* 1987), we examine temporal and spatial distributions over this broad geographic range. We augment our data with relevant tag, otolith and parasite studies to hypothesize stock-specific distributions and migration patterns.

Methods

Collection of immature Chinook salmon

We obtained oceanic samples of immature Chinook salmon (N = 3 923) from three sources during 2005-2011: (1) U.S. National Oceanic and Atmospheric Administration (NOAA) observers enumerating salmon bycatch in US commercial groundfish fisheries in the southeastern Bering Sea and western Gulf of Alaska, (2) Japanese salmon research cruises (R/V *Wakatake maru* and R/V *Hokko maru*) in the central Bering Sea and North Pacific Ocean, and (3) a Russian research cruise (R/V *TINRO*) in the western Bering Sea and northwestern North Pacific Ocean (Table 1.2, Fig. 1.2). Japanese and Russian cruises were part of the cooperative efforts of North Pacific Anadromous Fish Commission nations who were conducting the Bering-Aleutian Salmon International Survey. More than than 90% of the sample material were smears of scales dried in coin envelopes; approximately 370 fin clips, preserved in 95% ethanol, were also included. Length, weight, date of capture, and location of capture (GPS coordinates for each haul) were recorded for each fish. Saltwater age was determined using a microfiche reader to count annual winter growth patterns (annuli) on scales (e.g., Davis *et al.* 1990) for all but approximately 200 samples from the *Hokko maru*. We do not report freshwater age due to the large number of scales with regenerated freshwater growth zones and the difficulty of accurately distinguishing between subyearling (freshwater age 0) and yearling (freshwater age 1) scale growth patterns in ocean samples of Chinook salmon. Ocean age is designated by an Arabic numeral; e.g., "ocean age 1" is a fish with one ocean annulus on its scale. We do not have maturity data for the NOAA observer samples, but the relatively small number of ocean age 4 and ocean age 5 fish in our samples leads us to believe that most were immature.

Laboratory analysis

We genotyped all fish using the panel of 43 SNPs that were in linkage equilibrium described in Templin *et al.* (2011). Prior to genotyping, genomic DNA was extracted

using a DNeasy 96 Tissue Kit (Qiagen, Valencia, CA). DNA from scale samples was then preamplified using the protocol described in Smith *et al.* (2011) to achieve a sufficient template for PCR assays. Genotyping was conducted using TaqMan (Applied Biosystems, Foster City, CA) assays in Biomark 96.96 Dynamic arrays (Fluidigm, South San Francisco, CA) according to the methods of Seeb et al. (2009a). We re-genotyped three out of every 95 (3.1%) samples at all loci to quantify discrepancy rate. Individuals with greater than three missing genotypes were removed from further analysis.

Construction of sample strata

Spatiotemporal strata were created to encapsulate areas of environmental homogeneity and facilitate temporal comparisons while achieving mixture sizes of 100 or greater (Table 1.2, Fig. 1.2, Table A1). Geographic strata generally represented distinct environmental areas such as continental shelves or oceanic basins but, when sampling was semi-continuous across a large area with no clear environmental barriers, we constructed strata based on major latitude or longitude divisions. For example, we stratified samples collected along the Alaskan continental shelf break and Aleutian basin using the 55°N latitude line and the 170 °W longitude line (Fig. 1.2). Collections for a given geographic stratum were then subdivided temporally based on season and year of collection. All strata but one that originated from the western Bering Sea (N = 22) approached or exceeded our minimum target sample size of 100 fish, a target set to maintain sufficient confidence to detect the presence of reporting groups with low contributions (Marlowe & Bucsack 1995). The lack of previous genetic stock identification estimates from this region prompted its inclusion despite the small sample size.

Genetic stock identification

Genetic stock identification was used to assign samples of unknown origin (mixtures) to genetically distinct regional groups (reporting groups) characterized in a genetic baseline. This analysis was conducted in the program BAYES (Pella & Masuda 2001) which uses a Bayesian Markov chain-Monte Carlo (MCMC) method to calculate proportional contributions of each reporting group to the mixture. For all mixture composition estimates, we implemented a flat Dirichlet prior which weighted all reporting groups and all stocks within reporting groups equally. For each estimate, we

ran five MCMC simulations with randomly assigned unique starting values for 50,000 iterations. The first 25,000 iterations were discarded, and the posterior distribution was formed by combining the last 25,000 iterations of each chain. The composition of each mixture was reported as the mean of the posterior distribution, and the 90% credibility interval (CI) was defined as the central 90% of the distribution. To ensure adequate convergence of posterior distributions across all chains for each estimate, shrink factors (Gelman & Rubin 1992) were calculated using BAYES. A shrink factor <1.2 indicates convergence across all MCMC simulations and confirms that the starting values of each simulation had no influence on the end result (see Kass *et al.* 1998).

The baseline dataset used for genetic stock identification was identical to that of Templin *et al.* (2011) with one modification (Fig. 1.1, Table 1.1): we removed the Meshik River, a small North Alaska Peninsula stock with the smallest sample size in the baseline (N = 43, baseline average N = 135). Preliminary estimates from the winter 2006 eastern Bering Sea shelf stratum showed shrink factors over 15 for the coastal western Alaska and north Alaska Peninsula reporting groups. Further investigation revealed that approximately 70% of the mixture was being assigned to the Meshik River. After removing this stock, shrink factors for this estimate dramatically improved, suggesting that this stock was not accurately characterized.

Nine of the 11 regional (subsequently referred to as fine-scale) reporting groups used for genetic stock identification were identical to those of Templin *et al.* (2011) who extensively evaluated their performance using proof tests from mixtures of known origin (Fig. 1.1). Two exceptions, the Southeast Alaska and British Columbia reporting groups, failed to converge in multiple estimates; this was likely due to the lack of genetic differentiation between the southern stocks in the Southeast Alaska group and the northern stocks in the British Columbia group. We used the raw genotyped data from Templin *et al.* (2011) to redesign these two reporting groups to more accurately reflect patterns of genetic differentiation: (1) Southeast Alaska and North British Columbia (Southeast Alaska stocks included with stocks from northern British Columbia: Nass, Skena, coastal northern British Columbia), (2) South British Columbia (southern British Columbia stocks: central and south British Columbia mainland, Vancouver Island, Thompson and Fraser Rivers).

To increase the accuracy of mixtures with sample sizes less than 100, we combined multiple fine-scale reporting groups to form two sets of broad-scale reporting groups (Table 1.3). One set was constructed for Gulf of Alaska estimates and another for estimates from the Bering Sea and western North Pacific Ocean. Each set contained four broad-scale reporting groups which were formed by pooling geographically proximate or sparsely represented stocks from the original 11 groups (Table 1.3). For example, we pooled all stocks from rivers draining into the Bering Sea in our Gulf of Alaska estimates because those stocks did not contribute more than 10% to any estimate in this region. Similarly, we pooled all stocks from rivers south of the Alaska Peninsula in estimates from the Bering Sea and western North Pacific Ocean because these reporting groups did not contribute significantly to mixtures in this region (see Table 1.3 for further description of broad-scale reporting groups). We did, however, use the full 11 fine-scale reporting groups on the eastern Bering Sea shelf where we had the largest sample sizes.

Results

Laboratory analysis

Our genotyping success rate was 91% *(*3 563 out of 3 923 samples). Failed genotypes often originated from samples that contained only a single scale which did not provide adequate template for PCR amplification. Additionally, failed genotypes may have been the result of DNA contamination between individuals that occurred on the sampling vessel. Such contamination can cause genotypes to fall outside heterozygote and homozygote clusters during scoring (see Smith *et al.* 2011). For example, if a homozygous individual and a heterozygous individual are combined in a contaminated sample, the genotype will appear intermediate. Our genotyping discrepancy rate, calculated from re-genotyping 95 samples (3.1% of total), was zero (all samples matched previous genotype).

Genetic stock identification

We first evaluated the shrink factor for each estimate as a test of convergence. Sixteen out of 22 estimates had shrink factors < 1.2 for all reporting groups, indicating convergence of all five chains to the same general posterior distribution (see Kass *et al.* 1998). Five estimates had shrink factors between 1.20 and 1.55 for at least one reporting group and did not converge completely (Table 1.4). Despite this imprecision, the

proportional differences for non-converging reporting groups were < 10% between MCMC simulations and are unlikely to affect our overall conclusions. One estimate from the eastern Bering Sea during winter 2006 had shrink factors close to eight for the coastal western Alaska and north Alaska Peninsula reporting groups (Table 1.4). Lack of convergence in this estimate is likely due to the presence of genetically intermediate stocks in the mixture that are not represented in the baseline. Regardless of this lack of convergence, both stock groups compose significant portions of each MCMC simulation. The clear presence of these stock groups in the mixture is therefore useful information to consider when postulating large scale migration and distribution patterns and prompted their inclusion.

Overview of mixtures

Extensive variability of stock composition estimates existed among geographic strata with estimates generally demonstrating large contributions from baseline stocks geographically proximate to sampling locations (Table 1.4, Fig. 1.3). In the broadest sense, the proportion of North American stocks decreased from east to west as Russian fish became the major stock in the western North Pacific Ocean and Bering Sea. Estimates from the Gulf of Alaska geographic stratum were almost completely composed of stocks from California to Southeast Alaska (Table 1.4, Fig. 1.3a). Despite their geographic proximity to the sampling area, relatively low contributions were estimated for stocks from the Gulf of Alaska, contrasting the general trend observed for most other stock-groups.

Coastal western Alaska stocks (i.e., Bristol Bay, Kuskokwim River, lower Yukon River, and Norton Sound) were the most prevalent stocks encountered in the Bering Sea, composing approximately 60% of mixtures from this area (Table 1.4, Fig. 1.3b,c,d). Southern stocks from California to Southeast Alaska and stocks from the north Alaska Peninsula were found in relatively high proportions in the eastern Bering Sea shelf but did not contribute significantly to estimates from any other Bering Sea regions (Table 1.4). In contrast, stocks from the middle and upper Yukon River were rare in the eastern Bering Sea shelf but were present in substantial numbers in all other Bering Sea strata (Table 1.4, Fig. 1.3b,c,d)**.**

Although Russian stocks made up the majority of estimates from the western Bering Sea, Kamchatka Peninsula and Commander Islands, North American stocks were also found in this region. Specifically, fish from coastal western Alaska, the north Alaska Peninsula and the middle and upper Yukon River were encountered in the western Bering Sea, and upper Yukon River fish were encountered in the Commander Islands (Table 1.4, Fig. 1.3d).

Seasonal and annual variability of mixtures

In the eastern Bering Sea shelf, stocks from coastal western Alaska and the north Alaska Peninsula made up a large proportion of mixtures from winter and fall but a much smaller proportion of mixtures from late spring and summer (Table 1.4, Fig. 1.3b). The decrease in proportions of these stocks in late spring and summer corresponded to an increase in stocks from the North Pacific Ocean (mostly fish from California to Southeast Alaska) during this time period. These seasonal patterns were temporally stable across multiple years.

The substantial presence of middle and upper Yukon River stocks on the central Bering Sea shelf during winter and fall contrasted with much smaller estimates for this reporting group in the summer (Table 1.4, Fig. 1.3c). An opposite pattern was observed with more southern stocks from California to Southeast Alaska which were present in moderate numbers on the central Bering Sea shelf during summer but nearly undetected in the fall and winter (Table 1.4). This pattern of increased abundance of southern stocks in the summer months was similar to that observed in estimates from the southeastern Bering Sea.

Discussion

The objective of this study was to use genetic stock identification to significantly improve the understanding of Chinook salmon distribution and migration over a broad and infrequently sampled geographic range. In general, mixture strata were primarily composed of geographically proximate spawning stocks, but there were major exceptions to this trend. For example, stocks from the west coast US periodically contributed significantly to mixtures from the eastern Bering Sea shelf and stocks from the upper Yukon River contributed significantly to mixtures from the Commander Islands. Estimates from this study corroborate patterns of distribution obtained using scale pattern

and coded wire tag estimates while providing additional spatial representation and stock-specific resolution (e.g., Myers *et al.* 1987; Myers *et al.* 2004).

Stock-specific distribution and migration patterns

We constructed graphical synopses of seasonal migration for four major stock groups of Chinook salmon using our data and data from past studies (Fig. 1.4). Stock groups were chosen based on those of Myers et al. (1987) to facilitate the use of data from numerous studies which adopted similar methods (e.g. Myers & Rogers 1988; Bugaev & Myers 2009). These synopses are included as visual generalizations of diverse data and are not meant to be used as quantitative models.

Decreasing proportions of Russian stocks from west to east

Stocks from Russia composed the majority of estimates from the western North Pacific Ocean and western Bering Sea**,** but their numbers generally diminished when moving east of the Russian exclusive economic zone. Despite this decrease, at least a portion these stocks were found in the Aleutian Islands and central Bering Sea basin during summer, where they intermix with fish from western Alaska and the Yukon River. Most fish from this stock group did not migrate as far east as the Alaskan continental shelf; they were undetectable in these areas.

Our data, scale pattern data (Myers *et al.* 1987), and parasite data (Urawa *et al.* 1998) provide evidence that immature Russian Chinook salmon migrate between spring-fall feeding grounds in the western and central Bering Sea basin and overwintering areas in the North Pacific Ocean (Fig. 1.4a,b). Although we do not have DNA data from winter in areas where Russian stocks are common, we hypothesize that Russian Chinook salmon mostly overwinter in the North Pacific Ocean based upon the widely accepted conceptual model of Asian salmon distributions supported by research vessel and catch data (Major *et al.* 1978).

Prevalence of coastal western Alaska stocks in the Bering Sea

Both contemporary genetic data and historical data from scale pattern analysis (Myers *et al.* 1987; Myers & Rogers 1988) show that stocks from western Alaska dominate collections from the Bering Sea. This stock group composed the majority of estimates from the eastern Bering Sea shelf during winter and fall, the central Bering Sea basin during summer (not sampled in winter), and the central Bering Sea shelf in all

seasons. Coastal western Alaska fish were also found in the western Bering Sea in summer, supporting scale pattern analysis results (Bugaev & Myers 2009) and indicating this region may be an important summer-fall feeding area for these stocks. The prevalence of coastal western Alaska fish in our Bering Sea estimates was expected given the stock's large run size (approximately 500 000) and proximity to our sampling strata.

This stock also appears to complete a seasonal migration based on DNA data, coded-wire tag recoveries, and scale pattern data. These data indicate that the spring-fall distribution of this stock is primarily in the Bering Sea over the deep Aleutian Basin, with migrations extending into the Commander Basin in some years; in winter this stock concentrates in the southeastern Bering Sea and Aleutian Islands in waters over Alaskan continental shelf/shelf break (Fig. 1.4c,d; e.g. Myers *et al.* 1987; Myers & Rogers 1988; Bugaev & Myers 2009). This seasonal habitat preference is also demonstrated by stable isotope analysis of muscle tissue from western Alaskan Chinook which revealed depleted levels of ^{13}C in the summer when they are feeding in pelagic (basin) habitats and enriched levels in the winter when they are feeding on the continental shelf/shelf break (Myers *et al.* 2010).

Seasonal westward migration of middle and upper Yukon River Chinook salmon

Chinook salmon from the middle and upper Yukon River overwintered on continental shelves/shelf breaks in the central Bering Sea and then traveled westward into the central Bering Sea basin, western Bering Sea and North Pacific Ocean during summer. Evidence for this movement pattern was provided by estimates from the central Bering Sea shelf which demonstrated larger proportions of middle and upper Yukon River stocks in winter and fall compared to summer. Similar high abundance of middle and upper Yukon River fish in central Bering Sea shelf habitats during winter and fall was also demonstrated by coded-wire tag recoveries of Canadian hatchery stocks (e.g., Myers *et al.* 2004) and by Murphy *et al.* (2009), who conducted genetic stock identification on juvenile (ocean age 0) Chinook salmon caught slightly inshore of our central Bering Sea shelf stratum during fall.

In summer, the highest proportions of middle and upper Yukon River Chinook salmon were found in the central Bering Sea basin. Additional strata with contributions from this stock group during summer included the western Bering Sea, Commander

Islands and Kamchatka Peninsula. We did not detect significant contributions from middle and upper Yukon River fish in samples from the eastern Bering Sea shelf, a geographically proximate stratum to their natal river. This observation, which is supported by genetic estimates from Murphy *et al.* (2009) and tag recoveries (e.g., Myers *et al.* 2004), suggests these stocks do not travel south from their river of origin. Chinook salmon from the upper Yukon River complete one of the longest freshwater migrations of any salmonid (Beacham *et al.* 1989). Although they generally display patterns of distribution similar to stocks from coastal western Alaska, these stocks were encountered in areas of the western Bering Sea and North Pacific Ocean where no other western Alaska stocks were found. It is possible that fish from the Yukon River travel to different areas than western Alaska stocks but further evidence is needed to confirm these patterns.

Widespread but patchy distribution of Gulf of Alaska stocks

Coded-wire tags from Gulf of Alaska stocks have been recovered from Oregon to the Aleutian Basin indicating that the area of distribution for these stocks is quite large. Despite this range, we detected only occasional and relatively small contributions from these stocks in our mixture samples. As expected, contributions from Gulf of Alaska stocks were the largest in our Gulf of Alaska geographic stratum but they were still relatively small compared those of Southeast Alaska and northern British Columbia stocks. The small estimates for the Gulf of Alaska reporting group compared to the Southeast Alaska and northern British Columbia reporting group in this region were not simply a reflection of differences in relative production because both of these broad-scale reporting groups have run sizes of similar magnitude.

Both our results and historical scale pattern analyses demonstrate sporadic and relatively small contributions of Gulf of Alaska stocks to mixtures from the Bering Sea (see Myers & Rogers 1988; Guthrie *et al.* 2012) but larger contributions to estimates from the North Pacific Ocean (see Myers *et al.* 1987). These data suggest that the central North Pacific Ocean, where Chinook salmon are relatively uncommon (e.g., Major *et al.* 1978; Myers *et al.* 1993; Nagasawa & Azumaya 2009), may be a major area of distribution for Gulf of Alaska stocks. In general, Gulf of Alaska stocks contribute

sporadically to mixtures from many regions but appear to be most prevalent in the central North Pacific Ocean and Gulf of Alaska.

Since seasonal distribution data for Gulf of Alaska stocks is not available, our synopsis incorporated a generalized pattern of northern migration in spring-fall and southern migration in winter (Fig. 1.4e,f; Myers *et al.* 1987; Myers *et al.* 2007). This synopsis is the least supported of the four and should be re-examined when additional data become available.

Seasonal northward migration of stocks from California to Southeast Alaska

According to our estimates, a significant portion of stocks from California to Southeast Alaska overwinter in the Gulf of Alaska then travel northward to the continental shelf region of the eastern Bering Sea during spring and summer (Fig. 1.4g,h). This migration pattern, which has been hypothesized for other species of Pacific salmon, is thought to be driven by warm summer temperatures in the Gulf of Alaska which promote northward movement towards the cooler and more productive Alaskan continental shelf (Myers *et al.* 2007). Empirical evidence of this seasonal movement was provided by coded-wire tag data from a few hatchery stocks of Chinook salmon (e.g., Myers *et al.* 1996), but our estimates from the eastern Bering Sea are the first to reliably demonstrate the pattern in multiple stock groups. With temperatures rising in the Gulf of Alaska due to climate change it is possible that this region will become even less hospitable to salmonids during the summer months, increasing the proportion of salmon stocks which spend the summer in the Bering Sea (Myers *et al.* 2007; Abdul-Aziz *et al.* 2011). Future research incorporating similar data could provide direct evidence of shifting salmonid migration patterns in response to climate change.

Comparison to high seas tag, coded-wire tag, and scale pattern analysis data

The databases of historical high seas tag, coded-wire tag, and scale pattern analysis can provide additional information to complement our genetic stock identification data. Coded-wire tag data have been used since the early 1980s to infer stock-specific distribution patterns in the Bering Sea and North Pacific Ocean. Tagging efforts are focused on hatchery production from Southeast Alaska to California but very few fish are tagged from the more remote but plentiful Bering Sea stocks (Nandor *et al.* 2009). Scale pattern analyses estimated stock proportions for stream-type Chinook

salmon from British Columbia and northward but typically did not include any samples from west coast US stocks (Myers *et al.* 1987). Finally, high seas experiments using external tags have been conducted for over half a century by the International North Pacific Fisheries Commission (1954-1992) and North Pacific Anadromous Fish Commission (1993-present) and can provide additional insights although these types of data are often limited by sample size. Therefore, comparisons of estimates from coded-wire tag and scale pattern analyses to those derived from genetic data can be useful but must be interpreted with caution. Coded-wire tag data generally support our findings of Canadian Yukon River fish on the Alaska continental shelf and stocks from Cook Inlet to California in the Gulf of Alaska and on the eastern Bering Sea shelf (Myers *et al.* 2004).

Concordance varies for our estimates and those derived from scale pattern analysis (Myers *et al.* 1987; Myers & Rogers 1988). Estimates were highly concordant for mixtures from the central Bering Sea basin summer but were increasingly divergent for mixtures from the eastern Bering Sea shelf fall and western Bering Sea summer (Myers *et al.* 1987; Myers & Rogers 1988). In the western Bering Sea, evidence from scale pattern analyses suggests considerable interannual variation in the relative proportions of immature Russian stocks during the past five decades, possibly related to variation in year-class strength of these stocks, as well as variation in distribution and abundance of the Chinook salmon forage base (e.g., Major *et al.* 1978; Bugaev & Myers 2009; Zavolokin 2009).

Major findings and future directions

The fine-scale resolution of stock structure from DNA data is continually improving. Hence, the ability to infer stock distributions and migratory patterns is likely to improve in the near future. Expansion of the number of SNP markers and the number of stocks included is ongoing with Chinook and other Pacific salmon (Seeb *et al.* 2011a). Further, recent advances in molecular techniques and DNA sequencing have made it possible to screen hundreds of individuals with thousands of genetic markers (reviewed in Hohenlohe *et al.* 2010; Seeb *et al.* 2011a). Using this technology, it is now possible to develop high throughput assays for markers useful in differentiating genetically similar stocks. Development of these "high-resolution" markers would be especially useful to

accurately distinguish the major drainages in the coastal Western Alaska reporting group, providing fisheries managers with more accurate estimates of stock-specific distributions.

However, salmon research on the high seas is expensive, and tissues for DNA or other analyses are limited. Bycatch samples, such as used in this study, can be a valuable data source; however, sampling locations vary between years and areas due to the inherent unpredictably of the fishing fleet (Stram & Ianelli 2009). Future investigations of salmon distribution and migration will improve with long-term data sets obtained from a combination of bycatch samples and through international agreements with nations conducting research on the high seas.

Our findings significantly improve the understanding of stock-specific distribution and migration of Chinook salmon in the North Pacific Ocean and Bering Sea. In addition to augmenting the results of earlier studies, we provide direct evidence for the previously hypothesized seasonal northward migration of southern (California to Southeast Alaska) stocks in the Bering Sea. We also quantify the abundance of west coast US stocks which was not possible with previous scale pattern analysis studies. Our study demonstrates that genetic stock identification using a species wide baseline can provide accurate stock composition data across a large geographic range. These data can then be used to infer migration patterns and can be incorporated into stock assessment models.

Tables

Table 1.1. Number of stocks and total number of individuals (N) for each fine-scale reporting group included in the SNP baseline. All reporting groups except Southeast Alaska and North British Columbia and South British Columbia are identical to those of (Templin *et al.* 2011). Approximate run size (catch + escapement) of each group is also given.

			Run Size	
Fine-Scale Reporting Group	Stocks	N	Estimate (Thousands)	Source
1. Russia	4	340	180	(Heard *et al.* 2007)
2. Coastal Western Alaska (Coastal W AK)	29	4 520	533	(Baker *et al.* 2006; Molyneaux & Brannian 2006; Heard *et al.* 2007; Howard *et al.* 2009)
3. North Alaska Peninsula (N AK Pen)	6	479	13	(Nelson *et al.* 2006)
4. Middle Yukon River (Mid Yukon)	8	1 097	14[a]	(Howard *et al.* 2009)
5. Upper Yukon River (Up Yukon)	13	1 798	100	(Howard *et al.* 2009)
6. Northwest Gulf of Alaska (NW GOA)	19	3 212	131	(Fair *et al.* 2010)
7. Copper River	11	1 181	80	(Fair *et al.* 2008)
8. Northeast Gulf of Alaska (NE GOA)	7	987	9	(McPherson *et al.* 2003)
9. Southeast Alaska and North British Columbia (SE AK and N BC)	37	5 047	181	(Henderson & Graham 1998; McPherson *et al.* 2003)
10. South British Columbia (S BC)	24	3 060	410	(Henderson & Graham 1998; Heard *et al.* 2007)
11. West Coast US (Coastal US)	14	1 548	967	(Myers *et al.* 1998)
Total	172	23 269	2 606	

[a] Probable underestimate due to lack of escapement data for middle Yukon River tributar

Table 1.2. Mixtures analyzed, sample size, source of mixtures, and months included in mixture. Estimates are stratified by geographic regions as described in Fig. 1.2. Sample source abbreviations are as follows: NOAA groundfish observers (NOAA), Japanese research vessels (Japanese RV) and Russian research vessel (Russian RV).

Mixture	N	Source	Months
Gulf of Alaska			
2006 Winter	219	NOAA	Jan, Feb, March
2005, 2006 Summer	62	NOAA	July, Aug, Sept
2005, 2006 Fall	127	NOAA	October
Eastern Bering Sea Shelf			
2006 Winter	415	NOAA	Jan, Feb, March
2007 Winter	176	NOAA	Jan, Feb, March
2008 Winter	253	NOAA	Jan, Feb, March
2005 June	118	NOAA	June
2006 June	105	NOAA	June
2005 Summer	279	NOAA	July, Aug, Sept
2006 Summer	223	NOAA	July, Aug, Sept
2005, 2006 Fall	234	NOAA	October
Central Bering Sea Shelf			
2007, 2008 Winter	74	NOAA	Jan, Feb, March
2005 Summer	290	NOAA	July, Aug, Sept
2006 Summer	180	NOAA	July, Aug, Sept
2005, 2006, 2008 Fall	141	NOAA	October
Central Bering Sea Basin			
2007, 2008 Summer	89	Japanese RV	June, July
2009 Summer	137	Japanese RV	June, July, Aug
2010, 2011 Summer	76	Japanese RV	July, Aug
Western Bering Sea			
2008 Summer	21	Russian RV	June
Kamchatka Peninsula			
2008 Summer	126	Russian RV	July, Aug
Commander Islands			
2008 Spring	140	Russian RV	May, June
Aleutian Islands			
2006, 2009-2011	78	NOAA and Japanese RV	Feb, June, July, Aug
Total	3563		

Table 1.3. Broad-scale reporting groups for estimates with < 100 fish and the fine-scale reporting groups that they incorporate. Broad-scale reporting groups differ between estimates from the Gulf of Alaska and all other geographic strata located in the Bering Sea and western North Pacific Ocean.

Broad-Scale Reporting Groups	Fine-Scale Reporting Groups Included
Gulf of Alaska Estimates	
Bering Sea	Russia, Coastal Western Alaska, Middle Yukon River, Upper Yukon River, North Alaska Peninsula
Gulf of Alaska	Northwest Gulf of Alaska, Copper River, Northeast Gulf of Alaska
Southeast Alaska and North British Columbia (SE AK and N BC)	Southeast Alaska and North British Columbia
South British Columbia and West Coast US (S BC and Coast US)	South British Columbia, West Coast US
Bering Sea and western North Pacific Ocean Estimates	
Russia	Russia
Western Alaska and North Alaska Peninsula (W AK and AK Pen)	Coastal Western Alaska, North Alaska Peninsula
Middle and Upper Yukon River (Mid and Up Yuk)	Middle Yukon River, Upper Yukon River
North Pacific Ocean (N Pac Ocean)	Northwest Gulf of Alaska, Copper River, Northeast Gulf of Alaska, Southeast Alaska and North British Columbia, South British Columbia, West Coast US

Table 1.4. Stock composition estimates and sample sizes for 22 mixtures in the Bering Sea and North Pacific. Mixtures are stratified into eight geographic strata (Fig. 1.2). Abbreviations for geographic strata are from Fig. 1.2. Abbreviations for fine-scale reporting groups are from Fig. 1.1. Fine-scale reporting groups were pooled into broad-scale reporting groups (gray regions) in estimates with <100 fish. Abbreviations and information for these broad-scale reporting groups are found in Table 1.3. Credibility intervals (90%) are shown in parentheses. Superscripted estimates had a shrink factor > 1.2 for at least one reporting group indicating lack of complete convergence. Shrink factors for all non-converging estimates are reported.

Mixture	N	Russia	Coastal W AK	North AK Pen	Mid Yukon	Up Yukon	NW GOA	Copper River	NE GOA	SE AK and N BC	S BC	Coastal US
					Gulf of Alaska							
2006 Winter[a]	219	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.58	0.29	0.12
		(0-0)	(0-0)	(0-0)	(0-0)	(0-0)	(0-0)	(0-0)	(0-0.01)	(0.49-0.65)	(0.23-0.37)	(0.09-0.17)
2005, 2006 Summer	62			0.06				0.15		0.50	0.29	
				(0.01-0.13)				(0.07-0.24)		(0.38-0.62)	(0.18-0.4)	
2005, 2006 Fall [b]	127	0.00	0.00	0.00	0.00	0.00	0.06	0.01	0.01	0.57	0.25	0.10
		(0-0)	(0-0.01)	(0-0)	(0-0)	(0-0)	(0.02-0.1)	(0-0.04)	(0-0.03)	(0.46-0.68)	(0.15-0.34)	(0.06-0.16)
					Eastern Bering Sea Shelf							
2006 Winter[c]	415	0.00	0.54	0.19	0.00	0.00	0.00	0.00	0.00	0.07	0.16	0.04
		(0-0)	(0.38-0.74)	(0.01-0.34)	(0-0)	(0-0)	(0-0)	(0-0)	(0-0)	(0.04-0.1)	(0.13-0.19)	(0.02-0.06)
2007 Winter	176	0.00	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.04	0.01
		(0-0)	(0.89-0.96)	(0-0)	(0-0)	(0-0)	(0-0)	(0-0)	(0-0)	(0-0.04)	(0.02-0.07)	(0-0.03)
2008 Winter	253	0.00	0.77	0.17	0.01	0.02	0.00	0.00	0.00	0.00	0.02	0.01
		(0-0.01)	(0.71-0.82)	(0.12-0.22)	(0-0.03)	(0-0.05)	(0-0.01)	(0-0)	(0-0)	(0-0)	(0.01-0.04)	(0-0.02)
2005 June	118	0.01	0.28	0.08	0.01	0.00	0.00	0.00	0.00	0.08	0.20	0.32
		(0-0.03)	(0.21-0.36)	(0.04-0.13)	(0-0.03)	(0-0.01)	(0-0.01)	(0-0.03)	(0-0.01)	(0.02-0.15)	(0.13-0.29)	(0.25-0.4)
2006 June	105	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.33	0.38

Mixture	N	Russia	Coastal W AK	North AK Pen	Mid Yukon	Up Yukon	NW GOA	Copper River	NE GOA	SE AK and N BC	S BC	Coastal US
		(0-0)	(0.03-0.12)	(0-0)	(0-0.01)	(0-0)	(0-0.01)	(0-0.01)	(0-0.02)	(0.14-0.29)	(0.25-0.41)	(0.31-0.47)
2005 Summer	279	0.00	0.30	0.06	0.00	0.00	0.07	0.02	0.00	0.17	0.27	0.11
		(0-0.01)	(0.24-0.35)	(0.03-0.1)	(0-0)	(0-0)	(0.03-0.11)	(0.01-0.03)	(0-0.01)	(0.13-0.22)	(0.23-0.32)	(0.08-0.14)
2006 Summer	223	0.00	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.22	0.13
		(0-0)	(0.47-0.58)	(0-0)	(0-0)	(0-0)	(0-0)	(0-0)	(0-0)	(0.08-0.16)	(0.17-0.27)	(0.09-0.17)
2005, 2006 Fall	234	0.00	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.17	0.12
		(0-0)	(0.6-0.71)	(0-0.01)	(0-0)	(0-0)	(0-0.01)	(0-0.01)	(0-0)	(0.01-0.07)	(0.13-0.22)	(0.09-0.16)
Central Bering Sea Shelf												
2007, 2008 Winter[d]	74	0.00	0.71		0.27		0.02					
		(0-0.01)	(0.6-0.81)		(0.17-0.38)		(0-0.05)					
2005 Summer	290	0.01	0.69	0.07	0.03	0.01	0.03	0.00	0.00	0.01	0.08	0.06
		(0-0.03)	(0.64-0.74)	(0.04-0.1)	(0.01-0.05)	(0-0.03)	(0.01-0.06)	(0-0)	(0-0)	(0-0.03)	(0.06-0.11)	(0.04-0.08)
2006 Summer[e]	180	0.00	0.86	0.00	0.04	0.00	0.00	0.00	0.00	0.03	0.05	0.02
		(0-0)	(0.8-0.92)	(0-0)	(0-0.09)	(0-0.03)	(0-0)	(0-0)	(0-0)	(0.01-0.05)	(0.02-0.08)	(0-0.03)
2005, 2006, 2008 Fall[f]	141	0.00	0.83	0.00	0.12	0.00	0.00	0.00	0.00	0.02	0.01	0.01
		(0-0.01)	(0.76-0.91)	(0-0)	(0.06-0.19)	(0-0.01)	(0-0)	(0-0)	(0-0)	(0-0.04)	(0-0.03)	(0-0.03)
Central Bering Sea Basin												
2007, 2008 Summer	89	0.08	0.55		0.34		0.03					
		(0.04-0.13)	(0.45-0.65)		(0.25-0.43)		(0-0.09)					
2009 Summer	137	0.20	0.60	0.00	0.05	0.13	0.00	0.00	0.00	0.00	0.00	0.01
		(0.15-0.26)	(0.52-0.67)	(0-0.01)	(0-0.12)	(0.07-0.2)	(0-0.01)	(0-0.01)	(0-0)	(0-0.01)	(0-0)	(0-0.02)
2010, 2011 Summer	76	0.26	0.45		0.28		0.01					
		(0.18-0.35)	(0.35-0.55)		(0.19-0.38)		(0-0.03)					
Western Bering Sea												
2008 Summer	21	0.69	0.22		0.05		0.04					

Mixture	N	Russia	Coastal W AK	North AK Pen	Mid Yukon	Up Yukon	NW GOA	Copper River	NE GOA	SE AK and N BC	S BC	Coastal US
		(0.52-0.84)	(0.08-0.39)		(0-0.15)		(0-0.15)					
					Kamchatka Peninsula							
2008 Summer	126	0.97	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(0.94-0.99)	(0-0.01)	(0-0)	(0-0.05)	(0-0.02)	(0-0)	(0-0)	(0-0)	(0-0.01)	(0-0)	(0-0)
					Commander Islands							
2008 Spring	140	0.87	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00
		(0.82-0.92)	(0-0.01)	(0-0)	(0-0.01)	(0.08-0.17)	(0-0)	(0-0)	(0-0)	(0-0.01)	(0-0)	(0-0)
					Aleutian Islands							
2006, 2009-2011	78	0.33	0.55		0.11		0.01					
		(0.24-0.42)	(0.45-0.65)		(0.05-0.18)		(0-0.03)					

[a] SE AK and N BC shrink factor = 1.49, S BC 1.51

[b] SE AK and N BC shrink factor = 1.55, S BC 1.47

[c] Coastal W AK 6.74, North Alaska Pen 7.86

[d] Coastal W AK combined with North Alaska Pen 1.24, Mid and Up Yuk 1.25

[e] Mid Yukon 1.22

[f] Coastal W AK 1.20, Mid Yukon 1.34

Figures

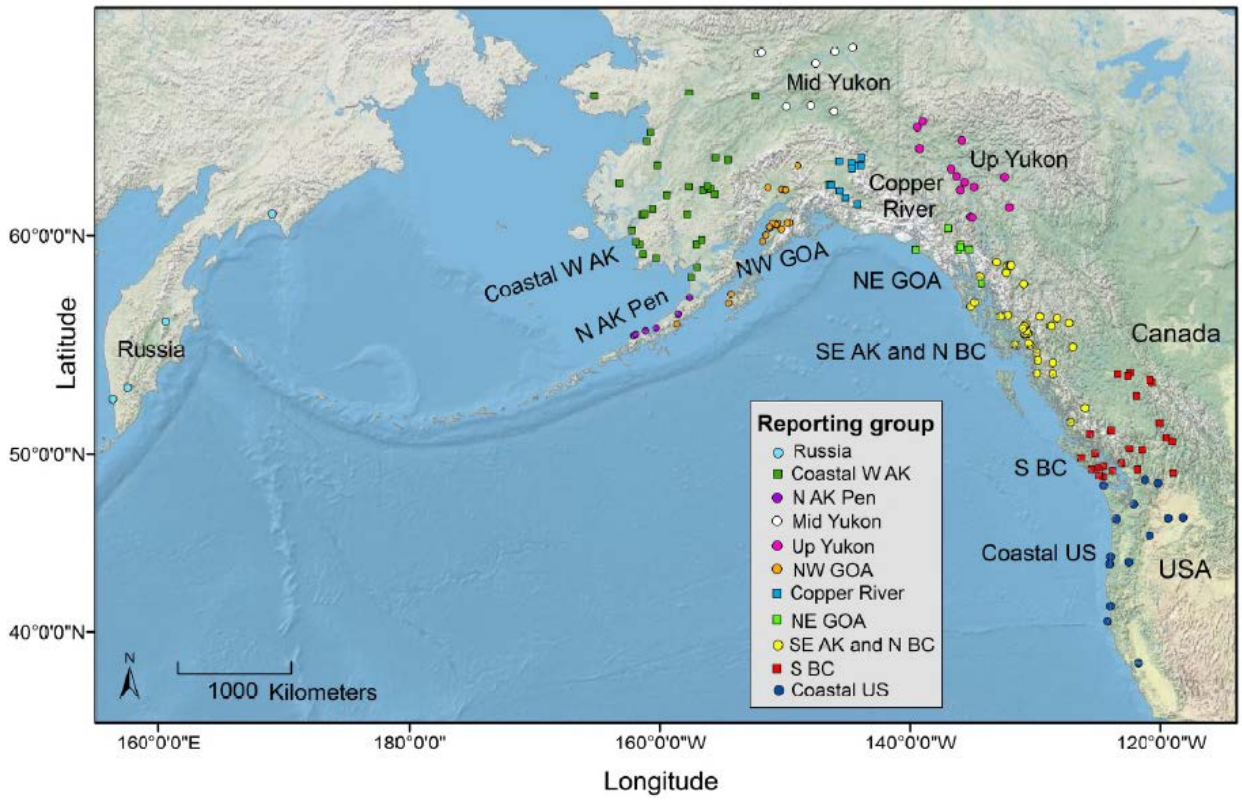

Fig. 1.1. Map of baseline collections used for genetic stock identification. Collections are colored by fine-scale reporting group. All fine-scale reporting groups except SE AK and N BC and South BC are identical to those of Templin *et al.* (2011). Full fine-scale reporting group names are found in Table 1.1.

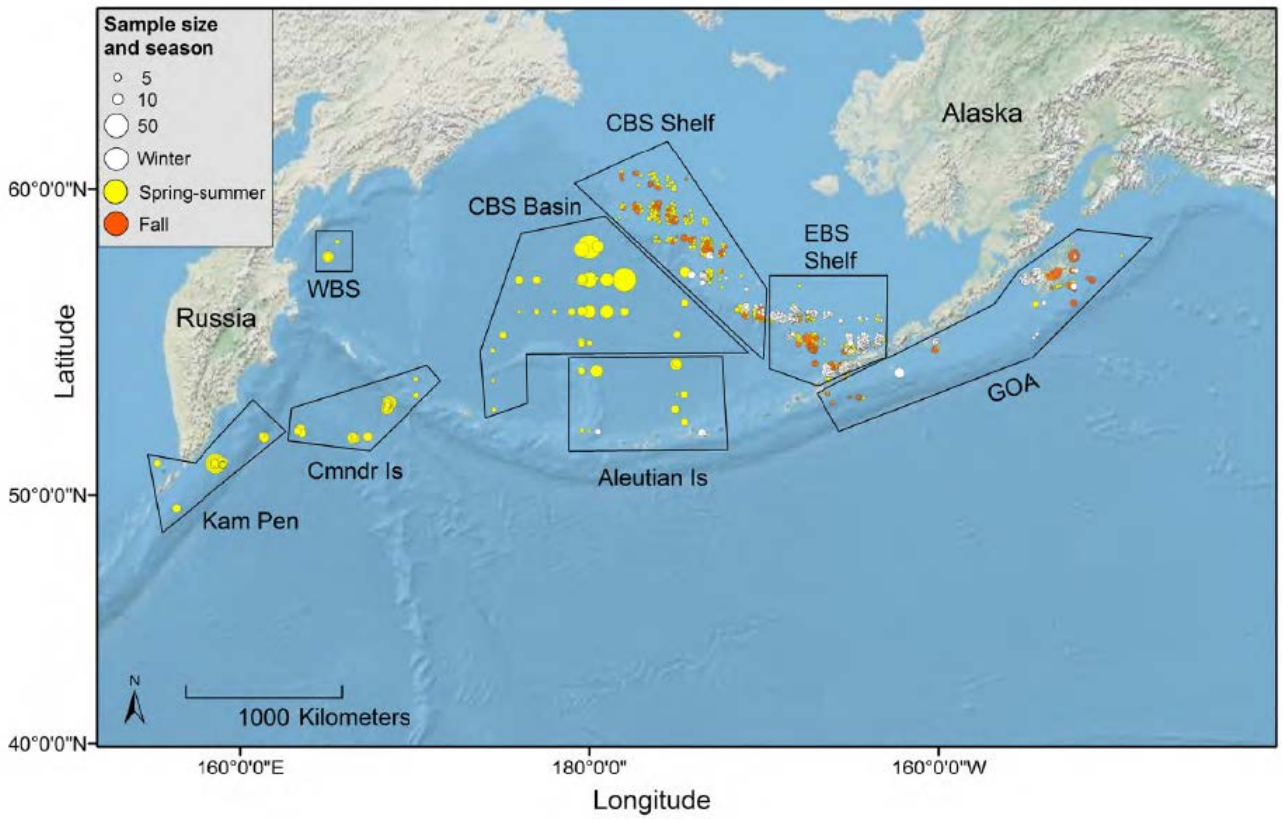

Fig. 1.2. Map of geographic strata and sampling locations with sample size and season of collection indicated by color and size of dot. Geographic strata are as follows: Gulf of Alaska (GOA), Eastern Bering Sea Shelf (EBS Shelf), Central Bering Sea Shelf (CBS Shelf), Central Bering Sea Basin (CBS Basin), Western Bering Sea (WBS), Kamchatka Peninsula (Kam Pen), Commander Islands (Cmndr Is) and Aleutian Islands (Aleutian Is).

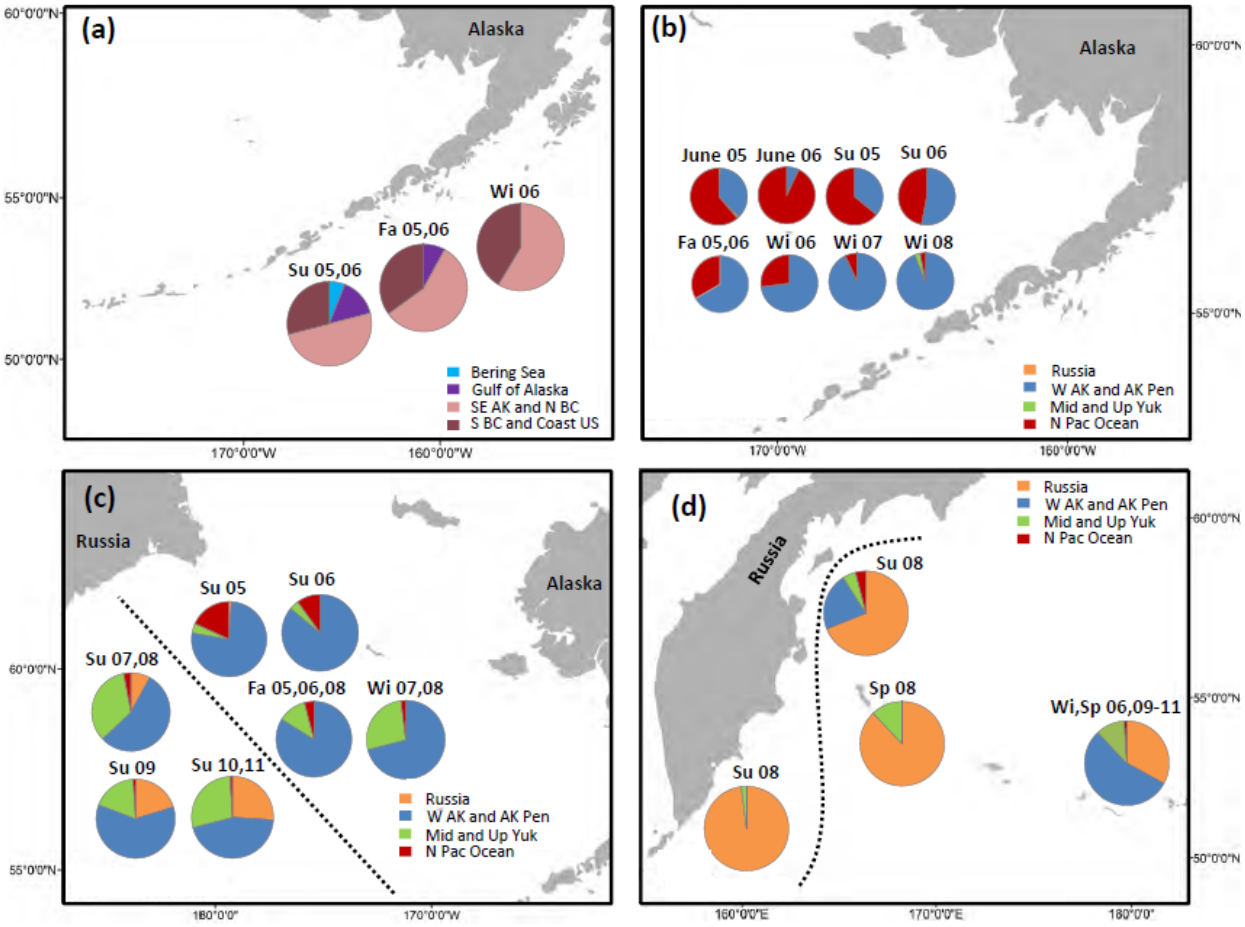

Fig. 1.3. Stock composition estimates of Chinook salmon caught in the Bering Sea and North Pacific Ocean. (a) Gulf of Alaska, (b) Eastern Bering Sea Shelf, (c) Central Bering Sea Shelf (top right), Central Bering Sea Basin (bottom left), (d) Western Bering Sea (top middle), Kamchatka Peninsula (bottom left), Commander Islands (bottom middle) and Aleutian Islands (bottom right). Each pie represents a separate mixture estimate. Pies are identified by season and year sampled. Seasonal abbreviations are spring (Sp), summer (Su), fall (Fa) and winter (Wi). Black dashed lines in panels (c) and (d) represent the edge of the continental shelf. Reporting groups are the broad-scale groups described in Table 1.3. Abbreviations for broad-scale reporting groups are found in Table 1.3.

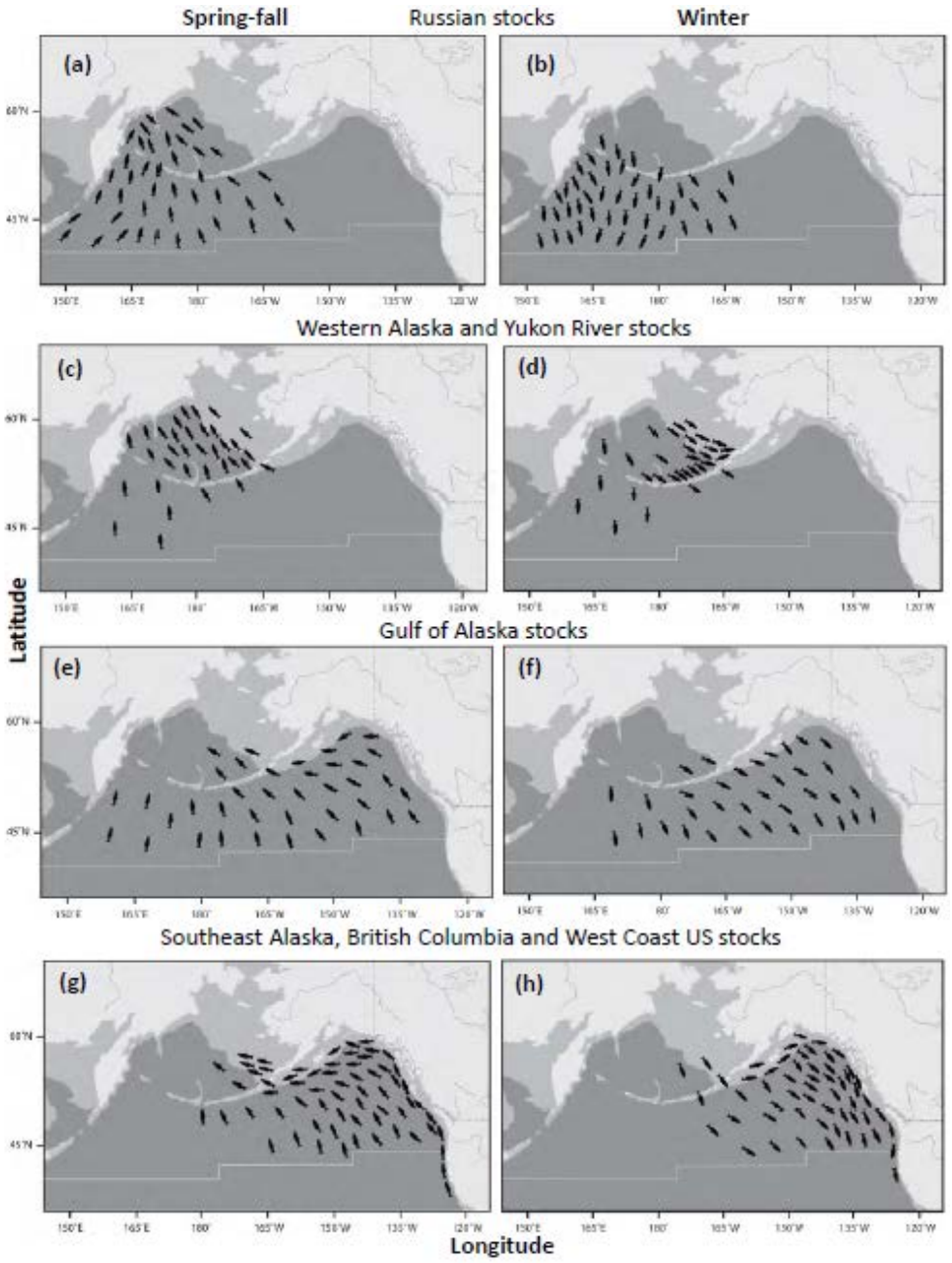

Fig. 1.4. Graphical synopses of seasonal distribution patterns of immature Chinook salmon in the Bering Sea and North Pacific Ocean. Orientation of fish signifies direction of seasonal movements. Distance between fish is proportional to density. Grey solid line represents farthest known southern occurrence of Chinook salmon (Major *et al.* 1978). Synopses are shown for four large stock aggregations: 1) Russian stocks: (a) spring-fall and (b) winter; 2) Western Alaska and Yukon River stocks: (c) spring-fall and (d) winter; 3) Gulf of Alaska stocks: (e) spring-fall and (f) winter; and 4) Southeast Alaska, British Columbia and West Coast US stocks (g) spring-fall and (h) winter.

Chapter 2

Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*)[2]

Abstract

Recent advances in population genomics have made it possible to detect previously unidentified structure, obtain more accurate estimates of demographic parameters and explore adaptive divergence, potentially revolutionizing the way genetic data is used to manage wild populations. Here, we identified 10,944 single-nucleotide polymorphisms using restriction-site-associated DNA (RAD) sequencing to explore population structure, demography and adaptive divergence in five populations of Chinook salmon (*Oncorhynchus tshawytscha*) from western Alaska. Patterns of population structure were similar to those of past studies but our ability to assign individuals back to their region of origin was greatly improved (> 90% accuracy for all populations). We also calculated effective size with and without removing physically linked loci identified from a linkage map, a novel method for non-model organisms. Estimates of effective size were generally above 1,000, and were biased downward when physically linked loci were not removed. Outlier tests based on genetic differentiation identified 733 loci and three genomic regions under putative selection. These markers and genomic regions are excellent candidates for future research and can be used to create high-resolution panels for genetic monitoring and population assignment. This work demonstrates the utility of genomic data to inform conservation in highly-exploited species with shallow population structure.

Introduction

Discrete management of genetically distinct populations can increase species-wide resilience and stabilize the productivity of ecosystems as a whole (Hilborn *et al.* 2003; Schindler *et al.* 2010). For over three decades, genetic data from 10-100 putatively-neutral markers was used to identify discrete populations, define conservation units and estimate demographic

[2] Full citation: Larson, W.A., L.W. Seeb, M.V. Everett, R.K. Waples, W.D. Templin, and J.E. Seeb. 2014. Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications*, **7**(3):355-369. Supplementary materials available from the online version of this manuscript (http://onlinelibrary.wiley.com/doi/10.1111/eva.12128/abstract).

parameters (Utter *et al.* 1974; Wirgin & Waldman 1994; Waples *et al.* 2008). The use of genetic data for management has been especially successful in Pacific salmon (*Oncorhynchus spp.*), which exhibit extensive population structure (Utter & Ryman 1993; Shaklee *et al.* 1999). However, applications have been limited for recently isolated populations of salmonids (Taylor *et al.* 1997) or marine species with little neutral structure (Waples 1998). In these circumstances, data from thousands of markers (genomic data) may be necessary to resolve population structure and aid management.

Genomic data can provide accurate estimates of neutral population structure (Avise 2010; Funk *et al.* 2012; Narum *et al.* 2013), identify genomic regions that display adaptive divergence (Allendorf *et al.* 2010; Angeloni *et al.* 2012) and provide increased accuracy when estimating demographic parameters (Allendorf *et al.* 2010). Genotypes from thousands of loci have been used to elucidate neutral structure in populations of Pacific lamprey (*Entosphenus tridentatus*, Hess et al. 2012) and to improve resolution of fine-scale structure in Atlantic salmon (*Salmo salar*, Bourret et al. 2013) . Additionally, genome scans have revealed adaptively important markers and genomic regions in sockeye salmon (*Oncorhynchus nerka*, Russello et al. 2012), Atlantic cod (*Gadus morhua*, Bradbury et al. 2013; Hemmer-Hansen et al. 2013), and lake whitefish (*Coregonus clupeaformis*, Renaut et al. 2012) . Although many studies have used genomic data to elucidate structure in non-model organisms, demographic parameters such as effective size are rarely estimated with these types of data.

Effective population size (N_e) is an important parameter in conservation biology (Frankham 2005), but methods to calculate N_e with genomic data are lacking (Waples & Do 2010). Specifically, many calculations of N_e require knowledge of linkage relationships, which are often unknown for non-model organisms. A possible solution to this problem is the use of high density linkage maps that can now be created rapidly for many species with genotyping by sequencing (e.g., Baxter *et al.* 2011; Miller *et al.* 2012; Gagnaire *et al.* 2013a). Using data from these maps, it is possible to obtain estimates of N_e that are not biased by physical linkage. To the best of our knowledge this method has only been implemented in populations of model organisms (Park 2011; Sved *et al.* 2013), but the increasing availability of linkage maps will facilitate N_e estimation in many species of conservation concern.

Chinook salmon (*Oncorhynchus tshawytscha)* from western Alaska represent an excellent system to explore the utility of genomics in a management context. Chinook salmon

inhabit four major regions in western Alaska: Norton Sound, the Yukon River, the Kuskokwim River and Bristol Bay, all of which vary significantly in size, hydrology and climate (Fig. 2.1, Olsen *et al.* 2011). The Kuskokwim and Yukon regions are composed of a mainstem river with many tributaries, whereas Norton Sound is composed of many unconnected and short rivers (mean length ~110 km; Olsen *et al.* 2011). The Bristol Bay region is composed of several river systems each with smaller tributaries (i.e. Nushagak, Togiak, Naknek rivers). Past studies using allozymes and single-nucleotide polymorphisms (SNPs) found evidence of structure in western Alaska but concluded that differences among populations in Norton Sound, the lower portions of the Yukon and Kuskokwim rivers, and Bristol Bay were insufficient to allocate mixture samples back to their region of origin (Gharrett *et al.* 1987; Templin *et al.* 2011). Returns of Chinook salmon to western Alaska over the past decade have been approximately 20% lower than their long-term average, renewing interest in the migration patterns and vulnerability of stocks to fisheries in this region (ADF&G 2013). Improved resolution of population structure would allow managers to investigate these questions using genetic tools. Additionally, estimates of N_e and other demographic parameters could help to inform conservation and management efforts across the region.

We used restriction-site-associated DNA (RAD) sequencing to investigate the population structure and demography of Chinook salmon from western Alaska. We identified over 10,000 SNPs in 270 individuals from five populations across western Alaska. Patterns of genetic variation were assessed using both population and individual-based methods and validated with assignment tests. We then aligned our RAD markers to a linkage map to calculate N_e with and without removing physical linkage. We also conducted outlier tests and used the linkage map to detect loci and genomic regions under putative selection. This approach defines an important way that genomics can be used to inform management of non-model species with high gene flow.

Materials and methods

Tissue sampling

Tissue samples from spawning Chinook salmon were available from four regions in coastal western Alaska and one in the upper Yukon River (Fig. 2.1, Table 2.1). We selected populations that did not have unusually small census sizes and that were genetically similar to proximate populations identified from previous studies (Olsen *et al.* 2011; Templin *et al.* 2011);

this approach ensured that our conclusions were based on populations that were representative of each region. Chinook salmon from the upper Yukon River are highly differentiated from those of western Alaska (Smith *et al.* 2005c; Templin *et al.* 2011) and were included to anchor inferences of population structure.

RAD sequencing, SNP discovery and genotyping

RAD libraries were prepared with the restriction enzyme *SbfI* following the methods of Baird et al. (2008) and Everett et al. (2012) and sequenced on an Illumina HiSeq2000 at the University of Oregon Genomics Core Facility. We constructed 18 libraries for single-end sequencing (100 bp target length) containing 12-24 individuals per library and one library for paired-end sequencing (100 x 2 bp target) containing 8 individuals to assemble longer sequence contigs for annotation. Pooled individuals were identified with unique 6 bp barcodes.

We used the *Stacks* software package, version 0.9999 (Catchen *et al.* 2011) and methods similar to Hohenlohe et al. (2013) to discover and genotype SNPs from the sequenced RAD tags. Quality filtering of raw reads and de-multiplexing based on barcode was conducted using *process_radtags*. Stacks of similar sequences were then assembled in each individual with *ustacks*, and a catalog of loci was created with *cstacks*. We included only the two individuals from each population with the greatest amount of sequence data when creating our catalog to reduce the detection of false polymorphisms. Including more individuals per population would have facilitated the detection of low frequency SNPs, but would not have added additional SNPs to the final dataset because these low frequency SNPs were filtered out in downstream analyses. Finally, we used *sstacks* and *populations* to combine the genotypes from each individual into a single *Genepop* formatted file.

SNP validation

Putative SNPs discovered using *Stacks* were filtered to remove possible sequencing errors, paralogous sequence variants (PSVs), and uninformative polymorphisms. First, we removed any putative SNP that failed to genotype in > 80% of individuals. We then removed those with a minor allele frequency < 0.05 in all populations. These polymorphisms are likely to be uninformative, are difficult to distinguish from sequencing errors, can distort signals of selection and drift in natural populations, and may bias tests for selection (Roesti *et al.* 2012). We also discarded putative SNPs that were found at RAD tag positions > 87 bp because these positions contained more polymorphisms on average than the rest of the sequence (138 per bp

for bp 1-87, 223 per bp for bp 88-93). This increase in putative SNPs per base pair is likely a result of sequencing errors as Illumina sequencing is more error prone towards the terminal positions of reads (Minoche *et al.* 2011). We kept only the putative SNP with the highest F_{ST} from each RAD tag to reduce linkage in our dataset. We also used the program *PLINK* version 1.07 (Purcell *et al.* 2007) to test for linkage disequilibrium between each pair of loci. If a locus pair had an r^2 value > 0.8 in three out of five populations we removed the locus that was genotyped in the fewest individuals.

PSVs, which are abundant in salmonids as a result of an ancient whole-genome duplication event (Allendorf & Thorgaard 1984; Seeb *et al.* 2011b), were removed from the dataset when possible. PSVs are closely related sequences from different genomic locations that do not segregate as single loci and are therefore difficult to genotype accurately (Gidskehaug *et al.* 2011). Haploid individuals can be used to identify PSVs because PSVs will appear heterozygous when all correctly segregating loci are homozygous (Hecht *et al.* 2013). To screen for PSVs in our data, we genotyped 50 haploid Chinook salmon from Washington, USA, at all putative SNPs discovered above, and loci with >10% heterozygosity were removed.
We also conducted exact tests of Hardy-Weinberg equilibrium in *Genepop* version 4 (Rousset 2008) and removed loci that were out of equilibrium in three or more populations ($P < 0.05$). We then removed individuals that were missing genotypes at > 15% of the SNPs. As a final filtration step we used *ML-Relate* (Kalinowski *et al.* 2006) to look for duplicated individuals in our data.

Paired-end assembly and *BLAST* annotation

We conducted a paired-end assembly with the P1 and P2 reads from each locus using *Velvet* (v.1.1.06, Zerbino & Birney 2008) and the methods of Etter et al. (2011) and Everett et al. (2012) to increase query lengths for *BLAST* annotation. Consensus sequences were then aligned to the Swiss-Prot database using the *BLASTX* search algorithm. Alignments with E-values of $\leq 10^{-4}$ were retained. If multiple alignments had E-values of $\leq 10^{-4}$ for the same locus, the alignment with the lowest E-value was retained.

Population structure and assignment tests

Initial analysis of population structure was conducted with an individual-based principal component analysis (PCA) implemented in the R package *adegenet* (Jombart 2008). The significance of each principal component was assessed by randomly permuting the data 1,000

times and comparing the observed eigenvalues to values generated by conducting PCA on the permuted data. PCA revealed five nonconforming individuals in the Anvik River collection that grouped between the Big Salmon River and Anvik River clusters (Fig. 2.2). These five individuals were removed from further analyses as they likely represent transient fish from middle or upper Yukon River populations. After removing these individuals, we calculated pairwise-F_{ST} values (Weir & Cockerham 1984) for each population and performed significance tests for genetic differentiation in *Arlequin* 3.5 (Excoffier & Lischer 2010) using an exact test with 10,000 permutations.

We conducted an analysis of molecular variance (AMOVA) in *Arlequin* 3.5 to examine the variation within and between groups of genetically similar populations. The hierarchy for this analysis was chosen based on the clustering from the PCA: (1) Koktuli River, Kogrukluk River and Anvik River (2) Tubutulik River and (3) Big Salmon River. Separate AMOVAs were conducted for (1) the entire dataset, and (2) all populations except the Big Salmon River. Finally, we calculated global and per-locus observed and expected heterozygosities for each population in *GenAlEx* 6.5 (Peakall & Smouse 2012).

We examined fine-scale structure in the closely related Anvik River, Kogrukluk River and Koktuli River populations with an individual PCA including only these three populations (see above for PCA methods). This analysis was conducted separately for the 10,944 RAD SNPs and 39 of the 43 SNPs from Templin et al. (2011) that were developed for Chinook salmon from expressed sequence tags. Of the four SNPs from Templin et al. (2011) that were not genotyped, two were removed because they were essentially monomorphic in other populations from western Alaska and two were removed because they were in linkage disequilibrium with another locus (Templin *et al.* 2011).

Assignment power of four panels was evaluated with leave-one-out tests in *GeneClass2.0* (Piry *et al.* 2004) to compare the influence of number of SNPs and relative divergence of SNPs on assignment accuracies. The four panels were (1) 39 SNPs from Templin et al. (2011), (2) 39 randomly chosen SNPs from the complete dataset of 10,944 RAD SNPs, (3) the complete dataset of 10,944 RAD SNPs, and (4) the full set of RAD SNPs with the 733 outlier SNPs that were found to be under putative selection removed. We did not construct a panel with only the most divergent RAD SNPs because this approach would have led to an upward bias in the predicated accuracy of assignment for that panel (Anderson 2010). Leave-one-out tests were conducted by

removing an individual from the baseline without replacement then assigning that individual back to a reference population using a Bayesian approach described in Rannala and Mountain (1997). Individuals were considered to be assigned to a population if the assignment probability to that population was higher than to any other population.

Alignment to linkage map

We aligned our filtered loci to a linkage map for Chinook salmon consisting of 3,534 RAD-derived SNPs distributed across 34 linkage groups ranging in size from 27.75 to 160.23 cM (map to be presented elsewhere). To conduct the alignments, we used *BLASTN* (Altschul *et al.* 1990) with the following parameters: minimum alignment length of 90 bp, 95% identity and no more than two mismatching bases. If a single locus aligned to multiple map loci, we discarded all alignments for that locus. We used relatively strict alignment parameters for this analysis because sequence alignment in tetraploid-origin salmonids can provide ambiguous results when alignment parameters are not sufficiently strict (Everett *et al.* 2011; Seeb *et al.* 2011b).

Calculating N_e and N_e/N

Estimates of N_e were performed with the linkage disequilibrium method (Hill 1981; Waples 2006) updated for missing data following Peel et al. (2013). This method assumes all loci in the analysis are physically unlinked then utilizes the observed linkage disequilibrium to estimate N_e. We removed comparisons between loci on the same linkage groups to obtain estimates that were unbiased by physical linkage (Park 2011; Sved *et al.* 2013). Additionally, we removed all loci that were putatively under selection as suggested by Waples (2006) (see below for description of tests for loci under selection). Calculations of N_e were conducted using *N_eEstimator* (Do *et al.* 2014) and *R* (R core development team). *N_eEstimator* was used to calculate r^2 values for each locus pair with the following parameters: a minimum allele frequency cutoff of 0.02 and a random mating model. We then implemented the methods described in Waples (2006) and Peel et al. (2013) in *R* to obtain N_e estimates and parametric 95% confidence intervals for each population (scripts available from W. Larson upon request). We calculated N_e for three datasets (1) all RAD SNPs that aligned to the linkage map, (2) all RAD SNPs that aligned to the map with pairwise comparisons between markers on the same linkage group removed and (3) the 39 SNPs from Templin et al. (2011) that were in linkage equilibrium.

We calculated the ratio of effective size to census size (N_e/N) using N_e calculated with RAD-derived SNPs after removing physical linkage and estimates of total escapement obtained from aerial surveys (Koktuli River, Anvik River, Tubutulik River) and weir counts (Kogrukluk River, Big Salmon River). Multiple aerial surveys were used to estimate total run size for the Anvik River and Koktuli River populations but only single aerial counts were available for the Tubutulik River population. Single aerial counts from a river near the Tubutulik River collection were approximately four times smaller than those taken from a counting tower; therefore, we multiplied the aerial count from our collection by four. We averaged the last ten years of data to obtain an approximate value of census size for each population (only last three years used for Koktuli River due to data availability).

Estimates of N_e for Chinook salmon populations are complicated by the fact that multiple cohorts are represented in each spawning group (Waples 1990a). Single-sample estimates of N_e therefore do not precisely reflect the effective number of breeders per year or the effective number of breeders per generation but instead represent some intermediate value. We calculated two N_e/N ratios to bracket these possible scenarios: N_e divided by the average census size (escapement) per year (N_e/N) and N_e divided by the total census size per generation (N_e/NG). Values of G for each population were obtained from the sources in Table 2.5 by averaging age compositions across one to 38 years of data depending on availability.

Detection of loci under putative selection

We identified putative loci under selection with *Arlequin* 3.5. This program uses coalescent simulations to create a null distribution of F-statistics then generates P-values for each locus based on this distribution and observed heterozygosities across loci (Excoffier *et al.* 2009). A hierarchical island model was selected to reduce false-positives introduced due to underlying population structure (Excoffier *et al.* 2009). The population hierarchy was the same as in the AMOVA. Settings for the analysis were 20,000 simulations, 10 simulated groups and 100 demes per group. Loci that fell above the 95% quantile of the F_{ST} distribution were considered candidates for directional selection.

Detection of candidate genomic regions under selection

We used a linkage map in conjunction with a sliding window analysis to identify highly divergent regions of the genome that may be under selection (c.f., Bourret *et al.* 2013). This analysis was conducted with a sliding window approach that compares the mean pairwise F_{ST} of

a small (5 cM) genomic region to a null distribution created by bootstrapping over the complete dataset (Hohenlohe *et al.* 2010; Bourret *et al.* 2013). For each window, we sampled *N* F_{ST} values with replacement from the entire dataset where *N* was the number of SNPs in the window. This resampling routine was repeated 1,000 times to generate a null distribution. Windows with mean F_{ST} values above the 95% quantile of the null distribution were candidates for directional selection. If a window mean was above 90% after 1,000 replicates, we increased the number of replicates to 5,000 to improve accuracy in the tails of the null distribution. We chose a sliding window size of 5 cM and frame shift value of 1 cM. We also required at least two SNPs to be present in a window to conduct the above test. After testing multiple window sizes, we found that a 5 cM window provided sufficient resolution for detecting divergent regions without introducing excessive variance. This value was also used by Bourret et al. (2013) for linkage groups with similar numbers of markers to ours. We conducted this analysis for all pairwise population comparisons.

Results

Sequencing, SNP discovery and filtration

We obtained RAD data from 289 individuals, and the number of sequences obtained for each individual ranged from 1,622,400 to 8,707,337 with an average of 3,796,368 (excluding low quality individuals, see below). Alignments using *Stacks* revealed 42,351 putative SNPs. Removing putative SNPs that were genotyped in < 80% of individuals eliminated more than half of these, leaving 20,296. After removing polymorphisms in bp 87-94 of each RAD-tag, removing all but one putative SNP from each tag and removing SNPs with minor allele frequency < 0.05, 12,585 SNPs remained. Screens for paralogous sequence variants revealed 845 loci that were potentially duplicated; these loci were eliminated. Significant deviations from Hardy-Weinberg equilibrium were observed in 397 SNPs, and these loci were also removed. Significant linkage disequilibrium in three or more populations was found for 399 SNPs, and one SNP from each pair was removed. The final filtered dataset consisted of 10,944 SNPs. We removed 17 individuals that were genotyped in <85% of SNPs, seven from the Kogrukluk River, four from the Anvik River and six from the Big Salmon River (adjusted sample sizes in Table 2.1). Relatedness analysis revealed two pairs of duplicated individuals (R > 0.9) from the Anvik River population, and one individual from each pair was removed. The final filtered dataset

consisted of 270 individuals genotyped at 10,944 SNPs. Summary statistics for each locus are available in Table S1, and histograms of overall and pairwise F_{ST} for each locus are in Fig. S1.

Paired-end assembly and *BLAST* Annotation

Paired-end assemblies produced 11,666 contigs with an average length of 268 bp (minimum 150 bp maximum 565 bp). *BLAST* annotation of these contigs yielded significant hits for 1,576 (14%) of 10,944 SNPs (Table S2). Of these hits, over one third aligned to transposable elements. Other common functional groups included DNA polymerases and transmembrane proteins.

Population structure

PCA analysis revealed that the Big Salmon River and Tubutulik River populations formed completely separate clusters while the Koktuli River, Kogrukluk River and Anvik River populations essentially formed a single cluster (Fig. 2.2). The overall F_{ST} of the full dataset was 0.041, and pairwise F_{ST} values ranged from 0.003 for the Koktuli River-Kogrukluk River comparison to 0.098 for the Big Salmon River-Tubutulik River comparison (Table 2.2). Genetic differentiation between all population comparisons was highly significant ($P < 0.001$). The results of these significance tests should, however, be interpreted with extreme caution due to the large number of loci, which may overestimate precision.

We conducted hierarchical AMOVA for the entire dataset and for a dataset without the Big Salmon River population (Table 2.3). Both analyses displayed much larger variation among groups than within groups. Levels of observed heterozygosity across populations ranged from 0.232 for the Big Salmon River to 0.260 for the Anvik River (Table 2.1).

When the Koktuli River, Kogrukluk River and Anvik River populations were analyzed separately with 10,944 SNPs, all populations generally formed discrete clusters, but some overlap was present between the Koktuli River and Kogrukluk River populations (Fig. 2.3a). Additionally, populations from the Anvik River and Kogrukluk River each contained a subset of 10-20 individuals that fell outside the main cluster. When PCA was conducted with the 41 SNPs from Templin et al. (2011), no clustering pattern was apparent (Fig. 2.3b).

The relatively small amount of variation (1-5%) explained by the first and second principal components (PCs) in our PCAs (Fig. 2.2 and 2.3) can be attributed to the large number of axes used. Each PCA contained as many axes as individuals plotted, so PCA using all populations contained 270 axes and the PCA with three populations contained 163. PCs one and

two in both PCAs each explained more than three times the variation of the average axis and explained significantly more variation than would be expected if no real correlation existed ($P < 0.001$), but because of the large number of axes, the actual proportion of variation explained was small.

Assignment accuracy was much higher using > 10,000 SNPs ($\geq$ 89% assignment to correct population) compared to 39 SNPs (~50% assignment to correct population, Table 2.4). Panels containing close to the same number of SNPs generally preformed similarly, but the 39 SNPs from Templin et al. (2011) did preform slightly better than the 39 randomly chosen RAD SNPs, and the panel containing all 10,944 RAD SNPs preformed slightly better than the panel with the 733 outlier SNPs removed (Table 2.4).

Alignment to linkage map

Of the 10,944 filtered loci, 1,156 were successfully placed on the linkage map (33% of loci on the map successfully aligned to a population locus, see Table S1 for map location of successful alignments). This proportion may seem small, but it is important to note that the map was constructed using a single Chinook salmon from Washington State. Chinook salmon from Washington State are substantially diverged from populations in western Alaska (Templin *et al.* 2011), therefore, it is likely that many RAD tags did not contain loci that were polymorphic in both the mapping cross and our study populations and were not useful for our analyses. Additionally, because only one individual was used for the mapping cross, our alignments were limited to the RAD tags containing SNPs that segregated in the mapped individual.

Demographic estimates

Estimates of N_e with the RAD-derived SNPs were highly variable across populations, ranging from close to 500 in the Anvik River to infinity for the Koktuli River (Table 2.5). These estimates were calculated using SNPs that were successfully aligned to the linkage map, providing over 500,000 pairwise comparisons between loci. Pairwise comparisons between SNPs located on the same linkage group represented about 20,000 of the 500,000 comparisons (6%). These 20,000 comparisons were removed to estimate N_e between physically unlinked loci. Estimates of N_e were consistently smaller for the dataset that included all comparisons (Table 2.5). This downward bias was not uniform, however, as estimates from Norton Sound appeared to be more affected by linkage than estimates for the other populations.

Estimates of N_e with the 39 SNPs from Templin et al. (2011) ranged from 209 for the Anvik River to infinity for the Koktuli River, Kogrukluk River and Tubutulik River populations. Confidence intervals for each estimate using 39 SNPs included infinity and were larger than confidence intervals around estimates from the RAD-derived SNPs.

Estimates of N_e/N and N_e/NG were extremely variable, ranging from 0.17 and 0.03 for the Kogrukluk River population to 0.59 and 0.11 for the Tubutulik River population (Table 2.5). We did not calculate N_e/N or N_e/NG for the Koktuli River or Big Salmon River populations because the CIs around N_e included infinity, suggesting our point estimates of N_e may not be completely representative.

Loci and genomic regions under putative selection

Outlier tests in *Arlequin* revealed 733 loci (6.7%) that were significant outliers at the 5% level and 178 (1.6%) that were significant at the 1% level. *BLAST* annotation of the outliers at the 5% level revealed 96 significant hits (13% success rate). Transposable elements represented over one third of the significant hits which is consistent with the pattern from the complete dataset.

The number of genomic regions under putative selection for each population pair ranged from 20 to 25 and generally increased when the Big Salmon River population was included (Fig. 2.4, Table 2.2). Overall, these regions appeared to be scattered randomly throughout the genome and were often significant in only one or two population comparisons. Despite this pattern, three genomic regions on separate linkage groups (LG) were candidates for selection in more than half of the population comparisons (Fig. 2.4). These regions are LG2 at 70-78 cM, LG4 at 2-8 cM and LG21 at 7-12 cM.

Discussion

We used RAD sequencing to characterize the genetic structure, genomic divergence and demography of five populations of Chinook salmon from western Alaska. Patterns of genetic differentiation were similar to, but more identifiable than in past studies (Gharrett *et al.* 1987; Templin *et al.* 2011). Estimates of population N_e ranged from 516 to infinity and appeared to be biased downward when loci that were physically linked were not removed. Regions of putative adaptive divergence appeared to be randomly distributed across the genome with few shared areas of high divergence across populations, but we did find three genomic regions that displayed high divergence in multiple populations. Using genomic data, we were able to conduct

individual assignment in populations where it was previously unfeasible, discover genomic regions under putative selection, and estimate N_e in populations with > 1,000 individuals. Our approach therefore represents a significant improvement over previous studies employing fewer markers and no linkage map.

Population structure

The largest genetic differentiation between populations in our dataset existed between the Big Salmon River from the upper Yukon and all other coastal populations. Chinook salmon from the upper Yukon are thought to have genetically diverged from coastal populations after being isolated during the last glacial maximum (Olsen *et al.* 2011). Our results support this hypothesis and are consistent with those based on allozymes, microsatellites and SNPs (Gharrett *et al.* 1987; Olsen *et al.* 2010; Templin *et al.* 2011).

We also found high levels of divergence between the Tubutulik River in Norton Sound and all other populations. This divergence was likely facilitated by the Nulato Hills, a small mountain range that separates the tributaries of Norton Sound from those of the Yukon River (Fig. 2.1), but could have also been influenced by environmental characteristics such as precipitation (Olsen *et al.* 2011).

Populations from the lower Yukon, Kuskokwim and Bristol Bay regions (Anvik River, Kogrukluk River and Koktuli River) were least divergent, displaying pairwise F_{ST} values < 0.01 for all population comparisons. The relatively small divergence we observed is consistent with other salmonids in the region (Olsen *et al.* 2011; Garvin *et al.* 2013) and is somewhat expected given the surrounding environment. Western Alaska is characterized by moisture laden tundra and dynamic rivers that frequently change paths. When such stream captures occur, gene flow is facilitated between populations that were previously isolated. It is therefore likely that substantial historic gene flow and possibly continuing low-level gene flow has largely restricted genetic differentiation in this region (Seeb & Crane 1999b).

Nevertheless, we found genetic structure among the Anvik River, Kogrukluk River and Koktuli River populations using both individual-based clustering methods and assignment tests. The Anvik River population displayed the highest levels of divergence, forming a completely isolated cluster. This population may have diverged more quickly as a reflection of its relatively small estimated census and effective sizes (N=1700, N_e=516). The Kogrukluk River and Koktuli River populations, on the other hand, are at least four times larger than the Anvik River

population and may not have been effected as substantially by genetic drift. Individuals that did not fall within major clusters were found in the Kogrukluk River and Anvik River populations. These individuals may represent evidence of gene flow from genetically diverged upriver populations but could have also resulted from within population variation. Individual-based PCA using 39 SNPs from Templin et al. (2011) did not resolve the population structure that was observed with the RAD data and displayed no apparent clustering pattern. These results emphasize the utility of genome wide data when attempting to elucidate patterns of population differentiation.

Assignment accuracies with both panels containing over 10,000 SNPs were $\geq$ 89% for all populations while assignment accuracies with the panels containing 39 SNPs were close to 50% on average per population. Additionally, the inclusion of outlier loci only slightly improved assignment accuracy. These results indicate that a large number of neutral SNPs was sufficient to achieve precise assignment and that, for our analysis, the number of SNPs used seemed to have more influence on assignment accuracies than the resolution of those SNPs. Unfortunately, we were unable to evaluate the effectiveness of small panels of high-resolution SNPs compared to large panels of neutral SNPs because this type of analysis requires the use of a training and holdout dataset (Anderson 2010), which was not feasible with the sample sizes in our study. The patterns of population divergence observed here are consistent with previous studies suggesting structuring of Chinook salmon populations on regional scales (Templin *et al.* 2011). Despite this, sampling additional populations from each region would likely improve estimates of population divergence and assignment accuracy.

Demography

Estimating and interpreting N_e in salmon populations using single samples can be difficult because multiple cohorts are often present (Waples 1990a). N_e estimates therefore reflect a value somewhere between the effective number of breeders in a given year and the effective number of breeders per generation. We divided N_e by the census size (escapement) per year (N) and the census size per generation (NG) to account for both of these possibilities when comparing N_e to census size. The N_e/N and N_e/NG ratios were highly variable across our populations, indicating that effective and census size are not well correlated in our study system. A meta-analysis of 251 estimates of N_e/N found a median value of 0.14 and also showed that N_e/N ratios are generally larger in smaller populations (Palstra & Ruzzante 2008). Larger N_e/N

ratios in smaller populations were also observed in our data. For example, the Anvik River had a census size of 1,700 and N_e/N of 0.30 while the Kogrukluk River had a census size of 12,000 and an N_e/N of 0.17. This trend, however, was not consistent in the Tubutulik River population which had a census size of 3,100 and N_e/N ratio of 0.62.

The large N_e/N ratio in the Tubutulik River population may have been due to gene flow from proximate populations which can introduce additional genetic diversity and inflate estimates of N_e (Palstra & Ruzzante 2011). The Tubutulik River is a small river in Norton Bay, which contains at least five additional salmon producing rivers. Gene flow among these rivers may be quite common and could therefore have resulted in the larger than expected N_e/N estimates that we observed. Gene flow from proximate populations may also be inflating N_e estimates from the Koktuli River and the Big Salmon River as both of these collections have census sizes close to 5,000 but N_e estimates with confidence intervals including infinity. It is important to note that estimates of census size are approximate and may not be completely representative. Nevertheless, our results suggest that census size is not an adequate predictor of effective size, especially in populations that may belong to a larger metapopulation.

Removing comparisons between loci on the same linkage group appeared to have a non-uniform effect on estimates of N_e with larger estimates being more affected by removing linkage. For example, the estimate of N_e for the Anvik River population, the smallest population in the study, only changed by 10 when linked comparisons were removed whereas the estimate for the Big Salmon River changed by almost 9,000. This relatively small bias for small populations was also found by Sved et al. (2013), and is expected given that, in small populations, the signal of linkage disequilibrium due to genetic drift should be large compared to the signal due to physical linkage.

It also appears that N_e estimates for populations of similar size can be affected non-uniformly by physically linked loci. Specifically, estimates of N_e for the Tubutulik River displayed a larger downward bias when physically linked loci were included than estimates for the Kogrukluk River, even though the effective sizes for these populations were similar with unlinked loci. The non-uniform effects we observed when removing physically linked loci may be due to historic signals of N_e that have been preserved due to linkage (Hill 1981; Tenesa *et al.* 2007).

Estimating N_e in large populations (N_e > 1,000) with 10-100 genetic markers is extremely challenging due to the small amount of linkage disequilibrium caused by drift (Waples & Do 2010), but, with thousands of markers, accurately characterizing the signal of drift and estimating N_e may be feasible (Allendorf *et al.* 2010). Estimates of N_e from our study were infinite for three out of five populations with 39 SNPs but only infinite for one population with 1,118 SNPs. Additionally, all estimates with 39 SNPs but only two estimates with 1,118 SNPs displayed confidence intervals including infinity, and confidence intervals were consistently smaller with 1,118 SNPs. Our results indicate that genomic data can improve the accuracy of N_e estimates in large populations, aiding management in many species.

Putative adaptive divergence

We identified 6.7% of SNPs in our dataset as outliers, consistent with past studies identifying 5-10% of markers as candidates for directional selection (Nosil *et al.* 2009). In general, patterns of divergence observed from our outliers were similar to patterns obtained using neutral markers. *BLAST* annotation of outlier loci revealed a high frequency of transposable elements, similar to the overall dataset. These transposable elements are quite common in teleost fish and are generally assumed to behave as neutral markers (Radice *et al.* 1994) although some evidence suggests that they can be adaptively important (Casacuberta & González 2013).

Tests for genomic regions under putative selection revealed that these regions appeared to be spread randomly across the genome with few common "hotspots" among populations. This pattern is consistent with Bourret et al. (2013), who found a similar distribution across the Atlantic salmon genome. Despite the apparent randomness, three regions were differentiated in more than five out of ten population comparisons. These highly divergent regions may represent adaptively significant areas of the Chinook salmon genome and should be targets of future research. Population comparisons that included the Big Salmon River generally displayed the largest number of divergent regions. Although these regions likely represent adaptively significant areas of the genome, it is possible that at least a portion of them resulted from genetic drift as a result of isolation during the last glacial maximum (Olsen *et al.* 2011). Research aimed at disentangling signatures of drift from those of natural selection should therefore focus on systems with low neutral divergence across heterogeneous environments (Nielsen *et al.* 2009).

Management and conservation implications for western Alaska

Returns of Chinook salmon to western Alaska have fallen dramatically over the last decade compared to their long term average (ADF&G 2013). This precipitous decline has prompted multiple fisheries closures causing extensive economic hardship and threatening subsistence catches for natives of the western Alaska region. Some of these closures stem from the inability of fisheries managers to differentiate a late run that is of normal size from a small run that is returning at a normal date. One way that managers can differentiate these two scenarios is with stock composition estimates facilitated by panels of high-throughput SNPs. Specifically, stock composition estimates from mixed-stock fisheries and test fisheries on the high seas can be used to monitor the contribution of each stock in real-time, helping to inform the need for fisheries closures and generally improving fisheries management (Seeb *et al.* 2000; Smith *et al.* 2005c; Dann *et al.* 2013). Despite this potential utility, tools for genetic stock identification in marine waters of western Alaska have been severely hampered by lack of genetic divergence among regions (Templin *et al.* 2011). Our data provide the first evidence that assignment to region of origin is feasible in western Alaska despite low levels of divergence. Although it is not currently possible to screen 10,000 loci on thousands of individuals, a subset of our RAD loci that show high levels of divergence can be used to construct a high-throughput SNP panel to differentiate stocks in this region (c.f., Ackerman *et al.* 2011).

This high-throughput SNP panel could also be used to investigate the migration and distribution patterns of Chinook salmon on the high-seas (e.g., Murphy *et al.* 2009; Larson *et al.* 2013). Patterns of productivity in the marine environment are thought to be a major cause of the fluctuations in abundance observed in Chinook salmon from western Alaska (Farley *et al.* 2005). Despite this, most stock assessment models assume a constant marine mortality rate across all stocks. The ability to monitor stock-specific abundance on the high-seas could provide important information for stock assessment models which is currently unavailable. Additionally, stock composition estimates could be used to monitor the impact of Chinook salmon interception in the Bering Sea pollock fishery; this fishery has captured as many as 100,000 Chinook salmon in a single year (Gisclair 2009). In summary, our results represent the first step towards a panel of high-throughput SNPs that can be used to conduct genetic stock identification and improve stock-specific management in the western Alaska region.

Applicability to other study systems

Our study demonstrates the utility of genomic data when attempting to differentiate closely related populations and estimate demographic parameters. The methods we employed will be especially applicable in marine species, which are often characterized by low genetic differentiation and large population sizes (Waples 1998; Nielsen & Kenchington 2001). For example, individual-based analyses with thousands of markers can provide extremely accurate estimates of individual genetic variation. Additionally, this method can shed light on patterns of connectivity by identifying migrants and admixture within populations.

Estimates of N_e in large marine populations can also be improved using approaches similar to ours (e.g., Gruenthal *et al.* 2013). Dense linkage maps have already been developed for many marine species including cod (Hubert *et al.* 2010), flounder (Castano-Sanchez *et al.* 2010) and shrimp (Du *et al.* 2010). By combining these linkage maps with genomic data it may be possible to accurately estimate N_e and N_e/N in many economically important marine species. These estimates can provide important insights into the adaptive potential of marine populations and can be used to inform management (Hare *et al.* 2011).

Summary

Our results demonstrated fine-scale structure between regions in western Alaska. This structure allowed us to assign fish back to their region of origin with greater than 90% accuracy, representing a significant improvement over past studies. We also estimated N_e for each population using a novel method for non-model organisms. Estimates were generally large and provided some evidence that metapopulation dynamics influence demography in this region. Investigation of loci and genomic regions under putative selection found three potential regions of adaptive divergence. The methods described in our study will be particularly applicable to marine species or any species where large population size and shallow structure are common.

Tables

Table 2.1. Populations analyzed in this study with year sampled, sample size (N), observed heterozygosity (H_O), and expected heterozygosity (H_E).

Sampling location	Region	Year	GPS coordinates	N	H_O	H_E
Tubutulik River	Norton Sound	2009	64.740, -161.888	56	0.248	0.252
Anvik River	Lower Yukon R	2007	62.681, -160.214	54	0.260	0.261
Kogrukluk River	Kuskokwim R	2007	60.841, -157.846	57	0.251	0.258
Koktuli River	Bristol Bay	2010	59.935, -156.427	56	0.256	0.259
Big Salmon River	Upper Yukon R	2007	61.867, -134.917	47	0.232	0.232

Table 2.2. Pairwise F_{ST} values calculated using 10,944 SNPs and number of genomic regions that were under putative selection (in parentheses). All pairwise comparisons are significantly differentiated ($P < 0.01$).

	Tubutulik River	Anvik River	Kogrukluk River	Koktuli River
Anvik River	0.030 (20)			
Kogrukluk River	0.027 (20)	0.005 (20)		
Koktuli River	0.028 (20)	0.006 (23)	0.003 (20)	
Big Salmon River	0.098 (24)	0.075 (25)	0.075 (21)	0.077 (25)

Table 2.3. Results from two AMOVAs with 10,944 SNPs.

Source of Variation	d.f.	Percentage of variation
(1) All populations		
Among groups	2	5.26
Among populations within groups	2	0.45
Within populations	529	94.32
(2) Big Salmon River excluded		
Among groups	1	2.41
Among populations within groups	2	0.43
Within populations	436	97.18

Table 2.4. Results of leave-one-out tests for individual assignment with four SNP panels. Panels are: 1) **39 EST**: 39 SNPs previously developed for Chinook salmon from expressed sequence tags (ESTs, Templin *et al.* 2011), 2) **39 RAD**: 39 randomly chosen SNPs from the complete dataset of 10,944 RAD SNPs, 3) **10,944 RAD**: the complete dataset of RAD SNPs, and 4) **10,211 RAD no outliers**: the full set of RAD SNPs with the 733 outlier SNPs that were found to be under putative selection removed. Individuals were considered to be correctly assigned if the assignment probability to population of origin was higher than to any other population. See Table S3 for assignment probabilities for each individual.

	% Correct Assignment			
Regions	39 EST	39 RAD	10,944 RAD	10,211 RAD no outliers
Tubutulik River	67	65	100	100
Anvik River	46	30	91	89
Kogrukluk River	34	30	93	93
Koktuli River	29	30	98	95
Big Salmon River	96	87	100	100

Table 2.5. Estimates of effective population size (N_e) for five populations calculated with 1,118 RAD-derived SNPs that were placed on the linkage map and 39 of the 43 SNPs that were in linkage equilibrium from Templin et al. (2011). Estimates with RAD SNPs are calculated using only comparisons between loci on different linkage groups (N_e linkage removed) and all comparisons (N_e all data). The ratio of effective population size to census size (N_e/N) and effective population size to census size multiplied by generation length (N_e/NG) for each population is also reported (G is generation length and N is an approximate value of yearly escapement for each population, see methods). The N_e used for these calculations is N_e linkage removed (column 2). We did not calculate N_e/N or N_e/NG for the Bristol Bay and upper Yukon populations because confidence intervals included infinity, suggesting our point estimates may not be completely representative.

Population	N_e linkage removed	N_e all data	N_e 39 SNPs	G	N	N_e/N	N_e/NG	Source of N	Source of G
Tubutulik River	1909 (1,295-3,602)	808 (674-1,009)	Inf (174-Inf)	5.43	3,100	0.62	0.11	(Banducci *et al.* 2007)	(Lingnau 1996)
Anvik River	516 (451-604)	505 (443-586)	209 (65-Inf)	5.48	1,700	0.30	0.06	(Howard *et al.* 2009)	(Sandone 1995)
Kogrukluk River	2,026 (1,375-3,825)	1,723 (1,233-2,842)	Inf (134-Inf)	5.20	12,000	0.17	0.03	(Williams & Shelden 2011)	(Howard *et al.* 2009)
Koktuli River	Inf (6,055-Inf)	26,071 (3,733-Inf)	Inf (Inf-Inf)	5.13	6,000	N/A	N/A	(Woody 2012)	(Howard *et al.* 2009)
Big Salmon River	13,101 (1,505-Inf)	4,243 (1,806-Inf)	520 (70-Inf)	5.65	5,000	N/A	N/A	(Mercer & Wilson 2011)	(Howard *et al.* 2009)

Figures

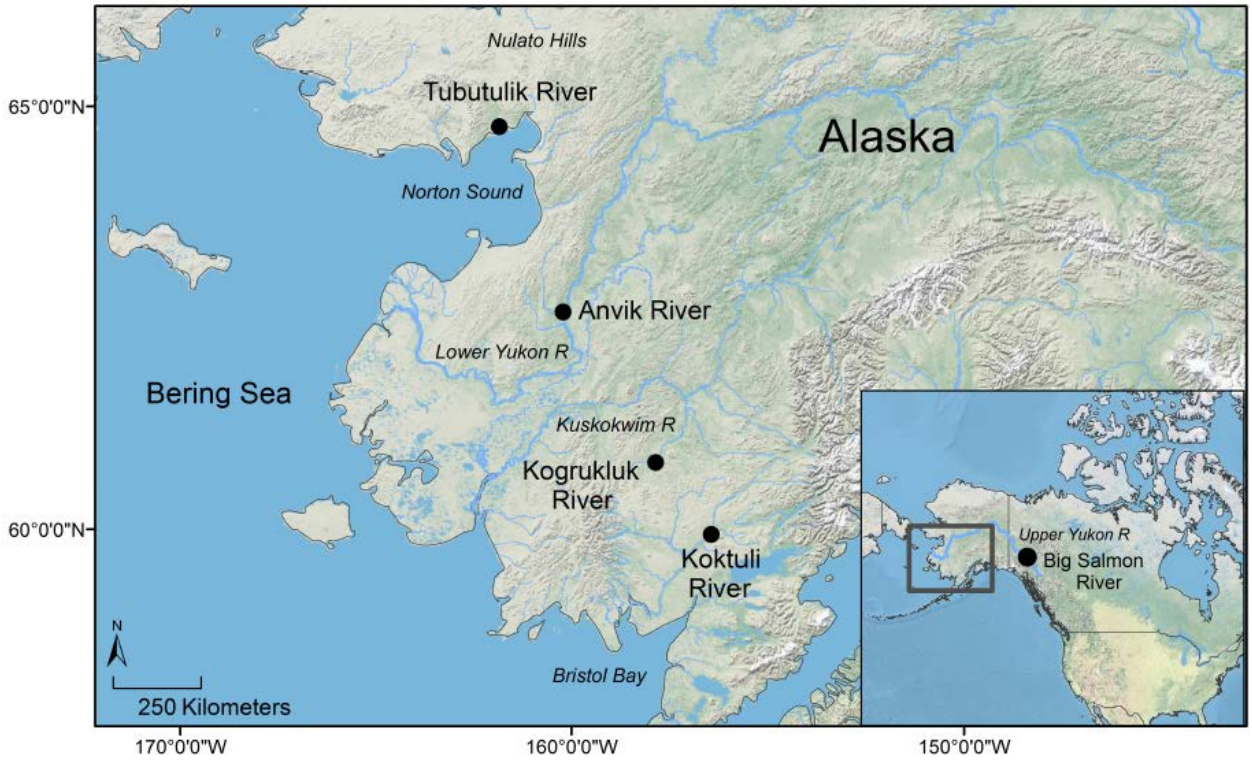

Fig. 2.1. Map of sampling locations. See Table 2.1 for additional details about each sampling site.

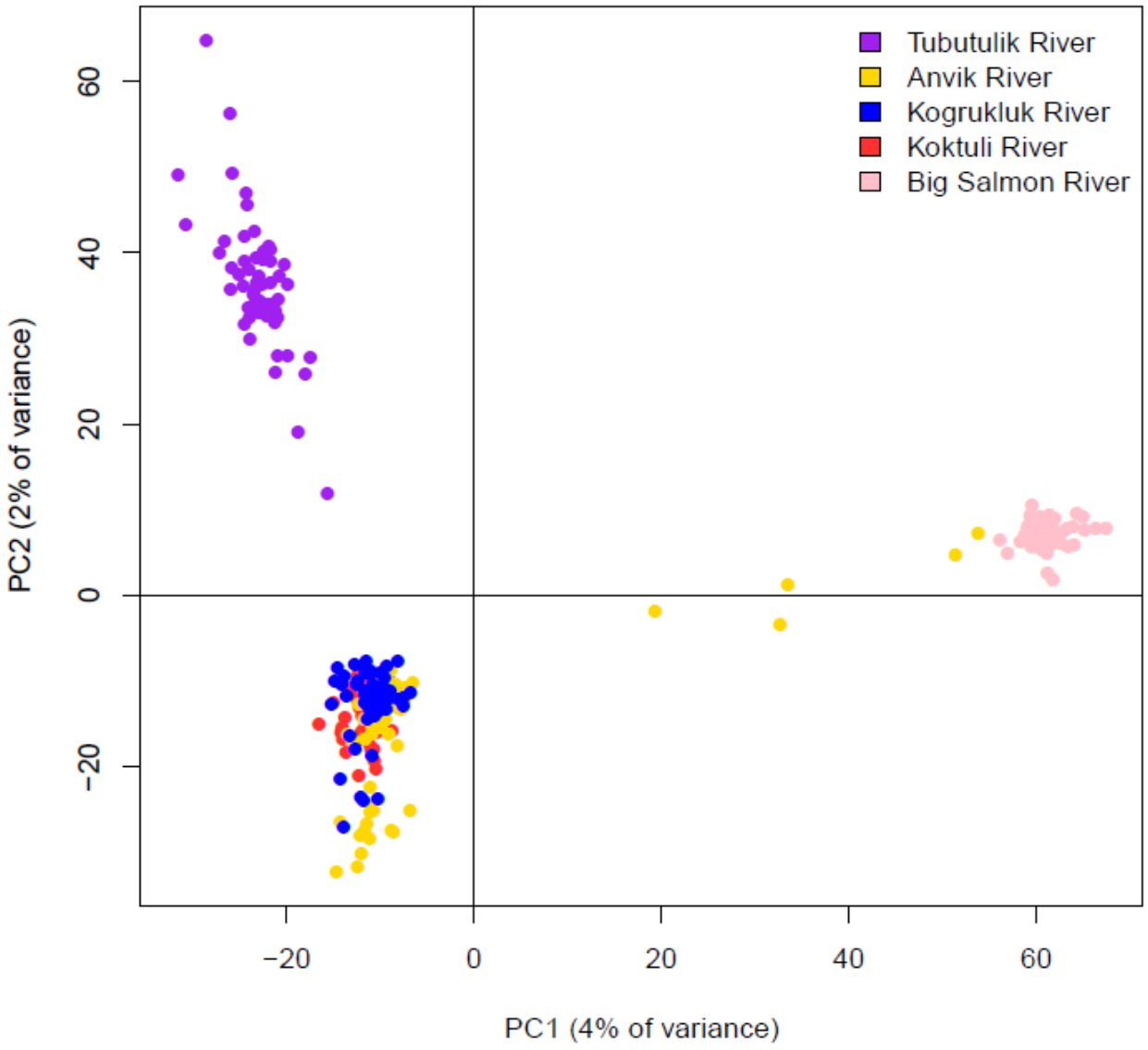

Fig. 2.2. Individual-based principal component analysis for all populations and 10,944 SNPs. The five intermediate individuals from the Anvik River were removed from further analyses (see methods).

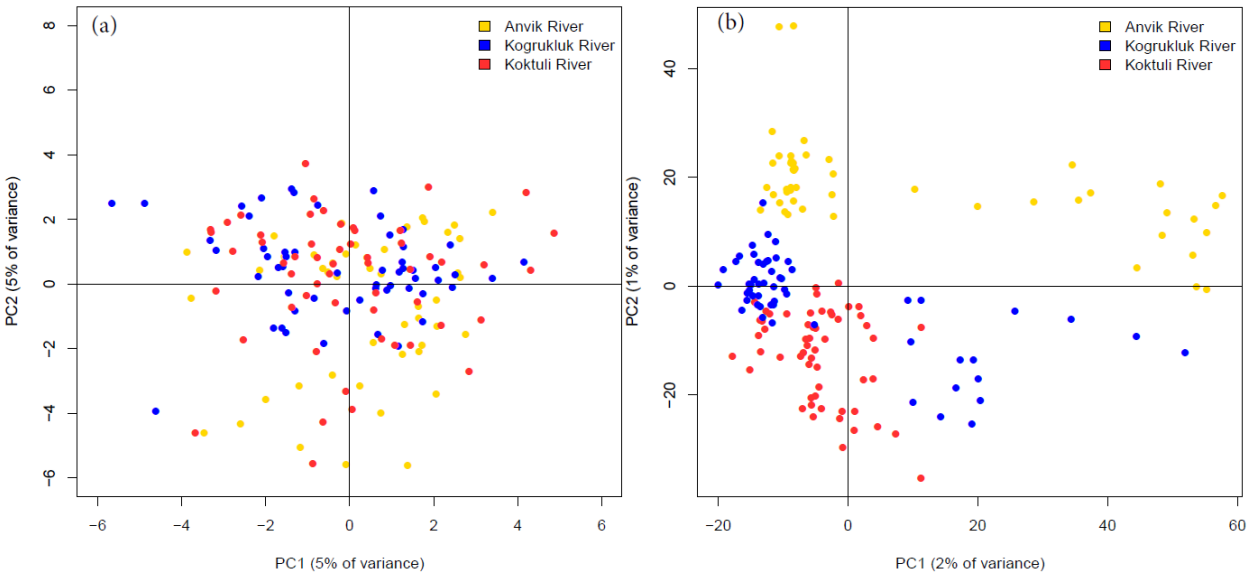

Fig. 2.3. Individual-based principal component analysis for the Anvik River, Kogrukluk River and Koktuli River populations using (a) 39 SNPs from Templin et al. (2011) and (b) 10,944 RAD SNPs.

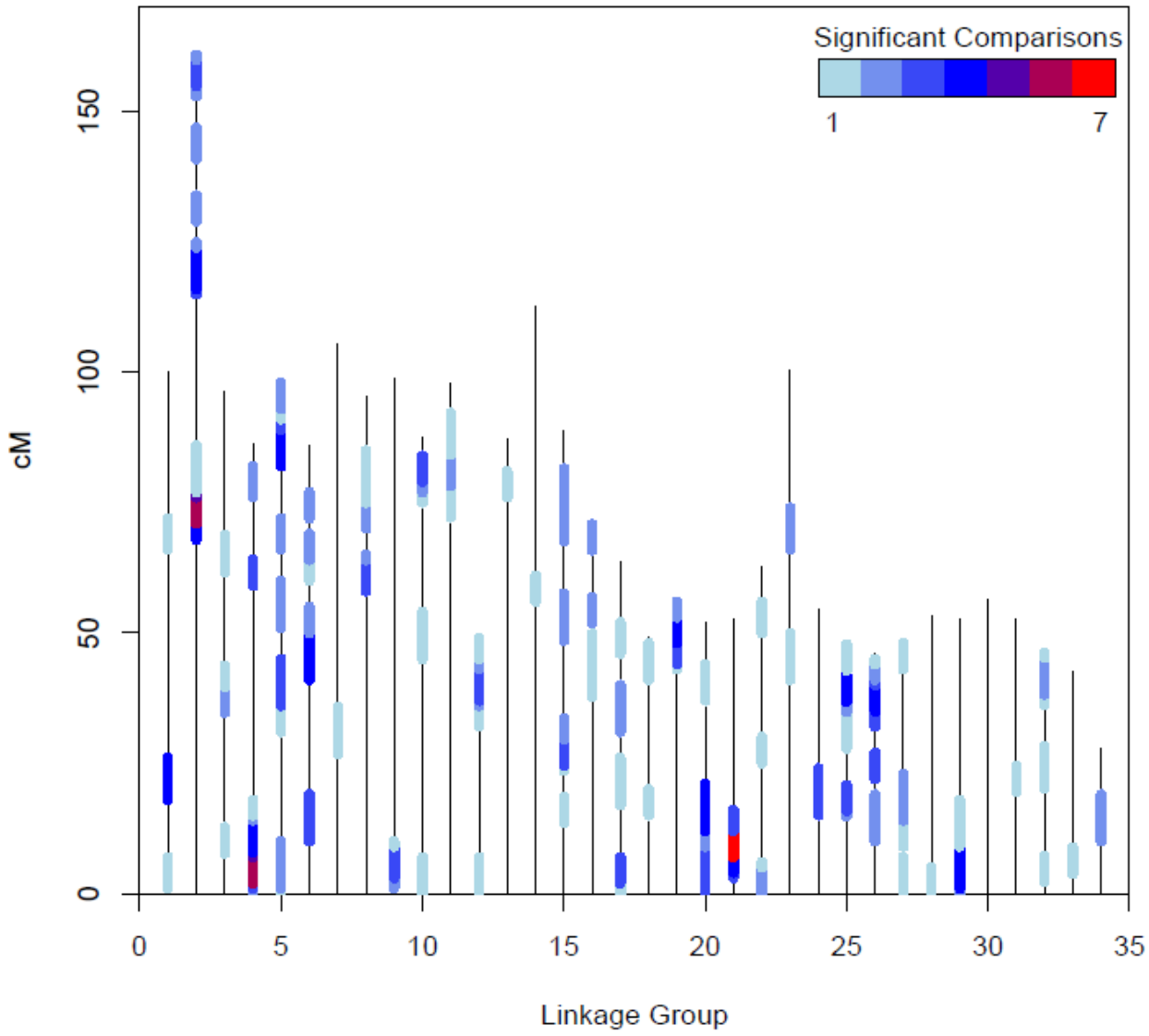

Fig. 2.4. Regions of the genome under putative selection as inferred by pairwise F_{ST} across all population pairs. Each vertical line represents a linkage group and the length of the line is proportional to the size of the linkage group in cM. Shaded areas indicate regions which are significantly diverged in at least one population pair indicating putative selection. The color of the shading corresponds to the number of significant pairwise population comparisons with red and purple indicating over half of the population pairs are divergent in the given region.

Chapter 3

SNPs identified through genotyping-by-sequencing improve genetic stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) from western Alaska[3]

Abstract

Genetic stock identification (GSI), an important tool for fisheries management that relies upon the ability to differentiate stocks of interest, can be difficult when populations are closely related. Here we genotyped 11 850 single-nucleotide polymorphisms (SNPs) from existing DNA sequence data available in five closely-related populations of Chinook salmon (*Oncorhynchus tshawytscha)* from western Alaska. We then converted a subset of 96 of these SNPs displaying high differentiation into high-throughput genotyping assays. These 96 SNPs (RAD96) and 191 SNPs developed previously (CTC191) were screened in 28 populations from western Alaska. Regional assignment power was evaluated for five different SNP panels including a panel containing the 96 SNPs with the highest F_{ST} across the CTC191 and RAD96 panels (F_{ST}96). Assignment tests indicated that SNPs in the RAD96 were more useful for GSI than those in the CTC191 and that increasing the number of reporting groups in western Alaska from one to three was feasible with the F_{ST}96. Our approach represents an efficient way to discover SNPs for GSI and should be applicable to other populations and species.

Introduction

Genetic tools have been used to document biodiversity and to manage wild populations for over four decades (Utter 2004; Waples *et al.* 2008). These techniques are particularly applicable to Pacific salmon (*Oncorhynchus spp.*); salmon return to their natal streams with high

[3] Full citation: Larson, W.A., J.E. Seeb, C.E. Pascal, W.D. Templin, and L.W. Seeb. 2014. SNPs identified through genotyping-by-sequencing improve genetic stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) from western Alaska. *Canadian Journal of Fisheries and Aquatic Sciences*, **71**(5): 698-708. Supplementary material available from the online version of this manuscript (http://www.nrcresearchpress.com/doi/abs/10.1139/cjfas-2013-0502#.Vc0hhPlViko).

fidelity, promoting local adaptation and the formation of genetically distinct populations (Shaklee *et al.* 1999; Stewart *et al.* 2003; Neville *et al.* 2006). Discrete management of these populations minimizes extirpation of lineages with smaller population sizes and preserves the resiliency of the species as a whole (Hilborn *et al.* 2003; Schindler *et al.* 2010).

As genetic techniques improved, genetic stock identification (GSI) became a commonly utilized tool for managing discrete populations of Pacific salmon (Dann *et al.* 2013). GSI uses the observed allelic frequencies of baseline populations sampled on the spawning grounds to infer the natal origin of fish captured in mixed-stock fisheries (Milner *et al.* 1985; Utter & Ryman 1993; Beacham *et al.* 2012). Population-specific assignment is rarely feasible; therefore, baseline datasets are often partitioned into reporting groups composed of genetically similar populations. The proportional contribution of each reporting group to mixed-stock samples is then estimated. GSI has been used to investigate the migration and distribution patterns of many Pacific salmonids (e.g., Habicht *et al.* 2010; Tucker *et al.* 2011; Larson *et al.* 2013) and to inform in-season management of mixed-stock fisheries (e.g., Seeb *et al.* 2000; Beacham *et al.* 2008b; Dann *et al.* 2013).

The genetic marker of choice for GSI has evolved dramatically over the past three decades with allozymes being replaced by microsatellites and, most recently, by single-nucleotide polymorphisms (SNPs, Schlotterer 2004; Hauser & Seeb 2008). Compared to microsatellites, SNPs can be developed and assayed more quickly, and the resulting genotypes are easily transferred among laboratories (Seeb *et al.* 2011a). Recent advances in genomic techniques have made it possible to screen thousands of putative SNPs in hundreds of individuals (reviewed in Allendorf *et al.* 2010; Narum *et al.* 2013). Researchers can then select SNPs that display elevated levels of differentiation among populations of interest and convert them to high-throughput genotyping assays for screening thousands of individuals. This type of approach has already been used to assess hybridization between two species of trout (Hohenlohe *et al.* 2011; Amish *et al.* 2012) and promises to be extremely applicable to the development of SNP panels for GSI (Storer *et al.* 2012).

Chinook salmon (*Oncorhynchus tshawytscha)* from western Alaska represent an excellent opportunity to apply genomic techniques towards the development of a SNP panel for GSI. Chinook salmon primarily spawn in drainages in four major regions in western Alaska: Norton Sound, Yukon River, Kuskokwim River, and Bristol Bay (Templin *et al.* 2011). Recent

returns to all four regions have been substantially lower than their long-term average, renewing interest in the migration patterns and relative vulnerability of these stocks to both targeted and bycatch fisheries (Stram & Ianelli 2009; ADF&G 2013). GSI could be used to investigate the above questions, but a lack of significant genetic differentiation among these regions has prevented its use. Specifically, evidence of shallow genetic structure among regions has been reported with allozyme (Gharrett *et al.* 1987) and SNP (Templin *et al.* 2011) data while microsatellite data showed no apparent structure (Olsen *et al.* 2011). The limited discriminatory power of these existing baseline datasets necessitated the pooling all four regions in Western Alaska into a single reporting group for GSI estimates (Templin *et al.* 2011; Larson *et al.* 2013). Nevertheless, given the apparent substructure suggested by previous studies, a search for additional SNPs that can differentiate the major regions in western Alaska is warranted.

Our goals were (1) to use genotyping-by-sequencing to develop a new set of 96 information-rich SNPs for western Alaska, (2) to compare the resolving power of these 96 new SNPs to 191 existing SNPs, and (3) to construct the best panel of 96 SNPs for GSI from all available SNPs. Panel sizes of 96 were selected because this represents the maximum number of SNPs that can be assayed simultaneously using the most prevalent genotyping platform for Pacific salmon management, the Fluidigm 96.96 dynamic array (Fluidigm, South San Francisco, California).

We identified 11 850 putative SNPs in five populations of Chinook salmon from western Alaska using data from restriction-site-associated DNA (RAD) sequencing obtained by Larson *et al.* (2014c). We then developed high-throughput assays for 96 RAD-derived SNPs showing high levels of differentiation. The 96 RAD-derived SNPs along with 191 SNPs developed previously for Chinook salmon were genotyped in 28 populations from across western Alaska. From these data, we compared the resolving power of the 96-RAD derived SNPs to the 191 previously developed SNPs, identified the 96 SNPs that displayed the highest levels of differentiation across these two panels, and tested the utility of the top 96 SNPs for GSI. Using the top 96 SNPs we were able to increase the number of reporting groups for GSI in western Alaska from one to three. Based on these results, we believe that SNP discovery using genomic techniques can improve GSI in populations characterized by low genetic divergence.

Materials and methods

Tissue sampling

Fin clips preserved in 100% ethanol were available from 28 populations of Chinook salmon collected throughout coastal western Alaska and the middle and upper Yukon River (2 275 fish total, 21 populations shared with Templin *et al.* 2011, Table 3.1, Fig. 3.1). Five populations that spanned the study area were RAD sequenced by Larson *et al.* (2014c). These five ascertainment populations did not have unusually small census sizes and were genetically similar to proximate populations (Templin *et al.* 2011). All 28 populations were then used to evaluate the resolving power of the RAD-derived and previously developed SNPs. Chinook salmon from the upper and to a lesser extent middle Yukon River are highly differentiated from those of coastal western Alaska (Smith *et al.* 2005c; Templin *et al.* 2011). We included populations from this region to anchor inferences of population structure and ensure that GSI outside of coastal western Alaska was feasible with the SNPs discovered in this study. Collections from multiple years were pooled if sample sizes were < 48 following recommendations of Waples (1990b).

Quality filtering and SNP discovery

Raw RAD sequence data (single-end, 100 base pair target length) were available from Larson *et al.* (2014c). Quality filtering, SNP discovery, and genotyping were performed on these data using a modified version of the pipeline first described in Miller et al. (2012) and adapted by Everett et al. (2012). The last base pair of each read was trimmed, and reads with < 90% chance of being error-free were discarded. A separate file was then created for each individual containing all of their unique sequences and the number of times they occurred. Sequences occurring < 6 or > 200 times were removed. We only used the 16 individuals with the most data from each population for SNP discovery to reduce the frequency of false positives in our dataset. Putative SNPs within each individual were identified with the program NOVOALIGN 2.07 (www.novocraft.com) using the following alignment parameters: maximum of 10 alignments returned per unique sequence and a maximum alignment score of 245. Alignments for each individual were filtered using the methods described in Miller et al. (2012) to retain RAD tags with a single putative SNP that did not align closely to any other sequence. Polymorphism data from each individual were combined to form a catalog of RAD tags, each containing a single, bi-allelic putative SNP. This catalog was aligned to each individual using Bowtie V0.12.9

(Langmead *et al.* 2009), and sequence counts for each allele were tabulated using the methods of Miller et al. (2012). Genotypes were obtained from allele counts using a two-allele maximum likelihood approach following the framework of Hohenlohe (2010) with a static error rate based on the published value for Illumina HiSeq data (0.0016, Minoche *et al.* 2011).

As an initial screen for paralogous sequence variants (PSVs), we genotyped 33 individuals from a haploid family, available from another study (Everett & Seeb 2014), and removed loci with > 10% heterozygosity. PSVs are closely related sequences from different genomic locations that are abundant in salmonids as a result of an ancient whole-genome duplication event (Allendorf & Thorgaard 1984; Seeb *et al.* 2011b). Although PSVs are difficult to genotype accurately because they do not segregate as single loci (Gidskehaug *et al.* 2011), haploid individuals can be used to differentiate true SNPs from PSVs because true SNPs will be homozygous in all haploid individuals whereas PSVs will often be heterozygous (Hecht *et al.* 2013).

Allele frequencies and sample sizes for each putative SNP were calculated using GENEPOP 4 (Rousset 2008) to enable the removal of uninformative or unreliable loci. Putative SNPs that failed to genotype in > 80% of individuals and those with minor allele frequencies < 0.1 in all populations were removed. As a final filtration step, we removed individuals with < 10X average coverage across the filtered SNPs because these individuals likely contained a substantial amount of missing data that could cause genotyping errors.

It is important to note that SNPs discovered in this study are not necessarily the same as those discovered in Larson *et al.* (2014c) because Larson *et al.* (2014c) used the STACKS software package (Catchen *et al.* 2011; Catchen *et al.* 2013) for SNP discovery.

Paired-end assembly and BLAST annotation

Paired-end data (100 x 2 base pair target length) were available from eight Chinook salmon collected in coastal western Alaska (Larson *et al.* 2014c). Paired-end assemblies for each locus were conducted using the methods of Etter et al. (2011) and adapted by Everett et al. (2012) to increase query lengths for BLAST annotation and template length for assay design. We used the program VELVET 1.1.06 (Zerbino & Birney 2008) to create a consensus sequence for each locus using all the paired and single-end reads that aligned to that locus. Consensus sequences for each locus were aligned to the Swiss-Prot database using the BLASTX search

algorithm. Alignments with E-values of $\leq 10^{-4}$ were retained. If multiple alignments had E-values of $\leq 10^{-4}$ for the same locus, then the alignment with the lowest E-value was retained.

Construction of high-throughput assays from RAD data

We selected 150 RAD-derived SNPs that displayed high levels of differentiation in our ascertainment populations for conversion to the 5'-nuclease reaction (Holland *et al.* 1991) with TaqMan chemistry (Life Technologies, Grand Island, New York), a chemistry commonly used on high-throughput genotyping platforms (Seeb *et al.* 2009a). Genetic differentiation among our ascertainment populations was estimated across all RAD-derived loci with overall F_{ST} values (Weir & Cockerham 1984) calculated in GENEPOP. We also calculated pairwise F_{ST} values and conducted exact tests of Hardy-Weinberg equilibrium for each locus in GENEPOP.

The 150 SNPs were chosen in an iterative fashion. First we choose the 150 SNPs with the highest overall F_{ST} across the Bristol Bay (Koktuli River), Kuskokwim River (Kogrukluk River), and lower Yukon River (Anvik River) populations and tested this panel's assignment power with 100% simulations conducted in the program ONCOR (see below for further details). We then modified the panel by adding and removing SNPs until we were able to find the 150 SNPs that achieved the highest possible assignment accuracies for all ascertainment populations. We did not choose SNPs that differentiated the upper Yukon River (Big Salmon River) or Norton Sound (Tubutulik River) populations because these populations were highly differentiated from all others and could likely be resolved with any SNP panel (Table 3.2).

We limited our selections for conversion to the 5'-nuclease reaction to SNPs that were in Hardy-Weinberg equilibrium in at least three of the five populations ($P > 0.05$). Also, we chose SNPs that were located past base pair 34 of the RAD tag in order to accommodate the primer/probe configuration of the 5'-nuclease reaction. Paired-end data were used to increase template length for assay design if no primer/probe configuration was feasible with the single-end reads.

Successfully designed assays were genotyped on 24 fish from each of the four ascertainment populations from coastal western Alaska (populations 2, 6, 16, and 22; 96 fish total). Genotyping was conducted with preamplification according to the methods of (Smith *et al.* 2011). Assays that did not amplify or produce consistent results were discarded, and the 96 assays with the highest overall F_{ST} across the Bristol Bay, Kuskokwim River, and lower Yukon

River populations based on the RAD data were retained to form a panel of 96 RAD-derived SNPs, hereafter referred to as the RAD96.

Selection and evaluation of SNP panels

Two major goals of this study were (1) to evaluate the resolving power of the RAD96 compared to a panel of previously developed SNPs, and (2) to construct the best possible panel of 96 SNPs to discriminate stocks in western Alaska from all SNPs available. To achieve these goals, we first genotyped 2 275 fish from 28 populations throughout western Alaska for the RAD96 and 191 SNPs previously developed for Chinook salmon. The 191 previously developed SNPs (hereafter referred to as CTC191) were mainly chosen for applications south of Alaska as part of a project funded by the Pacific Salmon Commission's Chinook Technical Committee (Warheit *et al.* 2013); these originated primarily from expressed sequence tags (Smith *et al.* 2005a; Smith *et al.* 2005b; Smith *et al.* 2007; Campbell & Narum 2008; Campbell & Narum 2009; Clemento *et al.* 2011; Warheit *et al.* 2013).

SNPs from the CTC191 and RAD96 were genotyped using the 5'-nuclease reaction with pre-amplification (Smith *et al.* 2011), and the reproducibility of our results was quantified by re-genotyping four of every 95 (4%) fish at all loci. Individuals with > 5% missing genotypes were excluded from further analyses. Tests for deviation from Hardy-Weinberg and linkage equilibrium were conducted for each locus across all 28 populations in GENEPOP, and loci out of equilibrium in > 50% of the populations ($P < 0.05$) were removed. Observed and expected heterozygosities for each locus were calculated in GenALEx 6.5 (Peakall & Smouse 2012) and overall F_{ST} (Weir & Cockerham 1984) for each locus was calculated in GENEPOP. Calculations of locus-specific heterozygosity and F_{ST} were conducted using populations 1-24 (excluding population 2, see below for justification).

Genetic differentiation across all 28 populations was estimated separately for the CTC191 and RAD96 panels with pair-wise F_{ST} values (Weir & Cockerham 1984) calculated in GENEPOP to compare the patterns of population structure resolved by each panel. We then conducted principal coordinate analysis (PCoA) in GenAlEx for each panel to visualize patterns of population structure. Populations 25-28 from the middle and upper Yukon River were not included in the PCoA because these populations are highly differentiated from those of coastal western Alaska according to previous studies (Smith *et al.* 2005c; Templin *et al.* 2011).

Including these populations may have prevented us from detecting signals of differentiation among the remaining populations.

After comparing the CTC191 and RAD96 panels, all SNPs were ranked by overall F_{ST} across populations 1-24 (excluding population 2), and the top 96 SNPs were chosen to create the final panel for GSI, referred to hereafter as the F_{ST}96. The Tubutulik River (population 2) was excluded from this analysis because it was a genetic outlier (Fig. 3.2). SNPs were ranked by F_{ST} because this method produced the highest performing panels for population assignment in recent analyses of multiple ranking methods (Storer *et al.* 2012; Warheit *et al.* 2013).

The assignment accuracy of the complete dataset (281 SNPs), the F_{ST}96, the CTC191, the RAD96, and 96 randomly chosen SNPs from the complete dataset was evaluated with the 100% simulation method described in Anderson et al. (2008) and implemented in ONCOR (www.montana.edu/kalinowski/) with the default parameters. The simulation method implemented in ONCOR simulates a mixture sample where all individuals are from the same population and then uses maximum likelihood to determine the percentage of the sample that is correctly allocated back to the population and reporting group of origin. A minimum value of 90% correct assignment is typically required for a reporting group to be considered identifiable and robust for management applications (Seeb *et al.* 2000). Regional aggregations (reporting groups) for this analysis were Norton Sound, lower Yukon River, Bristol Bay-Kuskokwim River, middle Yukon River and upper Yukon River (Table 3.1). These groups were similar to the fine-scale reporting groups presented in Templin et al. (2011) with one exception: populations from Bristol Bay and the Kuskokwim River regions were combined into one reporting group because preliminary assignment tests were generally unable to differentiate these two regions. Assignment success between panels was compared with a Student's t-test. The small sample sizes for the Norton Sound and lower Yukon River collections prevented dividing datasets into separate training and holdout sets as suggested by Anderson (2010).

Results

RAD sequencing and SNP discovery

Sequence data from 284 Chinook salmon available from Larson *et al.* (2014c) were used to discover 26 567 putative SNPs. Filtration steps eliminated 1 602 potential PSVs and 13 115 loci with low minor allele frequencies and genotyping rates. Seventeen individuals with < 10X coverage across all filtered SNPs were removed (adjusted sample sizes in Table 3.1). The final

filtered dataset consisted of 267 individuals genotyped at 11 850 SNPs. The average depth of coverage across these individuals for the filtered SNPs was 29.1 (range 10.5 – 70.6).

Paired-end assembly and BLAST annotation

Paired-end assemblies produced 12 016 contigs with an average length of 268 bp (minimum 150 bp, maximum 565 bp). BLAST annotation of these contigs yielded significant hits for 1 466 of 11 850 SNPs, representing a 12% success rate. Of these hits, 547 (37%) aligned to transposable elements. Other common functional groups included DNA polymerases and structural proteins (Table S1).

Construction of high-throughput assays from RAD data

Assay design for the 5'-nuclease reaction was successful for 128 of the 150 assays attempted (Table 3.3). These 128 assays were tested in 96 fish, and 101 of them successfully amplified. The top 96 assays were retained to form the RAD96 panel (see methods). Paired-end data were required to design 47 of the 96 assays (49%), and BLAST annotations were successful for 9 of 96 assays (9%, see Table S2 for primer and probe sequences and BLAST annotations for the RAD96). A comparison of genotypes derived from RAD and 5'-nuclease data revealed 99% concordance between chemistries (Table 3.4). The most common type of error was a heterozygous 5'-nuclease genotype that was called a homozygote from RAD data, an expected result for data from next-generation genotyping (Nielsen *et al.* 2011).

Selection and evaluation of SNP panels

Our genotyping success rate for the 5'-nuclease reaction was 97% (2 275 of 2 355 samples), and our genotyping discrepancy rate, calculated from re-genotyping 4% of samples, was 0.03%. Four locus pairs were significantly out of linkage equilibrium in greater than half of the populations ($P < 0.05$). These marker pairs were *Ots_FGF6A* and *Ots_FGF6B_1* (28/28 populations), *Ots_RAD8200-45* and *Ots_RAD9480-51* (28/28 populations), *Ots_HSP90B-100* and *Ots_HSP90B-385* (24/28 populations) and *Ots_RAD11821* and *Ots_RAD3703* (16/28 populations). The marker with the highest F_{ST} for each pair was retained resulting in the removal of *Ots_FGF6A, Ots_RAD9480-51, Ots_HSP90B-100,* and *Ots_RAD11821* from further analyses. Significant deviations from Hardy-Weinberg equilibrium ($P < 0.05$) in more than half of the populations occurred for two loci, *Ots_111084b-619* (28/28 populations) and *Ots_111666-408* (28/28 populations); these loci were removed from further analyses. After removing SNPs that were out of Hardy-Weinberg and linkage equilibrium, 186 SNPs were retained from the

CTC191 and 95 were retained from the RAD96 (see Table S3 for summary statistics for each locus).

Patterns of population structure were similar between the CTC191 and RAD96 panels with populations from the Bristol Bay and Kuskokwim River regions forming a discrete cluster and populations from the lower Yukon River forming another cluster (Fig. 3.2). Populations from Norton Sound, however, did not form a single cluster and were generally distinct from all other populations. Populations from the middle and upper Yukon River (not shown in Fig. 3.2) were extremely differentiated from those of coastal western Alaska with both panels and displayed pairwise F_{ST} values that were at least two times larger than any within coastal western Alaska comparison (Table S4, S5). Although the CTC191 and RAD96 panels showed similar patterns of population structure, the mean F_{ST} and H_O were significantly higher for markers in the RAD96 compared to the CTC191 (CTC191: $H_O = 0.24$, $F_{ST} = 0.006$, RAD96: $H_O = 0.34$, $F_{ST} = 0.008$, $P < 0.0001$ for both Student's t-tests, Fig. 3.3).

After evaluating the CTC191 and RAD96 panels separately, we ranked all SNPs by overall F_{ST} across populations 1-24 (excluding population 2, F_{ST} ranks in Table S3). We then choose the top 96 to form the $F_{ST}96$ panel: 49 SNPs from the CTC191 and 47 SNPs from the RAD96. The $F_{ST}96$ panel was composed of 49% RAD-derived SNPs while RAD-derived SNPs composed 33% of the full dataset.

Assignment accuracies calculated with GSI simulations in ONCOR varied across panels but were generally highest with the $F_{ST}96$ and the complete dataset (Fig. 3.4, Table S6). The $F_{ST}96$ panel produced assignment accuracies to reporting group > 90% for 26 of 28 populations (88% for population 1, 89% for population 7) and the complete dataset produced accuracies > 90% for 25 of 28 populations (89% for population 1, 88% for population 5, 87% for population 7) while all other panels produced accuracies > 90% for fewer than 24 populations. Additionally, the $F_{ST}96$ and the full dataset significantly outperformed the panel of 96 randomly chosen SNPs, the CTC191, and the RAD96 ($P < 0.05$, Fig. 3.4, Table S6). Assignment rates were slightly higher for the complete dataset compared to the $F_{ST}96$ panel ($P = 0.04$) but the $F_{ST}96$ panel did outperform the complete dataset in three populations (3, 5, and 7). The CTC191 and RAD96 panels performed similarly ($P = 0.29$) despite the fact that the CTC191 panel contained almost twice as many SNPs.

Discussion

RAD sequencing for SNP development

We efficiently developed 96 novel high-throughput assays for GSI in western Alaska using data from RAD sequencing. Compared to previous methods for SNP discovery in Pacific salmon, mining RAD sequence data was quicker, required fewer validation steps, and facilitated directed SNP discovery for markers showing high levels of differentiation among populations. Specifically, mining RAD sequence data was much less time consuming than methods mining EST databases for putative SNPs (e.g., Smith *et al.* 2005b) and achieved an approximately 30% higher conversion rate to the 5' nuclease reaction (Smith *et al.* 2005a; Amish *et al.* 2012). This approach also represented a significant improvement over transcriptome-based methods, that require multiple validation steps and still achieve a conversion rate to the 5' nuclease reaction of less than 50% (Everett *et al.* 2011; Seeb *et al.* 2011b). Additionally, the discrepancy rate between genotypes obtained from RAD and 5' nuclease data was extremely low (1%).

Population Structure

General patterns of population structure in western Alaska were similar among the 11 850 RAD SNPs, the CTC191, and the RAD96 and are consistent with results from previous studies in the region (Olsen *et al.* 2011; Templin *et al.* 2011). The largest differentiation in all three datasets existed between populations from coastal western Alaska (populations 1-24) and those from the middle and upper Yukon River (populations 25-28, Table S4, S5). This pattern has been documented in numerous studies (e.g., Gharrett *et al.* 1987; Smith *et al.* 2005c; Beacham *et al.* 2008a) and is consistent with isolation during the last glacial maximum (Olsen *et al.* 2011). Within coastal western Alaska, populations from Norton Sound and the lower Yukon River displayed the highest levels of differentiation while populations from the Bristol Bay and Kuskokwim River regions appeared to be closely related. It is likely that the observed structure is the result of genetic drift in the lower Yukon River and Norton Sound facilitated by relatively small census sizes. Specifically, populations in the lower Yukon River and Norton Sound regions generally contain less than 2 000 spawners whereas many populations in the Bristol Bay and Kuskokwim River regions contain greater than 10 000 spawners (Molyneaux & Dubois 1999; Baker *et al.* 2006; Banducci *et al.* 2007; Heard *et al.* 2007; Howard *et al.* 2009). Different levels of effective migration within regions may also influence this pattern.

Comparison of panels for GSI

Previous studies demonstrate that the level of polymorphism (H_O) and differentiation (F_{ST}) of SNPs is positively correlated with their value for GSI (Ackerman *et al.* 2011; Bradbury *et al.* 2011; Storer *et al.* 2012). We observed a significantly higher average H_O and F_{ST} for the RAD96 panel compared to the CTC191 panel, and a higher proportion of SNPs from the RAD96 were chosen for the final F_{ST}96 panel. These results indicate that, on average, the SNPs in the RAD96 panel are likely to be more useful for GSI in populations from western Alaska than the SNPs in the CTC191 panel.

Assignment accuracies for all populations with both the full dataset and the F_{ST}96 panel were close to or above the 90% threshold necessary for management applications (Seeb *et al.* 2000). Assignment rates were lower for the CTC191 and RAD96 implying that GSI with our reporting groups would be less powerful with only one of these panels. Although both the CTC191 and RAD96 panels displayed similar assignment accuracies overall, there were major differences between the two panels for specific populations in the Norton Sound and lower Yukon River regions (e.g. population 1, 3, 6). These differences demonstrate the importance of obtaining a representative set of ascertainment populations when attempting to create a SNP panel for GSI.

SNP discovery and evaluation conducted in this study has increased the number of feasible reporting groups for GSI in western Alaska from one to three, but accuracy could be further improved by sampling additional populations from the lower Yukon River and Norton Sound regions. It is especially important to sample throughout Norton Sound because these populations were each genetically distinct from each other and all others in the study. Norton Sound is composed of many small, unconnected rivers with census sizes that are often under 1 000 (Banducci *et al.* 2007). These populations are likely able to quickly diverge from each other due to greater genetic drift in small populations and/or regional landscape features restricting gene flow. Dense sampling is therefore necessary to accurately characterize genetic variation in this region. Any additional populations could also be used as a holdout set to assess the assignment accuracy of our panels as suggested by Anderson (2010).

Ascertainment bias

Both the CTC191 and RAD96 panels exhibited similar patterns of population structure but also displayed evidence of ascertainment bias. Ascertainment bias occurs when genetic markers are chosen such that they are unrepresentative of genetic variation in all populations or

regions of interest (Smith *et al.* 2007). Ascertainment bias can distort estimates of population structure but can also increase assignment power in the region of interest (Bradbury *et al.* 2011). Two major sources of ascertainment bias were present in our data (1) regional ascertainment bias in the CTC191 and RAD96 panels and (2) population-specific bias in the populations that were RAD sequenced. The regional ascertainment bias in the CTC191 and RAD96 panels occurred because these two panels were largely created for regional applications (CTC191: south of Alaska; RAD96: western Alaska), a common occurrence with SNP panels developed for salmonid management (e.g., Seeb *et al.* 2011c). In this case, the regional bias in the RAD96 is helpful because it likely increases our power to differentiate populations in western Alaska. However, this bias may also decrease the power of these SNPs to differentiate populations outside the region of interest, possibly reducing the utility of the RAD96 across the species range (Smith *et al.* 2007). Assignment accuracies from the middle and upper Yukon River populations suggest that the SNPs developed in this study should be useful outside of the ascertainment area but further testing is needed to fully validate this assumption.

The second type of ascertainment bias present in our data was population-specific bias in the populations that were RAD sequenced. This bias likely occurred because dozens of SNPs showing high differentiation were chosen from thousands, causing the populations that were RAD sequenced to appear more differentiated than expected. This type of bias was especially apparent in the Anvik River (population 6) which clustered tightly with the two other lower Yukon River populations using the CTC191 but was highly diverged with the RAD96.

Population-specific ascertainment bias could lead to upwardly biased estimates of assignment accuracy and could distort phylogenetic relationships among populations. To reduce unwanted population-specific bias, we suggest that future studies with similar objectives sequence at least two ascertainment populations from each drainage/region. Hierarchical *F*-statistics could then be used to discover SNPs that are similar within but divergent among regions. For example SNPs with high values of F_{CT} (variation among reporting groups) and small values of F_{SC} (variation among populations within reporting groups) could be chosen.

Use of adaptively important markers for GSI

The accuracy of GSI in poorly differentiated populations can often be improved by including adaptively important markers that are undergoing divergent natural selection (Nielsen *et al.* 2012). For example, Ackerman et al. (2011) found that the addition of adaptively

important markers to a panel of neutral markers significantly improved assignment accuracy, and Russello et al. (2012) showed that assignment accuracies were much higher with a panel of adaptively important markers compared to a panel of neutral markers. Multiple studies using RAD sequencing have found signatures of natural selection (e.g., Hohenlohe *et al.* 2010; Gagnaire *et al.* 2013b), but strong signatures of selection were not apparent in our data. Specifically, patterns of population structure were similar with the RAD96 and the primarily neutral CTC191, and we observed relatively small F_{ST} values across most loci indicating that the majority of loci in our dataset were probably neutral. It is interesting to note that one locus from the CTC191, *Ots_MHC2*, had an overall F_{ST} of 0.431 in Chinook salmon from the Copper River and was found to be under strong divergent selection in this environment (Seeb *et al.* 2009b; Ackerman *et al.* 2013). *Ots_MHC2* also had one of the highest overall F_{ST} values in our study indicating that it may be adaptively important in western Alaska. Future studies attempting to improve GSI in our study region would likely benefit from the inclusion of additional adaptively important markers such as *Ots_MHC2*. For example, adaptively important markers might be useful for differentiating populations from the Kuskokwim River and Bristol Bay regions, something that was not possible with our current set of SNPs.

Management applications

The precipitous decline of Chinook salmon in western Alaska has prompted multiple fisheries closures, causing extensive economic hardship and threatening subsistence catches for natives of the western Alaska region (ADF&G 2013). Increased resolution for GSI facilitated by our study has the potential to significantly improve fisheries management in this region. Specifically, GSI can be used to monitor the contribution of different stocks in mixed-stock fisheries, informing fisheries management and preventing unnecessary fishery closures (Shaklee *et al.* 1999; Smith *et al.* 2005c; Dann *et al.* 2013). Additionally, SNPs developed in this study can be used to improve resolution in studies of migration and distribution patterns of Chinook salmon on the high seas (c.f., Tucker *et al.* 2009; Guthrie *et al.* 2013; Larson *et al.* 2013). The ability to measure stock-specific abundance on the high seas can provide important information for stock assessment models that is currently unavailable.

Conclusions

We increased the number of feasible reporting groups for GSI in coastal western Alaska from one to three using directed SNP discovery. The SNPs we developed from RAD data

displayed higher levels of polymorphism and differentiation compared to many previously developed SNPs and were more useful for GSI. RAD sequence data therefore provided an excellent tool for discovering high-resolution SNPs which can differentiate closely related populations. The increased resolution for GSI in coastal western Alaska facilitated by this study will facilitate research into migration patterns and vulnerability to fisheries of Chinook salmon in this region, aiding in the conservation of an extremely important economic and cultural resource.

Tables

Table 3.1. Collection location, sampling region, reporting group and sample size for each population in the study. Pop no. corresponds to the numbers in Fig. 3.1 and Fig. 3.4. Ascertainment populations that were RAD sequenced are in bold, and sample sizes for RAD sequencing are given in parentheses. RAD data were obtained from Larson *et al.* (2014c). The reporting group is the group that was used for assignment tests. The Bristol Bay and Kuskokwim River reporting group is abbreviated Bristol-Kusk.

Pop. no.	Location	Region	Reporting Group	Sampling year	Sample Size
1	Pilgrim River	Norton Sound	Norton Sound	2005, 2006	71
2	**Tubutulik River**	Norton Sound	Norton Sound	2009	85 (**56**)
3	North River	Norton Sound	Norton Sound	2010	60
4	Golsovia River	Norton Sound	Norton Sound	2006	59
5	Andreafsky River	Lower Yukon	Lower Yukon	2003	90
6	**Anvik River**	Lower Yukon	Lower Yukon	2007	52 (**51**)
7	Gisasa River	Lower Yukon	Lower Yukon	2001	81
8	Goodnews River	Kuskokwim Bay	Bristol-Kusk	2006	94
9	Arolik River	Kuskokwim Bay	Bristol-Kusk	2005	52
10	Kanektok River	Kuskokwim Bay	Bristol-Kusk	2005	93
11	Eek River	Kuskokwim-Mouth	Bristol-Kusk	2005	76
12	Kisaralik River	Kuskokwim-Lower	Bristol-Kusk	2005	94
13	Salmon River	Kuskokwim-Middle	Bristol-Kusk	2006	94
14	George River	Kuskokwim-Middle	Bristol-Kusk	2005	95
15	Kogrukluk River	Kuskokwim-Middle	Bristol-Kusk	2005	49
16	**Kogrukluk River**	Kuskokwim-Middle	Bristol-Kusk	2007	94 (**57**)
17	Necons River	Kuskokwim-Middle	Bristol-Kusk	2007	94
18	Gagaryah River	Kuskokwim-Middle	Bristol-Kusk	2006	94
19	Togiak River	West Bristol Bay	Bristol-Kusk	2009	94
20	Iowithla River	West Bristol Bay	Bristol-Kusk	2010	65
21	Stuyahok River	West Bristol Bay	Bristol-Kusk	2009	93
22	**Koktuli River**	West Bristol Bay	Bristol-Kusk	2010	94 (**56**)
23	Klutuspak Creek	West Bristol Bay	Bristol-Kusk	2009	94
24	Big Creek	East Bristol Bay	Bristol-Kusk	2004	65
25	Henshaw Creek	Middle Yukon	Middle Yukon	2001	88
26	Kantishna River	Middle Yukon	Middle Yukon	2005	94
27	Salcha River	Middle Yukon	Middle Yukon	2005	90
28	**Big Salmon River**	Upper Yukon	Upper Yukon	2007	71 (**47**)
Total					2 275 (**267**)

Table 3.2. Pairwise F_{ST} values for the five ascertainment populations calculated with 11,850 RAD-derived SNPs (overall F_{ST} = 0.041).

	Tubutulik River	Anvik River	Kogrukluk River	Koktuli River
Anvik River	0.030			
Kogrukluk River	0.026	0.005		
Koktuli River	0.028	0.006	0.002	
Big Salmon River	0.097	0.077	0.075	0.077

Table 3.3. Number of SNPs at each stage of SNP discovery. Validated 5' nuclease assays are those that successfully amplified and produced clean scatter plots.

Dataset	**Number of putative SNPs**
Unfiltered RAD	26,567
Filtered RAD	11,850
5' nuclease assays attempted	150
5' nuclease assays designed	128
5' nuclease assays validated	101
Top 96 assays	96

Table 3.4. Number of discrepancies between 5' nuclease and RAD genotypes across 254 individuals that were genotyped for both chemistries. The table is based on a bi-allelic locus with allele one designated by A and allele two designated by B.

5' nuclease genotype	**RAD genotype**	**Number**	**Proportion**
Concordance		23 955	0.990
Discrepancies			
AA	BB	0	0.000
AA or BB	AB	62	0.003
AB	AA or BB	182	0.007
Total discrepancies		244	0.010

Figures

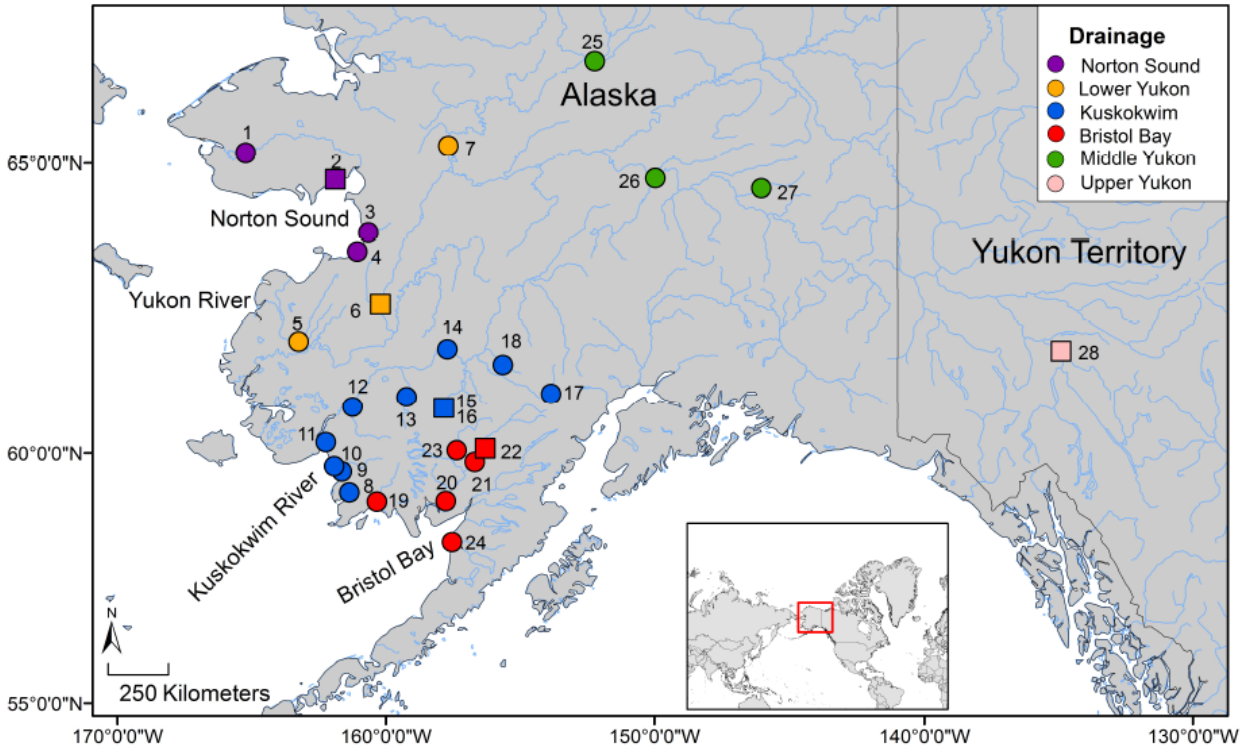

Fig. 3.1. Sampling locations for the 28 populations of Chinook salmon. Ascertainment populations that were RAD sequenced are denoted by squares. Table 3.1 provides additional details about each sampling site.

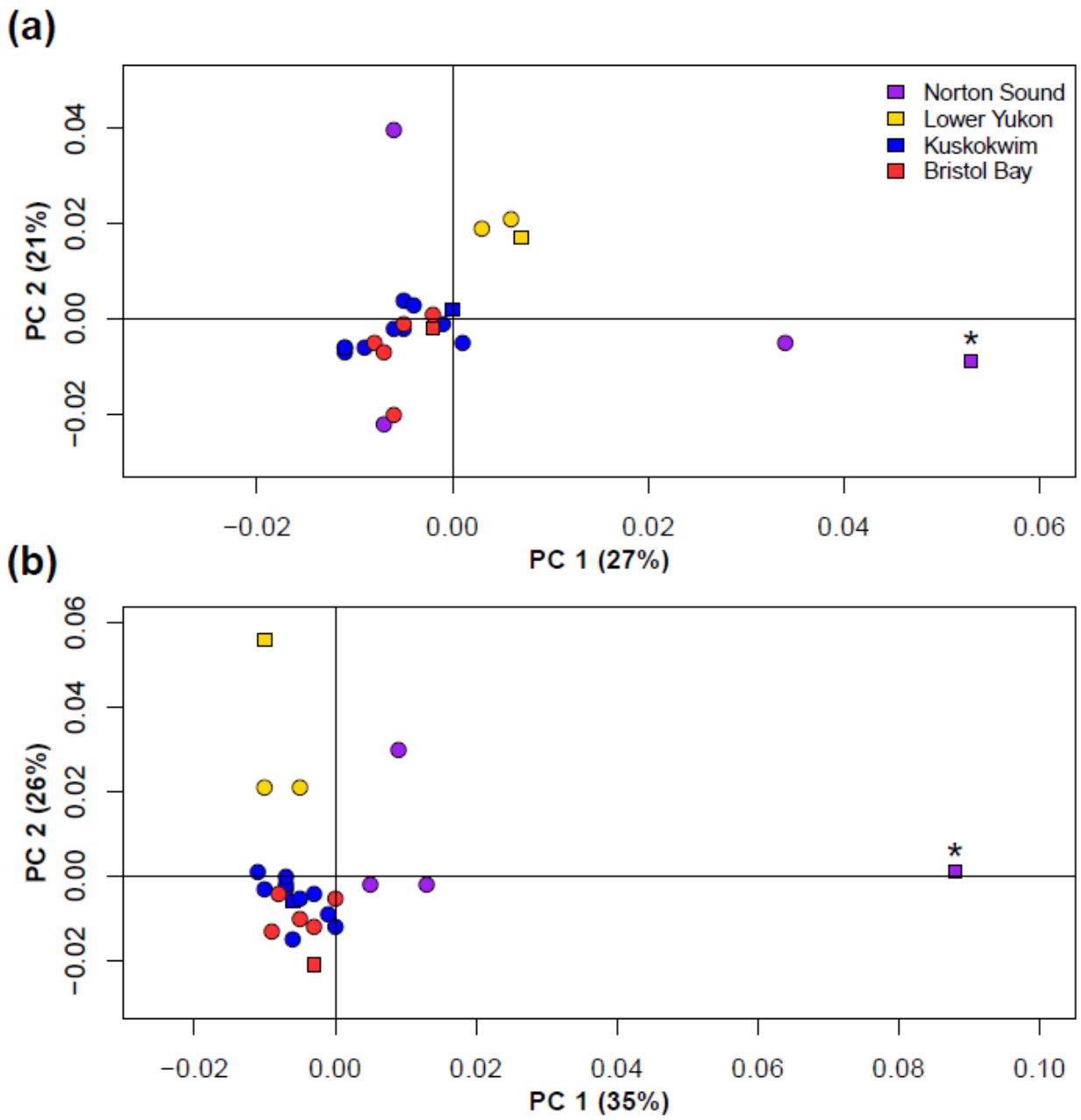

Fig. 3.2. Principal coordinate analysis (PCoA) of 24 populations from coastal western Alaska with: (a) CTC191, and (b) RAD96. Only SNPs that were in linkage and Hardy-Weinberg equilibrium were used in this analysis. The PCoA is based on pairwise-F_{ST} values. Squares are ascertainment populations that were RAD sequenced. Population 2 (Tubutulik River) is labeled with an "*" because it was a genetic outlier and was removed from some analyses (see text).

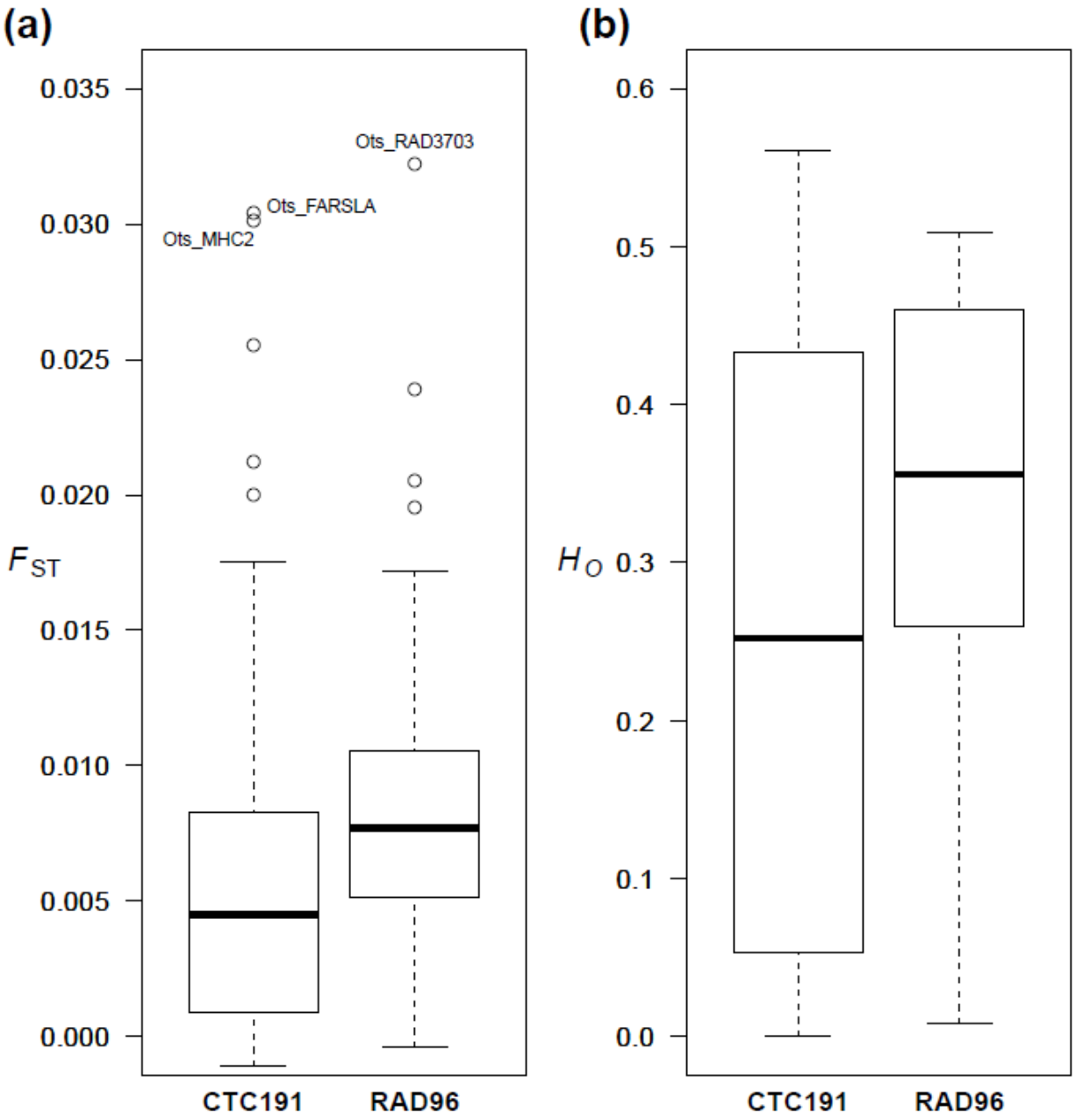

Fig. 3.3. Box and whisker plots of locus-specific overall (a) F_{ST} and (b) H_O for two SNP datasets. Datasets are CTC191 (average H_O = 0.24, average F_{ST} = 0.006) and RAD96 (average H_O = 0.34, average F_{ST} = 0.008). Only SNPs that were in linkage and Hardy-Weinberg equilibrium were used in this analysis. A Student's t-test indicated that the two datasets have significantly different distributions of H_O and F_{ST} ($P < 0.0001$). Loci with F_{ST} values above 0.03 are labeled in plot (a). Each dataset includes populations 1-24 (excluding population 2, see text).

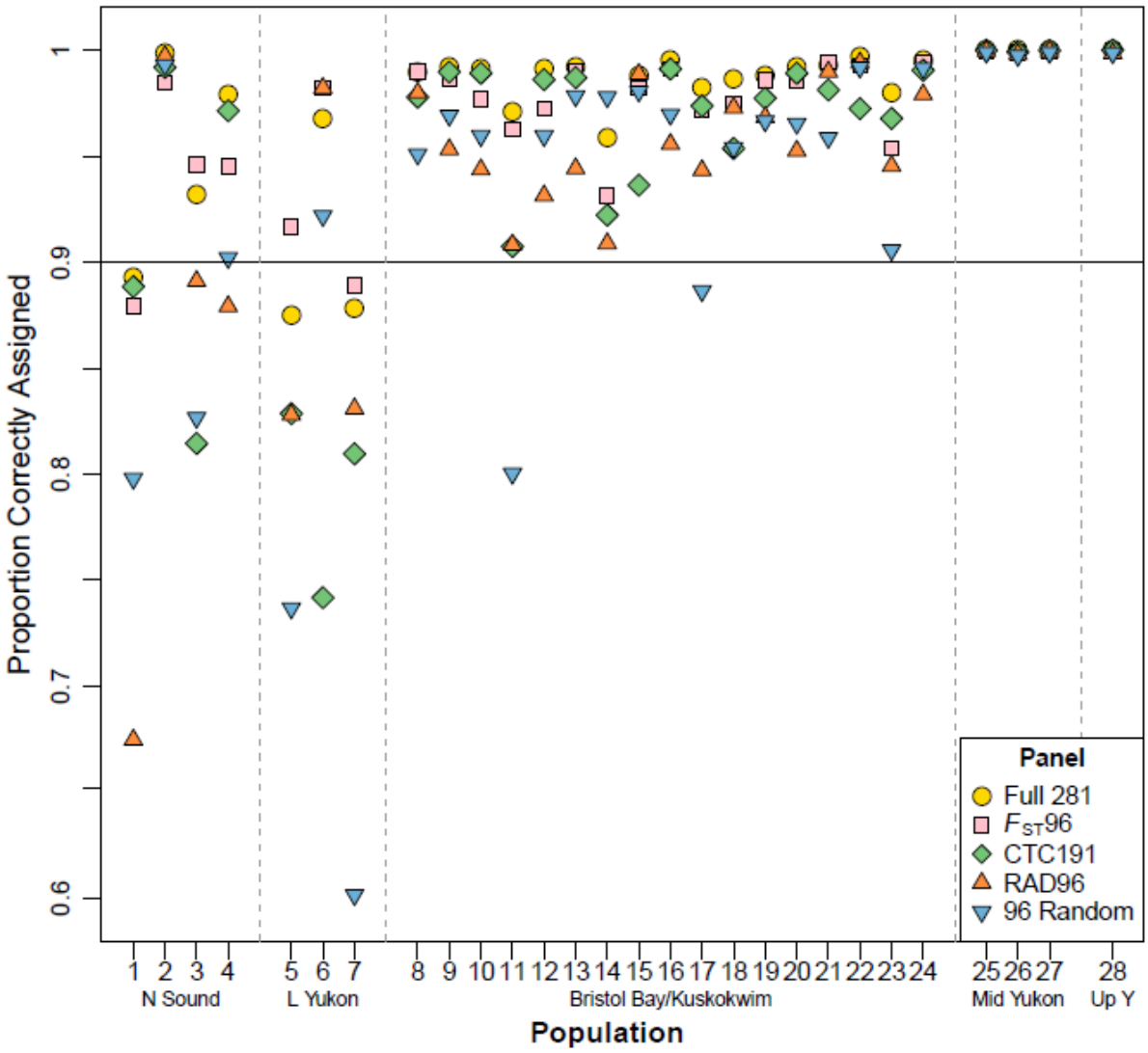

Fig. 3.4. Assignment probabilities to reporting group for the full dataset of 281 SNPs (Full 281), the 96 SNPs with the highest overall F_{ST} (F_{ST}96), the CTC191, the RAD96, and 96 randomly chosen SNPs (96 random). Only SNPs that were in linkage and Hardy-Weinberg equilibrium were used in this analysis. Population numbers correspond to those in Table 3.1. Reporting groups (X-axis) are separated by gray dashed lines. Abbreviations are Norton Sound (N Sound), lower Yukon River (L Yukon), Bristol Bay and Kuskokwim River (Bristol Bay/Kuskokwim), middle Yukon River (Mid Yukon), and upper Yukon River (Up Y). Bristol Bay and Kuskokwim River populations were combined into a single reporting group for this analysis (see text). The line at 0.9 represents a common value used to consider an assignment robust for management applications (Seeb *et al.* 2000). Confidence intervals for each assignment probability are reported in Table S6.

Chapter 4

Signals of heterogeneous selection at an MHC locus in geographically proximate ecotypes of sockeye salmon[4]

Abstract

The genes of the major histocompatibility complex (MHC) are an important component of the vertebrate immune system and can provide insights into the role of pathogen-mediated selection in wild populations. Here we examined variation at the MHC class II peptide binding region in 27 populations of sockeye salmon (*Oncorhynchus nerka*), distributed among three distinct spawning ecotypes, from a complex of interconnected rivers and lakes in southwestern Alaska. We also obtained genotypes from 90 putatively neutral SNPs for each population to compare the relative roles of demography and selection in shaping the observed MHC variation. We found that MHC divergence was generally partitioned by spawning ecotype (lake beaches, rivers, and streams) and was 30 times greater than variation at neutral markers. Additionally, we observed substantial differences in modes of selection and diversity among ecotypes, with beach populations displaying higher levels of directional selection and lower MHC diversity than the other two ecotypes. Finally, the level of MHC differentiation in our study system was comparable to that observed over much larger geographic ranges, suggesting that MHC variation does not necessarily increase with increasing spatial scale and may instead be driven by fine-scale differences in pathogen communities or pathogen virulence. The low levels of neutral structure and spatial proximity of populations in our study system indicates that MHC differentiation can be maintained through strong selective pressure even when ample opportunities for gene flow exist.

[4] Full citation: Larson, W.A., J.E. Seeb, T.H. Dann, D.E. Schindler, and L.W. Seeb. 2014. Signals of heterogeneous selection at an MHC locus in geographically proximate ecotypes of sockeye salmon. *Molecular Ecology*, **23**(22) 5448-5461. Supplementary materials available from the online version of this manuscript (http://onlinelibrary.wiley.com/doi/10.1111/mec.12949/abstract).

Introduction

The genes of the major histocompatibility complex (MHC) provide insights into the role of pathogen-mediated selection in wild populations. These genes encode proteins that bind foreign pathogens and present them to T cells (Potts & Wakeland 1990; Matsumura *et al.* 1992; Ekblom *et al.* 2007). Different variants of MHC genes confer resistance to different suites of pathogens (e.g., Hill *et al.* 1991; Langefors *et al.* 2001), making MHC genes an important component of the host-pathogen evolutionary arms race. MHC genes are some of the most polymorphic in vertebrates (Piertney & Oliver 2006), and the most diverse region of these genes is the peptide binding region (PBR) which has been characterized in a variety of taxa including mammals (Knapp *et al.* 1998), birds (Miller & Lambert 2004), and reptiles (Miller *et al.* 2010). It is widely hypothesized that MHC diversity is maintained through pathogen-mediated balancing selection (e.g., Landry & Bernatchez 2001; Niskanen *et al.* 2013); however, support for this hypothesis is not universal (reviewed in Bernatchez & Landry 2003; Piertney & Oliver 2006). For example, multiple studies have found that MHC differentiation among populations is larger than would be expected under a model of neutrality or balancing selection (e.g., Ekblom *et al.* 2007; Gomez-Uchida *et al.* 2011; Ackerman *et al.* 2013), while others have found that patterns of MHC differentiation are consistent with neutrality (e.g., Seddon & Baverstock 1999; Campos *et al.* 2006). Together, these results indicate that selection at the MHC is extremely variable depending on environment, and that neutral processes such as genetic drift can influence MHC variation. It is also important to note that, while many studies provide evidence to suggest that the MHC is under strong selection, very few studies have been able to demonstrate a direct association between pathogen communities and MHC diversity (but see Dionne *et al.* 2007; Evans & Neff 2009).

Salmon represent an excellent model for studying MHC variation because of their distinct life history. Salmon return to their natal streams with high fidelity, facilitating local adaptation to a variety of environmental characteristics including pathogen communities (Stewart *et al.* 2003; Zueva *et al.* 2014). Additionally, salmon express a single copy of the MHC class I and MHC class II, simplifying analysis compared to other species that have many copies (Lukacs *et al.* 2010). MHC variation has been shown to directly influence pathogen resistance in salmon (e.g., Consuegra & Garcia de Leaniz 2008; Evans & Neff 2009), providing substantial evidence that these genes are important for local adaptation and overall fitness.

Past studies of MHC variation in wild populations of salmon have demonstrated different types of selection depending on species, MHC class, and environment (reviewed in Bernatchez & Landry 2003). For example, McClelland *et al.* (2013) documented both balancing and directional selection at the MHC in 70 populations of sockeye salmon (*Oncorhynchus nerka*) sampled across the species range, and Evans et al. (2010) found that the MHC class I and class II genes were undergoing different types of selection in the same populations of Chinook salmon (*Oncorhynchus tshawytscha*). These studies emphasize MHC variation at broad spatial scales and among classes; however, the influence of fine-scale environmental heterogeneity on MHC variation has not been thoroughly investigated.

Here we examined variation at the MHC in 27 populations of sockeye salmon from the Wood River basin in southwestern Alaska. The Wood River basin is a series of five interconnected lakes that encompasses a drainage area of 3 590 km^2 and is relatively free from anthropogenic impacts. Sockeye salmon from this system comprise a large component of one of the most valuable fisheries in the United States (Schindler *et al.* 2010). Extensive life history and phenotypic diversity exists in these populations, including the presence of three distinct ecotypes associated with the habitat used for spawning: lake beaches, rivers between lakes, and streams (Hilborn *et al.* 2003).

Spawning habitats for these three ecotypes differ in a number of geomorphological characteristics that may influence pathogen communities. Beach spawning environments are relatively homogenous, with generally stable temperatures and fairly low spawning densities. On the other hand, stream and river habitats are more heterogeneous, can experience larger temperature fluctuations (Lisi *et al.* 2013) and often have higher spawning densities (ASP, unpublished data; cf., Braun & Reynolds 2011). Additionally, juveniles from river and stream populations migrate to lakes to rear for 1-3 years whereas this migration is not obligate for juveniles from beach populations (McGlauflin *et al.* 2011). Differences also exist between river and stream environments; these include gravel size, water depth, water velocity, temperature, and predation (Quinn *et al.* 2001; Pess *et al.* 2013).

The three ecotypes of sockeye salmon in this system exhibit differentiation in a variety of traits, such as body size and shape (Quinn *et al.* 2001), egg morphology (Quinn *et al.* 1995), and spawn timing (Schindler *et al.* 2010). For example, sockeye salmon that spawn in rivers and on beaches are much larger than individuals that spawn in streams. Additionally, sockeye salmon

from beach and river ecotypes possess larger eggs and spawn two to six weeks later than individuals from the stream ecotype. Despite this extensive phenotypic and life history differentiation, there is minimal evidence of genetic structure among populations from the three ecotypes at neutral markers (McGlauflin *et al.* 2011). However, McGlauflin *et al.* (2011) did observe extensive differentiation at two single-nucleotide polymorphisms (SNPs) in the MHC class II region, suggesting that selection may be influencing the observed patterns of variation.

In this study we obtained sequence data from the MHC class II region for 27 populations of sockeye salmon to investigate the hypothesis that the MHC is undergoing strong and variable selection across the Wood River basin. We also obtained data from 90 putatively neutral SNPs to compare patterns of differentiation between the MHC and neutral loci and investigate the relative role of selection and neutral processes in shaping the genetic variation that we observed. Our study is one of the first to investigate fine-scale differentiation at the MHC and provides evidence that pathogen-mediated selection can promote MHC differentiation at extremely small spatial scales (as little as 1 km).

Materials and methods

Tissue sampling and DNA extraction

Axillary processes were collected from sockeye salmon from 27 spawning populations throughout the Wood River basin during 2001-2013 (Table 4.1, Fig. 4.1). Removing the axillary process also represents a visual mark that ensured the same individual was not resampled. Populations were classified into three ecotypes based on spawning environment. Beach spawners were sampled in 1-2 m of water along the shores of lakes, river spawners were sampled from large (> 50 m wide) tributaries that drain major lakes, and stream spawners were sampled from small (2-16 m wide) tributaries that drain adjacent hillsides, tundra or small lakes and empty into major lakes (Quinn *et al.* 2001; Lisi *et al.* 2013). Populations were also grouped into one of five nursery lakes (Kulik, Beverley, Nerka, Lynx, and Aleknagik) based on where the majority of the population's juveniles likely rear. Collections from multiple years were pooled to increase sample size as suggested by Waples (1990b). Genomic DNA was extracted from fin clips with a DNeasy 96 Tissue Kit (Qiagen, Valencia, California).

Genotyping neutral SNPs

We utilized genotypes from the 96-SNP panel described in Elfstrom *et al.* (2006) and Storer *et al.* (2012) to construct a neutral dataset for all 27 populations. Genotypes were

available for 20 of the 27 populations as part of the dataset in Dann *et al.* (2012). Data for the remaining seven populations were obtained using Biomark 96.96 Dynamic arrays (Fluidigm, South San Francisco, California) following the methods of (Smith *et al.* 2011). We removed six SNPs from the 96 panel that were either in linkage disequilibrium (LD) with another locus according to Dann et al. (2012) (*One_GPDH-187, One_Tf_ex11-750*), produced scatter plots that displayed inconsistent separation among genotypes (*One_SUMO1-6, One_U1016-115*), or were found within the MHC (*One_MHC_190, One_MHC_251*). Three mtDNA SNPs were combined and scored as a single locus following the methods of Habicht *et al.* (2010). Individuals with greater than four missing genotypes were removed from further analysis. Finally, we re-genotyped four out of every 95 (4.1%) individuals to quantify genotyping discrepancy rate.

To validate the neutrality of our putatively neutral dataset, we conducted a test to detect loci under selection from F-statistics across all populations using ARLEQUIN 3.5 (Excoffier & Lischer 2010) with the following parameters: finite island model, 20 000 simulations, 100 demes. No loci were candidates for directional selection at the 1% significance level (data not shown), and we proceeded with these 90 SNPs as the neutral dataset.

Genotyping the MHC locus

MHC sequence data was obtained by PCR amplifying and Sanger sequencing a 373 base pair fragment of the MHC class II *β1* locus described in Miller and Withler (1996). DNA was amplified with the forward primer from Miller et al. (2001, 5'-CCGATACTCCTCAAAGGACCTGCA-3') and a reverse primer located 110 nucleotides into intron 2 (5'-TTAATCCCTGAATCTCCACCATCA-3'). PCR was conducted in a 20 µL volume containing approximately 10 ng of DNA, 1X PCR Gold Buffer (Life Technologies, Carlsbad, California), 1 mM MgCl2, 200 mM dNTPs, 0.5 units of AmpliTaq Gold DNA polymerase (Life Technologies), and 0.2 µM of each primer. Thermal cycling was performed as follows: 95 °C hold for 10 min, followed by 40 cycles of 95 °C for 10 s, 56 °C for 30 s, 72 °C 60 s, and a final extension of 72 °C for 7 min. PCR products were sequenced in both directions on a 3730 DNA Analyzer (Life Technologies).

Sequence chromatograms were aligned, and polymorphisms were scored and visually confirmed with SEQUENCHER 4.10 (Gene Codes Corporation, Ann Arbor, Michigan). We chose a consensus sequence that facilitated the inclusion of as many polymorphic sites as possible without sacrificing sample size. The consensus sequence began at codon 34 of the *β1*

exon and ended at the last codon (codon 94). This 180 bp sequence contained all but two polymorphic positions previously discovered in the *β1* exon in sockeye salmon (Miller & Withler 1996) and all polymorphisms in the putative PBR (Miller *et al.* 2001). Polymorphisms were scored using the IUPAC nucleotide ambiguity codes.

We used the PHASE program (Stephens *et al.* 2001) implemented in DnaSP 5.10.01 (Librado & Rozas 2009) to reconstruct the most likely MHC haplotypes for each individual from the sequence data. PHASE implements a coalescent-based Bayesian method to infer haplotypes from sequence data and has been shown to be robust to departures from Hardy-Weinberg equilibrium (Stephens *et al.* 2001). The input parameters for PHASE were a 100 iteration burn-in followed by 1 000 iterations. Individuals with less than a 95% probability of correct haplotype reconstruction were removed from further analysis. Haplotypes throughout the study were labeled with an arbitrary number corresponding to the order in which they were discovered (e.g. *H2* was the second haplotype discovered). Each haplotype was aligned to the NCBI nucleotide database using BLASTN to facilitate comparisons to previous studies.

Summary statistics

Exact tests for deviations from Hardy-Weinberg and linkage equilibrium were conducted for each population and locus in the neutral dataset using GENEPOP 4 (Rousset 2008). Hardy-Weinberg tests were also conducted for each population in the MHC dataset. The initial significance level for these tests was 0.05, and we applied a sequential Bonferroni correction (Rice 1989) to correct for multiple tests. Summary statistics, including allele frequencies, allelic richness (A_R) and observed and expected heterozygosities (H_O, H_E), were calculated for each population in both datasets in FSTAT 2.9.3 (Goudet 1995) and GenAlEx 6.5 (Peakall & Smouse 2012).

Population structure

Genetic relationships among populations were visualized for both datasets separately with neighbor-joining trees based on Nei's D_A distance (Nei *et al.* 1983) constructed in POPTREE2 (Takezaki *et al.* 2010). We also calculated overall and pairwise F_{ST} values (Weir & Cockerham 1984) in GENEPOP and conducted exact tests of genetic differentiation (Raymond & Rousset 1995; Goudet *et al.* 1996) for each dataset in ARLEQUIN 3.5 (Excoffier & Lischer 2010) with the default parameters and a significance level of 0.01. Finally, we split samples that

were taken from the same population over multiple years and conducted exact tests of genetic differentiation among years to ensure that genetic structure was temporally stable.

Selection on the MHC

We quantified the number of synonymous (D_S) and non-synonymous (D_N) substitutions in the MHC dataset in MEGA 6 (Tamura *et al.* 2011) to test the hypothesis that $D_N > D_S$, indicating historical balancing selection. We also conducted an Ewens-Watterson homozygosity test (Ewens 1972; Watterson 1978) in ARLEQUIN to quantify recent selection within each population. This test compares observed haplotype frequencies (designated as observed *F*) to simulated haplotype frequencies for a gene in migration drift equilibrium (designated as expected *F*) to determine if the population is undergoing balancing selection (observed *F* < expected *F*), directional selection (observed *F* > expected *F*), or no selection (observed *F* ~ expected *F*). Two methods are implemented in ARLEQUIN to estimate the significance of the Ewens-Watterson distribution, the original Ewens-Watterson test (Ewens 1972; Watterson 1978) and a slightly modified version developed by Slatkin (1996). P-values < 0.05 for either method indicate strong evidence for balancing selection and P-values > 0.95 indicate strong evidence for directional selection. Marginally significant P-values (P < 0.1, P > 0.9) were also considered as evidence for selection as this test can fail to return significant results in cases of weak to moderate selection (Ewens 1972).

Lake and ecotype effects

We used three analyses to test the influence of nursery lake and ecotype on patterns of genetic differentiation, diversity, and selection in our dataset.

First, we conducted an analysis of molecular variance (AMOVA) in ARLEQUIN to determine the amount of genetic variation partitioned within and among different groupings of populations. This analysis was conducted for the neutral and MHC datasets with two different population groupings: (1) all populations grouped by ecotype, (2) all populations grouped by nursery lake.

We then constructed box and whisker plots of pairwise F_{ST} to further investigate the effect of nursery lake and ecotype on genetic differentiation. For this analysis, estimates of pairwise F_{ST} within ecotypes were grouped into among lake and within lake comparisons. We also used box and whisker plots to quantify differences in genetic diversity (H_O) among

ecotypes. Both box and whisker analyses were conducted separately for the neutral and MHC datasets.

Finally, we plotted ratios of observed *F* to expected *F* values from the Ewens-Watterson test to visualize trends in selection among lakes and ecotypes.

Results

Genotyping neutral SNPs and MHC

Genotyping success rate for the 90 neutral SNPs was 96% (2515/2691), and the combined genotyping discrepancy rate from this study and Dann *et al.* (2012) was 0.08%.

Sequencing the MHC revealed 16 polymorphic sites, and all sites produced sequence that could be scored unambiguously. Phasing sequences into haplotypes based on these polymorphic sites was successful for 1236 out of 1248 sequences (99%). In total, 27 unique haplotypes were found, and haplotype occurrences ranged from 580 occurrences for haplotype *H7* to one occurrence for haplotypes *H19*, *H25*, *H26*, and *H27*. Alignments to the NCBI nucleotide database were successful for 13 of the 27 haplotypes (see Table S1 for alignment results and sequences for each haplotype).

Summary statistics and population structure

We calculated summary statistics to test for departures from linkage and Hardy-Weinberg equilibrium and to quantify genetic diversity and variation within and among populations. No locus pairs showed significant deviations from linkage equilibrium in the neutral dataset, and no loci or populations deviated significantly from Hardy-Weinberg equilibrium for the neutral or MHC datasets ($P < 0.05$, sequential Bonferroni correction for all tests). Estimates of genetic diversity within populations (A_R and H_O) were higher on average and more variable in the MHC dataset than the neutral dataset (Table 4.1). Overall genetic differentiation (F_{ST}) was approximately 30 times higher for the MHC dataset compared to the neutral dataset (neutral $F_{ST} = 0.01$, MHC $F_{ST} = 0.27$). Results from a hierarchical AMOVA also displayed much higher levels of variation among groups and among populations within groups for the MHC dataset as compared to the neutral dataset (MHC: 29%, neutral: 1%, Table 4.2).

Genetic structure within the neutral and MHC datasets was generally partitioned by spawning ecotype, and each ecotype contained one to two dominant MHC haplotypes (Fig. 4.1, 4.2). Evidence of structuring by lake was also present, especially in beach populations. For example, the Silverhorn Beach population in Lake Beverley and the North Shore Kulik Beach

population in Lake Kulik were each nearly fixed for alternate haplotypes (Fig. 4.1). All populations adhered to the general lake and ecotype pattern described above except for the Grant River population, a stream population that grouped closely with a neighboring beach population rather than other stream populations in the system.

Tests for genetic differentiation between populations were not significant for any population pair with the neutral dataset but were significant for over half of population pairs with the MHC dataset (Table S2, S3). No significant genetic differentiation was found between samples from the same population taken over multiple years in the neutral SNP or MHC datasets, a similar result to Gomez-Uchida *et al.* (2012).

Selection and diversity at the MHC

All polymorphisms in the MHC dataset were non-synonymous (also reported by Miller *et al.* 2001), providing substantial evidence for historical balancing selection. Despite this, results from Ewens-Watterson homozygosity tests for each population suggested that recent directional selection was more prevalent than recent balancing selection in our study system. Evidence for directional selection at significant or marginally significant levels ($P > 0.9$) was present in 13 populations (48%) while evidence for balancing selection at these levels ($P < 0.1$) was present in only two populations (7%, Table 4.1).

Comparisons among ecotypes and lakes

Hierarchical AMOVAs with populations grouped by either nursery lake or ecotype (Table 4.2) revealed that 11.6% of the variation in the MHC dataset was partitioned among ecotypes and 15.6% was partitioned among lakes (69% of variation occurred within populations). In contrast, almost 99% of variation in the neutral dataset existed within populations.

Box and whisker plots of pairwise F_{ST} revealed high genetic differentiation among beach populations from different lakes for MHC dataset and moderate differentiation for the neutral dataset (Fig. 4.3). Beach populations within lakes, however, were not highly differentiated. A similar but less pronounced pattern of high genetic differentiation among lakes and low differentiation within lakes was also observed for stream populations in the MHC and neutral datasets (Fig. 4.3). This pattern was accentuated by the Grant River population, which was highly differentiated from all other stream populations at the MHC. River populations showed

similar levels of differentiation among and within lakes, but low sample size may have affected our power to differentiate these groupings.

Levels of H_O were similar among all three ecotypes in the neutral dataset but were generally lower for beach populations compared to river and stream populations in the MHC dataset (Fig. 4.4, Table 4.1). Additionally, beach and stream populations both displayed larger variation in H_O at the MHC than rivers (beaches range = 0.021 – 0.886, streams range = 0.280 – 0.911, rivers range = 0.663 – 0.902).

Results from Ewens-Watterson tests for each population indicated that most beach populations showed signals of directional selection, whereas the mode of selection in stream populations was variable and possibly associated with lake of origin (Fig. 4.5). Stream populations from lakes Beverley and Nerka (populations 5, 6, 8, 10, 12, 14, 15, 17, 18) generally displayed signals of balancing selection or neutrality, whereas stream populations from lakes Kulik and Aleknagik (populations 2, 21-24, 26) displayed signals of directional selection. River populations appeared to be evolving neutrally with one exception, the Agulowak (population 19).

Discussion

Population structure

Levels of MHC differentiation among populations inhabiting the Wood River basin were similar to those observed in previous studies of sockeye salmon spanning much larger geographic areas. For example, overall genetic differentiation at the MHC in our study, which spanned a river basin of about 3 590 km^2, was comparable to that observed in a study spanning nearly the entire range of sockeye salmon (F_{ST} = 0.27 in current study and 0.34 in McClelland *et al.* 2013) . Additionally, the amount of genetic variation partitioned among populations at the MHC was similar in our study and a study spanning the entire Fraser River that drains an area of 220 000 km^2 (29% of variation partitioned among populations in current study and 25% in Miller *et al.* (2001)) . These findings suggest that MHC differentiation does not necessarily increase with spatial scale. Instead, it is likely that historical balancing selection has maintained similar MHC alleles across large geographic scales, but the frequency of these alleles is highly variable on small scales due to differences in pathogen communities or their virulence among proximate habitats.

Relatively few studies have examined MHC variation at both large and small spatial scales in non-salmonids. One exception is a study by Alcaide *et al.* (2008), who examined

patterns of MHC variation in kestrel populations across a large geographic area and found high levels of divergence among proximate populations. However, similar MHC alleles were maintained over the entire study region suggesting that balancing selection likely occurred. These findings provide further evidence to support our hypothesis that pathogen-mediated selection causes high levels of MHC differentiation at small spatial scales while balancing selection maintains similar MHC alleles over large geographic areas. Nevertheless, future research is necessary to confirm that the patterns of MHC variation we observed are present in other organisms and study systems.

Genetic differentiation among populations within the Wood River basin was 30 times higher at the MHC compared to neutral markers and was generally partitioned by spawning ecotype. For example, beach populations from Lake Beverley were more similar to beach populations from Lake Nerka than they were to stream populations that spawn only a few km away. Similar, but less resolved patterns of population differentiation, were also observed by McGlauflin *et al.* (2011), who examined variation at 45 SNPs in this same lake system. Genetic differentiation of spawning ecotypes is likely driven by a pattern of isolation by adaptation (Nosil *et al.* 2008), where low reproductive success of migrants among ecotypes promotes differentiation among these ecotypes at both neutral loci and loci under selection (Lin *et al.* 2008a; Peterson *et al.* 2014). A similar pattern of differentiation between ecologically distinct populations at neutral markers and the MHC was observed in the great snipe (Ekblom *et al.* 2007), although this differentiation occurred over large geographic scales.

Despite the general pattern of genetic structuring by ecotype, genetic differentiation within ecotypes among lakes also existed. This pattern was especially prevalent in beach populations which displayed extremely high levels of MHC differentiation and moderate levels of neutral differentiation among lakes. Genetic differentiation among lake populations at neutral markers and the MHC has also been found in sockeye salmon from two other drainages in southwestern Alaska (Creelman *et al.* 2011; Gomez-Uchida *et al.* 2011). However, both of these studies found significant isolation by distance relationships which are not present in the Wood River basin (McGlauflin *et al.* 2011). These results provide further evidence that isolation by adaptation caused by phenotypic differences among ecotypes is a more important driver of genetic differentiation in the Wood River basin than isolation by distance.

Genetic differentiation within nursery lakes was relatively small for beach and stream populations at neutral markers and at the MHC, indicating that migration within ecotypes likely occurs. This pattern was also observed by Lin *et al.* (2008b), who suggested that gene flow among populations in three proximate streams in Lake Aleknagik was sufficient to homogenize genetic structure. Similar patterns of low MHC differentiation among spatially proximate populations have also been found for a variety of species including water voles (Oliver *et al.* 2009) and mice (Huang & Yu 2003). These results suggest that patterns of differentiation at the MHC are often influenced by metapopulations dynamics.

Although migration between ecotypes seems to be rare, migrants from river populations do appear to colonize beach habitats. Evidence for this is provided by the two small and ephemeral beach populations from Lake Aleknagik (populations 20, 25) that are genetically similar to proximate river populations and share similar spawn timing. Migration between ecotypes may also explain the genetic similarity between the North Shore Kulik Beach and the Grant River populations. The Grant River is an isolated stream that is surrounded by large beach populations. This stream is also relatively wide and deep, likely reducing selection against migrants from proximate beach populations. These factors indicate that the Grant River was either recently recolonized by beach spawners from Lake Kulik or receives enough gene flow from proximate beach populations to homogenize genetic structure between stream and beach habitats.

Patterns of population structure were generally similar between the MHC and neutral datasets but some inconsistencies did exist. For example, both populations from Lake Kulik were genetically similar to stream populations with neutral SNPs but were similar to beach populations at the MHC. Geomorphological features suggest that Lake Kulik was isolated from the other lakes in the system until recently (DES, personal observation). Our results indicate that, when this lake was reconnected with the rest of the system, it was colonized by stream spawners. Strong selection pressure then likely caused populations from Lake Kulik to become nearly fixed for a common beach allele at the MHC. This result demonstrates the utility of combining data from adaptive and neutral markers and indicates that selection may cause rapid divergence in MHC allele frequencies.

Diversity and selection at the MHC

The fact that all substitutions in the MHC were non-synonymous provided substantial evidence for historic balancing selection, a similar result to previous studies in salmon (Miller *et al.* 2001; Evans *et al.* 2010) and other taxa (Van der Walt *et al.* 2001; Jarvi *et al.* 2004). However, the highly elevated levels of genetic differentiation (F_{ST}) at the MHC compared to the neutral SNPs indicates that strong divergent selection among populations has recently occurred. It is important to note that highly variable loci such as the MHC locus in this study often display lower levels of genetic differentiation than SNPs (Hedrick 2005), therefore the 30 fold increase in differentiation between the neutral SNPs and the MHC provides strong evidence for divergent selection among populations.

Within populations, signals of recent selection were common and patterns of diversity were extremely variable. For example, we found that over half of the populations in our study showed signatures of either recent balancing or directional selection. The frequency of selection in our system was higher than both Miller *et al.* (2001) and McClelland *et al.* (2013) who found signatures of selection in less than half of their study populations with the same analyses and statistical thresholds that we employed. We also documented a large range of H_O at the MHC that was comparable to studies spanning much larger geographic areas (e.g., Miller *et al.* 2001; McClelland *et al.* 2013). The patterns of H_O at the MHC were not correlated with any patterns of H_O at neutral markers, indicating that variation in MHC H_O was unlikely to be caused by neutral processes alone.

In general, mode of selection and level of diversity at the MHC was partitioned by ecotype. Beach populations most commonly displayed low levels of diversity and frequent signatures of directional selection while river and stream populations displayed higher levels of diversity and variable selection that was dependent on lake of origin. In total, only two populations showed significant signals of balancing selection while 13 displayed signals of directional selection.

Directional selection at the MHC should be favored when one or a few alleles provide the highest resistance to the local pathogen community, whereas balancing selection should be favored when pathogen communities are diverse and require many alleles to maximize resistance (Bernatchez & Landry 2003). Signals of directional selection at the MHC have been found in a variety of study systems (e.g., Oliver *et al.* 2009; Fraser *et al.* 2010); however, these signals are generally rare compared to signals of balancing selection. The unusually high prevalence of

directional selection in our study system may be explained by high environmental heterogeneity and possibly low levels of pathogen diversity within environments. Nevertheless, we do see high levels of diversity and signals of balancing selection in some stream populations, suggesting that certain environments in our study system may contain diverse pathogen communities.

Our study is one of the first to demonstrate that the mode of selection acting on the MHC is highly variable across extremely small spatial scales. For example, we observed balancing and directional selection as well as neutral evolution in populations of a highly mobile species that were separated by as little as 1 km. These findings suggest that the long-standing hypothesis that selection at the MHC is most often driven by pathogen-mediated balancing selection may not be accurate for many study systems. Instead the mode of selection acting on the MHC cannot be generalized and is likely to be highly variable, even among proximate populations.

What is driving MHC differentiation in the Wood River basin?

Sockeye salmon inhabit multiple fresh and saltwater environments throughout their lifecycle, thereby complicating our understanding of the impacts of MHC variation on pathogen resistance. However, we observed strong differentiation among spawning populations suggesting that pathogen-mediated selection at the MHC is likely occurring in spawning or early rearing habitats. Past studies have shown that disease related mortality of adults on the spawning grounds can extirpate a large portion of the population (reviewed in Miller *et al.* 2014). Additionally, pathogens can cause significant mortality events while juveniles are rearing in their natal habitat (Quinn 2005).

Spawning habitats in our study system differ in a number of geomorphological characteristics that could influence pathogen communities and the MHC variation that we observed. For example, the apparent homogeneity and stability of beach spawning habitats could be an important factor underlying the low MHC diversity and prevalence of directional selection in these populations. Stream and river spawners, on the other hand, experience high levels of environmental heterogeneity and must migrate between environments, suggesting that maintenance of a larger number of MHC alleles would be advantageous to successfully respond to a variety of pathogens. It is also important to note that spawning densities in stream and river environments are generally higher than in beach habitats. High density is often correlated with increased prevalence of pathogens (reviewed in Reno 1998), therefore the spread of pathogens

may be faster and pathogen mediated selection may be stronger in river and stream environments.

Differences in pathogen virulence among spawning habitats may also contribute to the MHC variation that we observed. Pathogen virulence is influenced by a variety of environmental factors including water quality, nutrient availability, and temperature (Miller *et al.* 2014). Temperature, in particular, is one of the most important abiotic factors influencing fish and pathogen physiology (Fry 1971; Marcogliese 2001) and has been shown to significantly increase risk of infection and pre-spawn mortality in salmon (Martins *et al.* 2012). Spawning environments in the Wood River basin display extreme thermal heterogeneity and can vary by as much as 16°C (Armstrong *et al.* 2010). This thermal heterogeneity may create spatial variation in parasite virulence that could influence the magnitude of selection on the MHC (Dionne *et al.* 2007). For example, temperatures in Aleknagik streams are generally colder and more stable than streams from lakes Nerka and Beverley (Lisi *et al.* 2013), possibly explaining the differences in MHC diversity and mode of selection that we observed between these groups of streams.

Although variation in pathogen communities or spatial variation in their virulence has likely caused the patterns of MHC differentiation that we observed in this study, we did not provide any direct evidence to support this claim. Future studies could attempt to quantify pathogen communities across this system to confirm the relationships that we hypothesized. However, demonstrating a link between MHC diversity and pathogen communities is non-trivial and has only been accomplished in an exceedingly small number of study systems (e.g., Dionne *et al.* 2007; Evans & Neff 2009). These type of studies require strong collaborative efforts between pathologists and population geneticists but will likely prove extremely useful for understanding the mechanisms driving MHC variation in wild populations.

Conclusions and conservation implications

We demonstrated significant MHC variation among spatially proximate populations, even when little neutral structure existed. In general, MHC variation was partitioned by spawning ecotype, although significant variation among lakes did exist for populations from the same ecotype. We also demonstrated extensive variation in MHC diversity and mode of selection across populations. Overall genetic differentiation and the range of MHC diversity within our study system was similar to studies encompassing much larger geographic areas. This

indicates that MHC variation does not necessarily increase with increasing spatial scale and may instead be driven by fine-scale differences in pathogen communities or habitat-mediated expression of pathogen virulence.

Our study provides evidence that environmental heterogeneity can promote adaptive variation at small spatial scales. Adaptive variation has been shown to significantly increase ecosystem stability and resilience (Hilborn *et al.* 2003; Hutchinson 2008; Schindler *et al.* 2010). It is therefore important to maintain environmental heterogeneity in order to ensure species are resilient to fluctuating environmental conditions and changes in climate.

Tables

Table 4.1. Collection data and summary statistics for 27 populations of sockeye salmon. P-values are given for the Ewens-Watterson test for neutrality (EW P, Ewens 1972; Watterson 1978) and the Slatkin's exact P-test (Slat P, Slatkin 1996). P-values < 0.05 indicating balancing selection, and P-values > 0.95 indicating directional selection are in **bold**. Marginally significant values (P < 0.1 or > 0.9) are underlined.

				Neutral SNPs					MHC						
Pop #	Population	Lake	Ecotype	N	Years	A_R	H_O	H_E	N	Years	A_R	H_O	H_E	EW P	Slat P
1	North Shore Kulik	Kulik	Beach	90	2007	1.87	0.255	0.255	51	2007	5.58	0.157	0.167	**0.990**	**0.990**
2	Grant River	Kulik	Stream	81	2007	1.86	0.265	0.257	50	2007	5.52	0.280	0.298	0.941	**0.963**
3	Silverhorn Beach	Beverley	Beach	94	2007	1.88	0.262	0.260	47	2007	2.00	0.021	0.062	0.701	0.701
4	Hardluck Beach	Beverley	Beach	94	2007	1.88	0.260	0.260	45	2007	7.73	0.267	0.263	**0.997**	**0.996**
5	Uno Creek	Beverley	Stream	89	2009, 2013	1.85	0.262	0.260	47	2013	9.89	0.787	0.837	0.054	**0.033**
6	Moose Creek	Beverley	Stream	92	2009	1.85	0.248	0.252	45	2009	15.66	0.911	0.859	0.448	0.516
7	Agulupak River	Nerka	River	88	2001	1.87	0.256	0.257	47	2009	13.44	0.830	0.788	0.752	0.841
8	Kema Creek	Nerka	Stream	91	2009	1.86	0.254	0.257	48	2009	15.57	0.875	0.871	0.285	0.266
9	Anvil Beach	Nerka	Beach	92	2006	1.87	0.255	0.260	43	2011	10.88	0.674	0.677	0.868	**0.968**
10	Joe Creek	Nerka	Stream	93	2009	1.88	0.268	0.267	44	2009	14.77	0.841	0.786	0.840	0.753
11	Little Togiak River	Nerka	River	65	2008	1.89	0.254	0.256	42	2008	12.00	0.857	0.822	0.359	0.195
12	Pick Creek	Nerka	Stream	155	2008, 2010	1.88	0.254	0.256	44	2008	16.59	0.750	0.799	0.902	**0.987**
13	Lynx Beach	Nerka	Beach	95	2006	1.88	0.263	0.258	47	2011	10.76	0.766	0.728	0.723	0.644
14	Lynx Cold Tributary	Nerka	Stream	78	2009	1.86	0.256	0.258	47	2009	11.76	0.787	0.790	0.542	0.364
15	Lynx Creek	Nerka	Stream	189	2009, 2010	1.87	0.258	0.258	45	2009	14.79	0.889	0.879	0.127	0.126
16	Lynx Lake Beach	Lynx	Beach	91	2009	1.84	0.243	0.252	42	2009	5.00	0.238	0.237	0.936	0.942
17	Teal Creek	Nerka	Stream	95	2011, 2013	1.87	0.264	0.253	44	2011, 2013	11.90	0.773	0.803	0.475	0.308
18	Stovall Creek	Nerka	Stream	94	2009	1.87	0.264	0.263	47	2009	12.77	0.894	0.869	0.067	0.099

				Neutral SNPs					MHC						
Pop #	Population	Lake	Ecotype	N	Years	A_R	H_O	H_E	N	Years	A_R	H_O	H_E	EW P	Slat P
19	Agulowak River	Aleknagik	River	93	2001	1.88	0.264	0.259	49	2009	9.39	0.633	0.601	0.899	0.932
20	Sunshine Beach	Aleknagik	Beach	94	2012, 2013	1.89	0.251	0.253	43	2012, 2013	17.84	0.767	0.789	0.948	0.913
21	Happy Creek	Aleknagik	Stream	90	2001	1.89	0.267	0.263	45	2001	8.86	0.533	0.500	**0.952**	0.896
22	Hansen Creek	Aleknagik	Stream	94	2004	1.87	0.266	0.261	43	2013	4.95	0.349	0.333	0.857	0.907
23	Eagle Creek	Aleknagik	Stream	79	2007	1.86	0.258	0.258	44	2007	8.95	0.568	0.560	0.901	0.777
24	Bear Creek	Aleknagik	Stream	80	2001	1.86	0.260	0.256	45	2001	8.85	0.467	0.446	**0.971**	0.939
25	Yako Beach	Aleknagik	Beach	94	2006	1.87	0.262	0.263	44	2009	16.72	0.886	0.868	0.452	0.553
26	Whitefish Creek	Aleknagik	Stream	93	2011	1.87	0.265	0.262	47	2011	9.67	0.660	0.573	0.930	0.863
27	Wood River	Aleknagik	River	92	2009	1.88	0.266	0.264	51	2009	17.02	0.902	0.866	0.514	0.383

Table 4.2. Results from four AMOVA analyses. Populations are partitioned by spawning habitat (Ecotypes) and by nursery lake (Lakes). Population groupings are found in Table 4.1. Bold values indicate significance ($P < 0.05$).

	Ecotypes				Lakes			
Source of variation	d.f	SSQ	Var	% of Var	d.f	SSQ	Var	% of Var
Neutral SNPs								
Among groups	2	137.38	0.03	**0.24**	4	222.15	0.03	**0.28**
Among populations within groups	24	687.91	0.09	**0.79**	22	603.14	0.08	**0.73**
Within Populations	5 123	58 206.50	11.36	**98.97**	5 123	58 206.50	11.36	**98.98**
MHC								
Among groups	2	94.33	0.05	**11.64**	4	149.79	0.07	**15.58**
Among populations within groups	24	200.81	0.09	**19.11**	22	145.34	0.07	**14.97**
Within Populations	2 445	779.62	0.32	**69.24**	2 445	779.62	0.32	**69.45**

Figures

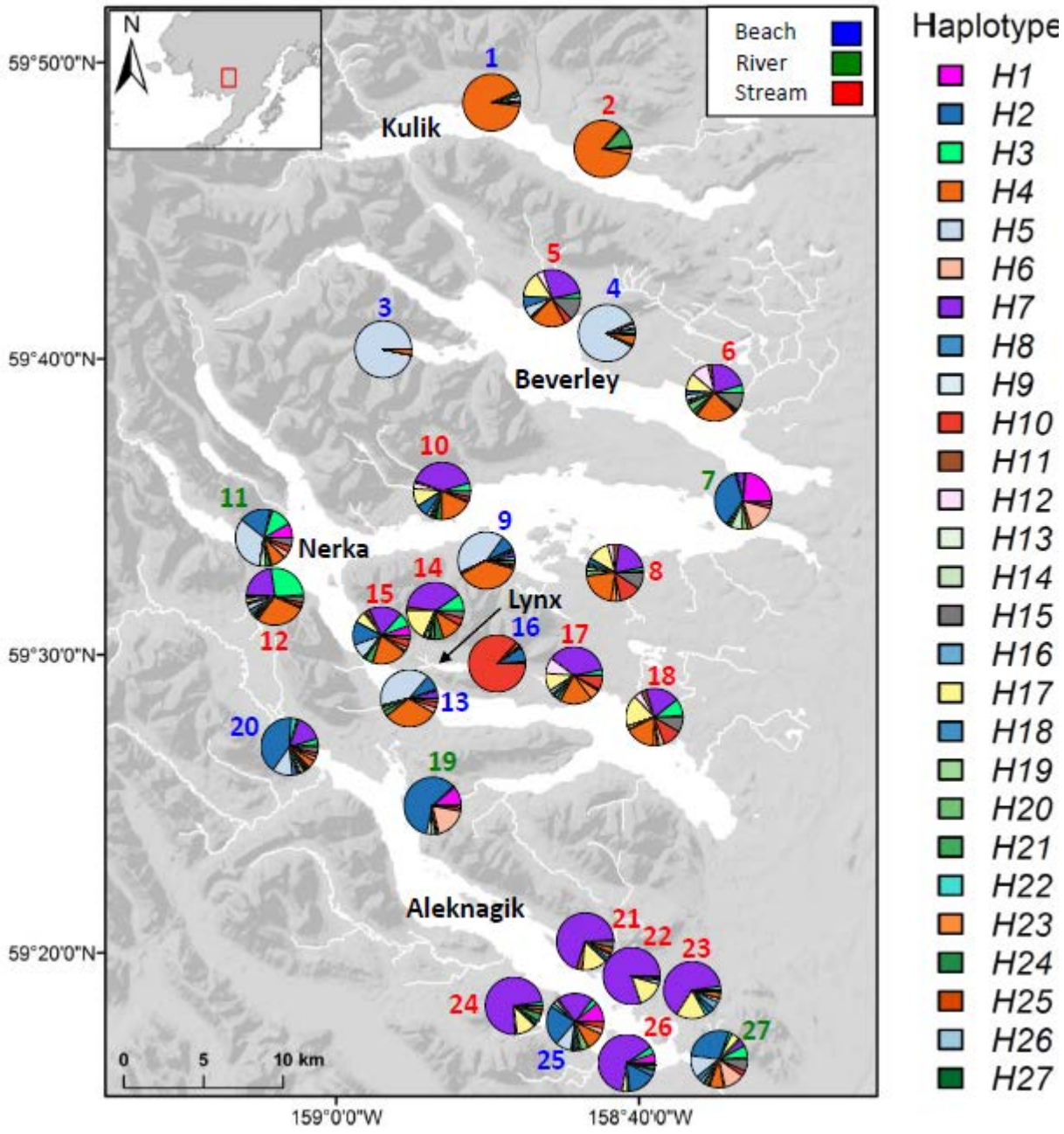

Fig. 4.1. Map of sample populations. MHC haplotype frequencies for each population are shown with pie charts, and population numbers correspond to those in Table 4.1. The color of each population number corresponds to the spawning ecotype of that population and nursery lakes are labeled in black.

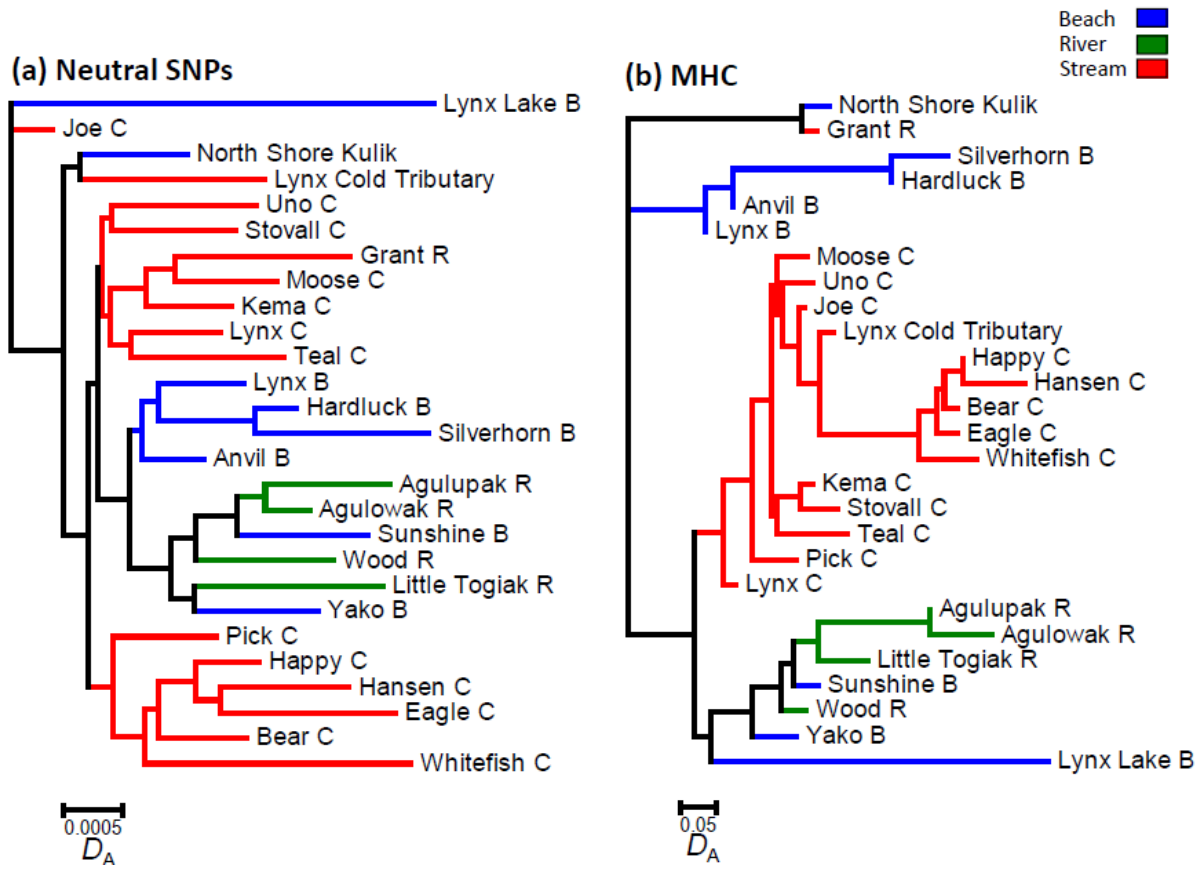

Fig. 4.2. Neighbor-joining trees based on D_A distance for (a) 90 neutral SNPs and (b) the MHC. Branches are colored by spawning ecotype, and population names correspond to those in Table 4.1. Black branches are composed of more than one ecotype. Population names are abbreviated as beach (B), river (R), and creek (C).

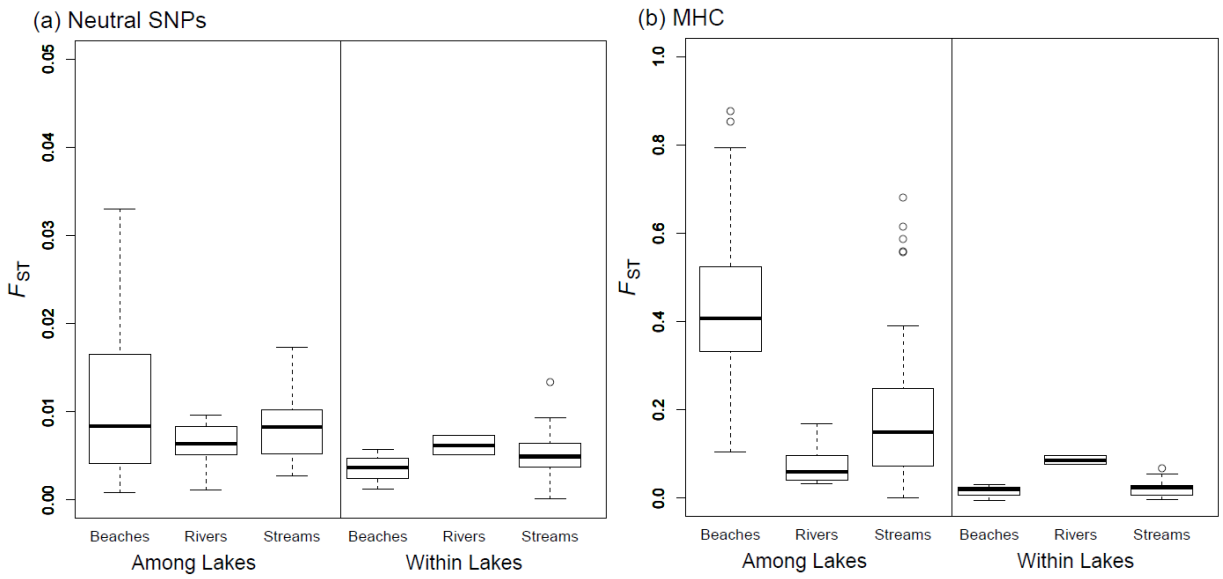

Fig. 4.3. Box and whisker plots of pairwise F_{ST} for comparisons within each ecotype. Comparisons among lakes and within lakes were pooled separately. Datasets for this analysis were (a) 90 neutral SNPs and (b) the MHC. Note the different scales on the y-axis for each plot. Ecotype and nursery lake information for each population is found in Table 4.1.

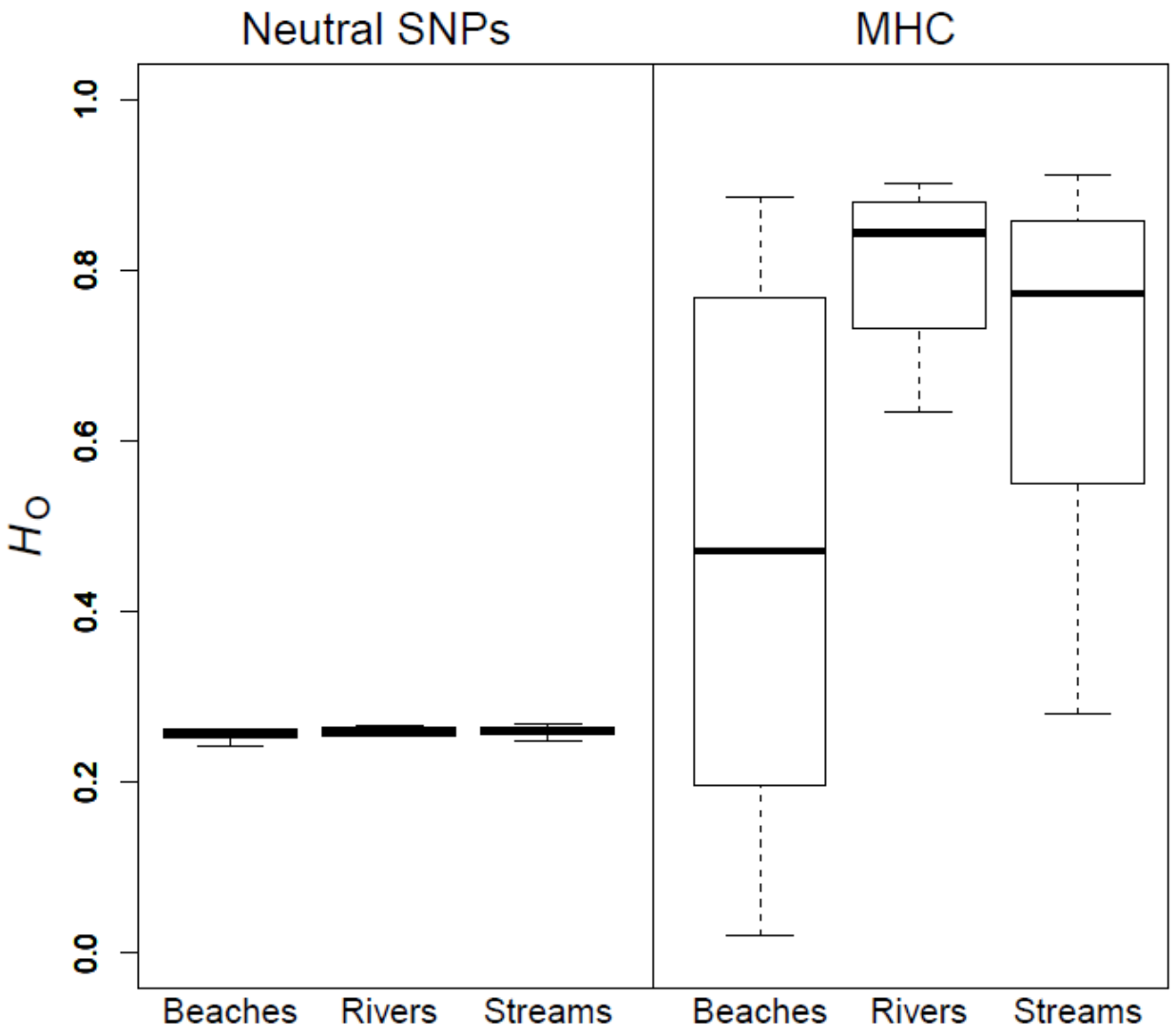

Fig. 4.4. Box and whisker plots of H_O by ecotype for 90 neutral SNPs and the MHC. Values of H_O for each population can be found in Table 4.1.

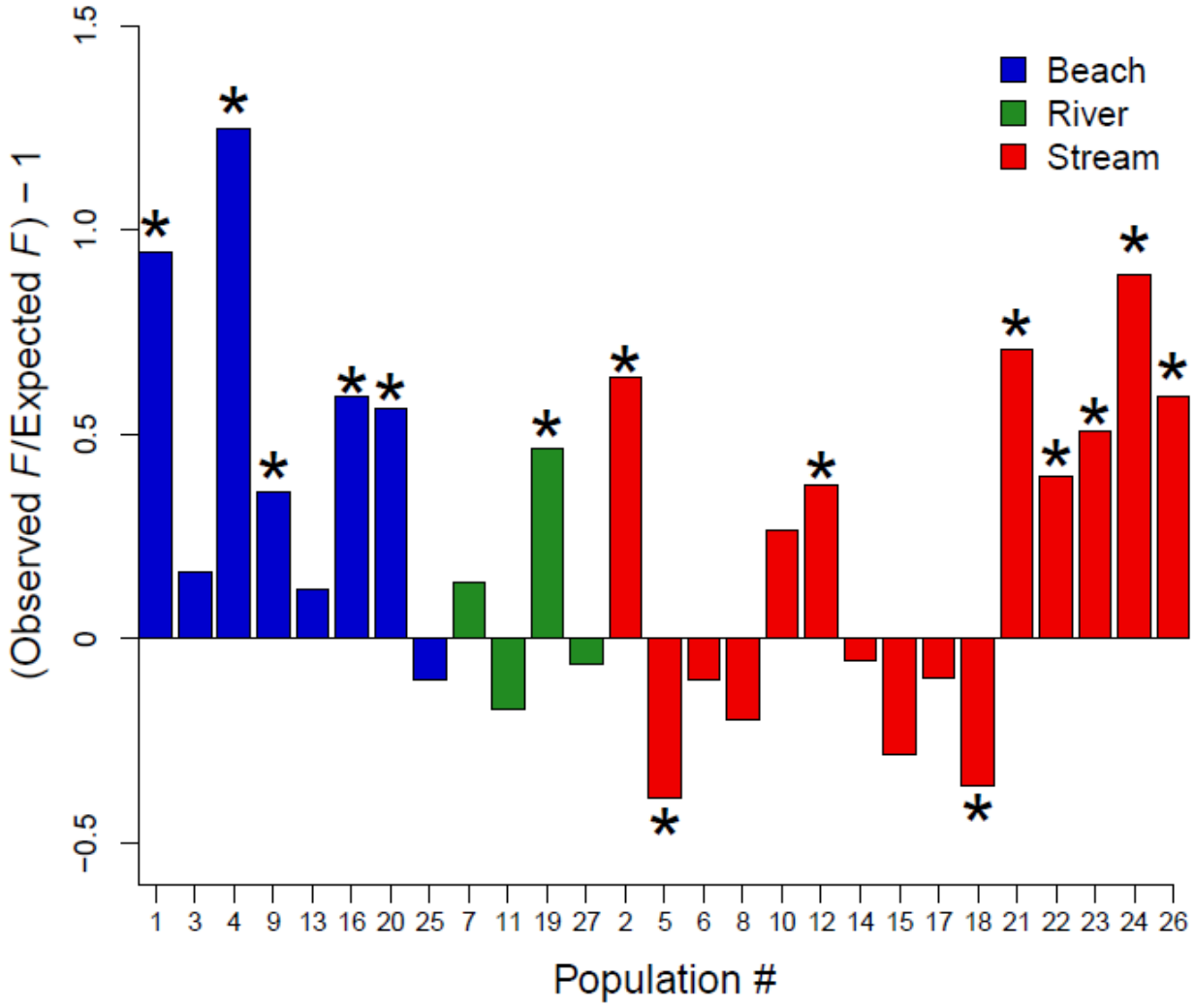

Fig. 4.5. Ratio of observed to expected *F* values from an Ewens-Watterson homozygosity test for each population. One was subtracted from each ratio to standardize the results to zero. Positive values suggest directional selection, negative values suggest balancing selection, and values close to zero suggest neutrality. Populations with significant or marginally significant P-values for the Ewens-Watterson test or the Slatkin's exact P-test ($P < 0.1$, $P > 0.9$) are designated with an *. Population numbers correspond to those found in Table 4.1, and populations within each ecotype are grouped geographically.

Chapter 5

Identification of multiple QTL hotspots in sockeye salmon (*Oncorhynchus nerka*) using genotyping-by-sequencing and a dense linkage map[5]

Abstract

Understanding the genetic architecture of phenotypic traits can provide important information about the mechanisms and genomic regions involved in local adaptation and speciation. Here, we used genotyping-by-sequencing and a combination of previously published and newly generated data to construct sex-specific linkage maps for sockeye salmon (*Oncorhynchus nerka*). We then used the denser female linkage map to conduct quantitative trait locus (QTL) analysis for four phenotypic traits in three F1 families from a wild population. The female linkage map consisted of 6,322 loci distributed across 29 linkage groups and was 4,082 cM long, and the male map contained 2,179 loci found on 28 linkage groups and was 2,291 cM long. QTL analysis using the female map revealed 26 QTL: six for thermotolerance, five for length, nine for weight, and six for condition factor. QTL were distributed non-randomly across the genome and were often found in hotspots containing multiple QTL for a variety of phenotypic traits. These hotspots may represent adaptively important regions and are excellent candidates for future research. Comparing our results with studies in other salmonids revealed several regions with overlapping QTL for the same phenotypic trait, indicating these regions may be adaptively important across multiple species. Altogether, this work demonstrates the utility of genomic data for quickly and efficiently creating genomic resources and studying the genetic basis of important phenotypic traits.

[5] Full citation: Larson, W.A., J.E. Seeb, M.T. Limborg, G.J. McKinney, M.V. Everett, and L.W. Seeb. (Submitted). Identification of multiple QTL hotspots in sockeye salmon (*Oncorhynchus nerka*) using genotyping by sequencing and a dense linkage map. *Journal of Heredity*. Supplementary tables and figures are available online from ProQuest LLC (http://www.proquest.com/) and legends for supplementary tables and figures are available at the end of this chapter.

Introduction

Understanding the genetic basis of phenotypic traits can provide important insights into how organisms adapt to their environment and if they will be able to adapt to changing environments in the future (Stinchcombe & Hoekstra 2008). A common method used to elucidate the genetic basis of phenotypic traits involves examining genotypes at a large number of markers to identify associations between markers or genomic regions and phenotypes of interest (Lynch & Walsh 1998). This method, termed quantitative trait locus (QTL) analysis, has proven useful for identifying loci associated with phenotypes of interest in many model organisms, especially those that are agriculturally important (reviewed in Dekkers & Hospital 2002). However, QTL analysis has historically been difficult to conduct in non-model organisms due the absence of genomic resources and the large number of genetic markers required (Slate 2005).

The increasing availability of genomic data provides a potential solution to this limitation (reviewed in Allendorf *et al.* 2010). Genotyping-by-sequencing (GBS) techniques now make it possible to screen thousands of markers in hundreds of individuals (Tonsor 2012). Additionally, genomic data facilitate the creation of high-density linkage maps that assign markers to specific genomic locations (Davey *et al.* 2011). These advances have enabled QTL studies in a variety of non-model organisms, including cichlid fish (*Tropheops sp*., Albertson *et al.* 2014), great tits (*Parus major*, Santure *et al.* 2013), and moths (*Heliothis sp.*, Groot *et al.* 2013).

Salmonids represent ideal candidates for QTL studies due to their cultural and economic importance, but a lack of genomic resources has historically limited QTL studies to two species of salmonids commonly used in aquaculture, Atlantic salmon (*Salmo salar*) and rainbow trout (*Oncorhynchus mykiss*). Previous QTL studies in these species have provided some important insights into the genetic basis of phenotypic traits such as thermotolerance, size, and condition factor (e.g., O'Malley *et al.* 2003; Perry *et al.* 2005; Reid *et al.* 2005). However, these studies generally employed a relatively small number of loci (< 300), suggesting that many QTL were not discovered due to inadequate coverage across the genome (Santure *et al.* 2013; Johnston *et al.* 2014).

The recent availability of genomic data has facilitated the creation of high density linkage maps for salmonids that provide extensive coverage of the genome and can be used for QTL analysis (see for example Gutierrez *et al.* 2014). These modern linkage maps often include

thousands of loci mapped in both sexes (e.g., Lien *et al.* 2011; Kodama *et al.* 2014) and contain both non-duplicated loci and duplicated loci resulting from an ancient whole genome duplication (WGD) in salmon (e.g., Brieuc *et al.* 2014; Waples *et al.* 2015). Additionally, since many of these maps are created using restriction site associated DNA (RAD) data from the same restriction enzyme (*SbfI*), maps can be easily aligned to discover orthologous regions and marker overlap between species and studies (e.g., Brieuc *et al.* 2014; Kodama *et al.* 2014). Loci on these maps can also be aligned to various genomic resources to investigate the functional significance of certain genomic regions (e.g., Everett & Seeb 2014; Waples *et al.* 2015; McKinney *et al.* Submitted). QTL studies using high-density linkage maps have revealed loci associated with growth and life-history type in rainbow trout (Hecht *et al.* 2012; Miller *et al.* 2012), temperature tolerance and size in Chinook salmon (*Oncorhynchus tshawytscha*, Everett & Seeb 2014), and ecotype in lake whitefish (*Coregonus clupeaformis*, Gagnaire *et al.* 2013a).

Sockeye salmon (*Oncorhynchus nerka*) are one of the most intensely managed species across the Pacific Rim because of their iconic stature, supporting both native cultures and valuable commercial fisheries (Schindler *et al.* 2010; Dann *et al.* 2013); but, the genetic basis of phenotypic traits in this species has rarely been studied. Here, we investigated the genetic basis of four phenotypic traits, thermotolerance, length, weight, and condition factor, in anadromous populations of sockeye salmon from southwestern Alaska. Thermotolerance is an important predictor of how sockeye salmon will respond to climate change (Eliason *et al.* 2011); size-related traits including length and weight are highly correlated with survival and reproductive success in sockeye salmon (Bradford 1995; Quinn 2005); and condition factor is associated with the ability of salmonids to survive stressful environmental conditions (Robinson *et al.* 2008).

The objectives of our study were to: (1) construct a dense linkage map for sockeye salmon using a combination of newly generated and previously published data (Everett *et al.* 2012; Limborg *et al.* Submitted), (2) conduct QTL analysis for four phenotypic traits in three F1 families from a wild population (families from Everett *et al.* 2012), (3) align our QTL with available genomic resources to find potential genes underlying phenotypic variation, and (4) compare our results to previous studies in closely related species. Our study emphasizes the utility of GBS for quickly and affordably creating genomic resources and identifying regions associated with phenotypic traits in non-model organisms. Additionally, the genomic resources

created here will prove extremely valuable for future research on the genetic basis of adaptively important traits in sockeye salmon.

Materials and Methods

Families used for linkage mapping and QTL analysis

Two gynogenetic haploid families (GH1, GH2), one gynogenetic diploid family (GD1), and three diploid families (D3-D5) were available for linkage mapping and QTL analysis (Table 5.1, Fig. 5.1). Families GH1 and GD1 (same female parent) were derived from the landlocked form of sockeye salmon (kokanee) sampled at the southern end of the species range, and families GH2 and D3-D5 were derived from anadromous sockeye salmon sampled at the northern end of the range (Table 5.1). Separate linkage maps have been previously constructed using families GH1 and GD1 (3,245 markers, Limborg *et al.* Submitted) and D5 (1,672 markers, Everett *et al.* 2012, designated HX13 and HX13-WL). Families D3 and D4 were described in Everett *et al.* (2012) (designated HX6, HX8 in that study) but sequencing was not conducted on these families. We sequenced families D3 and D4 and constructed an additional haploid cross (GH2) to increase the number of mapped markers derived from the northern extent of the range and provide additional replication for QTL analysis. The methods used to preserve samples, validate ploidy (families GH1, GD1), and conduct genotyping (family GD1) have been previously described by Limborg *et al.* (Submitted, families H1, G1) and Everett *et al.* (2012, families D3-D5).

Family GH2 was created by combining eggs with UV-irradiated sperm following the methods of Thorgaard *et al.* (1983). Embryos were preserved in 100% ethanol as close to hatch as possible, and DNA from the parents and offspring was isolated using QIAGEN DNAeasy 96 Tissue Kits (Qiagen, Valencia, California). To confirm ploidy, we genotyped the parents and offspring for 96 EST-derived 5'-nuclease assays (Elfstrom *et al.* 2006; Storer *et al.* 2012) following the methods of (Smith *et al.* 2011) and (Everett & Seeb 2014). Genotypes for these assays were also available for families GH1 and D3-D5 and were used for linkage mapping and QTL analysis.

Restriction site associated DNA (RAD) sequencing, SNP discovery, and genotyping

RAD sequencing was conducted using the enzyme *SbfI* following the methods of Baird *et al.* (2008) and Everett *et al.* (2012). Sequence data was then analyzed with the *STACKS* software package (version 1.20, Catchen *et al.* 2011; Catchen *et al.* 2013) and genotyping methods developed by Waples *et al.* (2015). Different parameters were used to genotype haploid and

diploid individuals (see supplementary methods). As a final step before linkage mapping, genotypes were filtered to remove individuals and SNPs with > 20% missing data. Additional information on RAD sequencing and genotyping can be found in supplementary methods.

Linkage mapping

Separate female and male linkage maps were constructed with the program *LepMap* (Rastas *et al.* 2013). *LepMap* is a fast and memory efficient program that utilizes data from multiple families simultaneously to construct consensus linkage maps. Parameters for *LepMap* analysis were identical to those of McKinney *et al.* (Submitted) with one exception: the LOD (log10 odds) score limit used to form linkage groups (LGs) in our study was 9.5 for the female map and four for the male map. We excluded data from diploid families for markers that were heterozygous for the same alleles in both parents because phase cannot be unambiguously determined in the offspring.

Gynogenetic diploids (half-tetrads) provide information about marker-centromere distances from recombination events during meiosis and facilitate placement of centromeres based on observed heterozygosity (also known as y, Thorgaard *et al.* 1983). We placed centromeres on the female and male linkage maps using genotype data available from the gynogenetic diploid family GD1 described and genotyped in Limborg *et al.* (Submitted). Centromeres on the female map were defined as the region of each LG containing all markers with heterozygosity < 0.1 (Limborg *et al.* Submitted). Centromeres on the male map were defined as the region containing all markers that were found to be centromeric in the female map. Arms for each LG on the female map were arbitrarily assigned based on centromere location and do not correspond to previous studies.

We compared our linkage map to existing maps for sockeye salmon and Chinook salmon to validate our results and establish orthologous relationships with other salmonid species. First, we compared our map to the map generated by Limborg *et al.* (Submitted) and named our LGs based on this map. No alignment step was necessary for this comparison because the locus names were identical across studies. We then compared our map to another existing map for sockeye salmon (Everett *et al.* 2012). Loci shared between studies were identified with *BLASTN* (parameters: minimum alignment length of 57 bp, 95% identity, and no more than two mismatching bases). Finally, we aligned our map to one for Chinook salmon (McKinney *et al.* Submitted) to establish orthologous relationships between the two species (*BLASTN* parameters:

minimum alignment length of 80 bp, 90% identity, and no more than four mismatching bases). Information from this alignment was combined with data presented in Brieuc *et al.* (2014) and Kodama *et al.* (2014) to establish and report orthologous relationships among sockeye salmon, Chinook salmon, coho salmon (*O. kisutch*), rainbow trout, and Atlantic salmon.

Thermal challenge and other phenotypic data

A thermal challenge was conducted on 96 offspring from families D3-D5 following methods similar to Everett and Seeb (2014). Prior to the thermal challenge, offspring from each family were raised for 30 days post hatch in separate aquaria kept at 11° C. Water in each aquarium was then gradually replaced with water preheated to 29° C until the temperature reached 25° C. The first 48 individuals from each family that lost equilibrium were removed and recorded as thermosusceptible. After 48 individuals lost equilibrium (approximately two hours), the remaining 48 individuals were sampled and considered thermotolerant. Total length and weight were recorded for each individual, and samples were preserved in RNALater (Life Technologies, Carlsbad, California). Condition factor (K), a standardized measure of fish health, was then calculated using the formula $K=W/L^3*10^5$ where W = weight in grams and L = length in mm (Bagenal & Tesch 1978).

Summary statistics for length, weight, and condition factor were calculated separately for thermotolerant and thermosusceptible individuals from each family, as well as the family as a whole. We conducted Student's t-tests to investigate the hypothesis that phenotypic distributions were significantly different between thermotolerant and thermosusceptible individuals and among families (alpha = 0.05). Finally, we plotted weight versus length for each family and visually examined the distribution of thermotolerant and thermosusceptible individuals in relation to the line of best fit derived for each family.

QTL analysis

QTL analysis for thermotolerance, length, weight, and condition factor was conducted separately for each diploid family with the R package *qtl* (Broman *et al.* 2003) and methods similar to Hecht *et al.* (2012). First, we identified single QTL using the function *scanone*. We then iteratively ran the *scanone* function, adding previously identified QTL as cofactors, until no additional QTL were detected. Experiment and LG-wide significance thresholds (alpha = 0.05) were determined with permutation tests (1,000 iterations) implemented in *scanone*. We

considered QTL with LOD scores > 3 that were also above the experiment or LG-wide significance threshold as significant (Lander & Kruglyak 1995).

Potential interactions between QTL discovered with *scanone* were investigated with the *addint* function. We then refined the positions of significant QTL using the *refineqtl* function. Finally, we fit a multiple-QTL model including interaction terms for all QTL found for a given phenotypic trait with the *fitqtl* function. QTL that were not significant in the context of the full model ($P > 0.05$) were removed and *refineqtl* and *fitqtl* were rerun until all QTL included in the full model were significant. The percentage of variation explained (PVE) by each QTL was obtained from the results of *fitqtl*. Approximate 95% confidence intervals for the position of each QTL were calculated with the LOD drop-off method implanted in the *lodint* function (1.5 LOD drop, Visscher *et al.* 1996; Dupuis & Siegmund 1999).

Paired-end assembly, alignment to genomic resources, and functional annotation

We conducted paired-end assemblies for each locus to increase query length for functional annotation and alignment to genomic resources. Paired-end sequences from the six parents of families D3-D5 were assembled with the alignment program *CAP3* (150 bp minimum alignment length, Huang & Madan 1999) following the methods of Etter *et al.* (2011) and Waples *et al.* (2015). Consensus sequences for each locus were aligned to Atlantic salmon genome scaffolds (ICSASB_v1 ; GenBank accession: GCA_000233375.3). Alignments were conducted with the longest sequence available from each locus using *BLASTN* (parameters: >90% identity, $\leq$ 4 mismatches per 100 bp, $\leq$ 1 gap per 100 bp, and alignment length > 80% of query sequence). Scaffolds were placed on the linkage map if at least three loci on the same LG aligned to the scaffold and the order of the loci on the linkage map was concordant with their order on the scaffold. Consensus sequences for each locus were also aligned to all expressed sequence tags (ESTs) for sockeye salmon in the cGrasp database (http://web.uvic.ca/grasp/) using *BLASTN* (parameters: >90% identity, $\leq$ 4 mismatches per 100 bp, $\leq$ 1 gap per 100 bp, and alignment length > 50% of query sequence). If multiple alignments met these parameters for a single locus, the alignment with the lowest e-value was retained.

Functional annotation was conducted by aligning paired-end consensus sequences for each locus to the Swiss-Prot database using *BLASTX*. The alignment with the lowest e-value < 10^{-4} for each locus was accepted as the annotation. Additional annotations were attempted for QTL peak loci that were placed on the Atlantic salmon genome by aligning 100,000 bp of 3' and

5’ flanking sequence for each locus to the Swiss-Prot database using BLASTX and the parameters described above. QTL peak loci that did not directly align to the genome but were found at map locations spanned by a scaffold were aligned to the scaffold with relaxed parameters. If the QTL could be placed in the correct scaffold, annotation was attempted using the methods described above.

Results

Sequencing, SNP discovery, and genotyping

RAD sequence data was obtained from 525 individuals across five families (Table 5.1). Sequencing depth varied substantially by family ranging from an average of 1,117,053 sequences per individual for family GH2 to 4,671,087 sequences per individual for family D5 (excluding low quality individuals). SNP discovery using the RAD data revealed 11,377 polymorphic loci that were genotyped in > 80% of individuals. We added 80 polymorphic 5’-nuclease assays that were genotyped in > 80% of individuals to this dataset. Finally, we removed 34 individuals that were genotyped at < 80% of loci resulting in a final dataset of 491 individuals genotyped at 11,457 loci.

Linkage mapping

We constructed a female linkage map containing 6,322 loci distributed across 29 LGs and a male linkage map containing 2,179 loci distributed across 28 LGs (Fig. 5.2, S5.1, Table 5.2, Table S5.1). The total length of the female map was 4,082 cM, and the total length of the male map was 2,291 cM. These maps contained 7,367 unique loci, with 1,143 loci found on both maps. Placement of centromeres using gynogenetic diploids was successful for all LGs in the female map, and we were able to place centromeres on 19 of 28 LGs in the male map using information from markers found to be centromeric on the female map. The female map contained six acrocentric and 23 metacentric LGs, and LG type was well conserved between male and female maps (excluding LG So9). LG So9 has been previously identified as the sex chromosomes for sockeye salmon (Limborg *et al.* Submitted) and is composed of two pairs of acrocentric chromosomes (X_1 and X_2) in females. In males, one copy each of X_1 and X_2 are fused into a single metacentric chromosome (Y) resulting in a single copy of both X_1, X_2 and Y in males (Thorgaard 1978; Faber-Hammond *et al.* 2012). We designated the two acrocentric sex LGs in the female map as So9 and So9.5.

As expected, marker order and LG designations were highly concordant between our map and two previous maps for sockeye salmon constructed using families included in this study (Everett *et al.* 2012; Limborg *et al.* Submitted, data not shown). However, some differences did exist. We identified two differences between our map and Everett *et al.* (2012): (1) all markers that we identified from LG 29 in Everett *et al.* (2012) were placed on LG So27 in our map and (2) LG 9 in Everett *et al.* (2012) was composed of two separate LGs in our female map. We also identified differences between our map and Limborg *et al.* (Submitted): (1) we were able to join LG 18a and 18b from Limborg *et al.* (Submitted) into a single metacentric LG, (2) we identified LG 17 as metacentric rather than acrocentric, and (3) we identified two additional homeologous relationships in this study that were not identified in Limborg *et al.* (Submitted) (So10a-So28b, So27a-So28a, Table 5.3).

We identified orthologous relationships between Chinook and sockeye salmon for all 52 LG arms on our maps and extended these relationships to coho salmon, rainbow trout, and Atlantic salmon (Table 5.2). Each orthologous relationship between sockeye and Chinook salmon was supported by one to 15 marker pairs (average of seven pairs per relationship, 353 total markers shared between the two species). All LG arms in Chinook salmon aligned to a single LG arm in sockeye salmon except for LG So17, where both the a and b arms aligned to a single arm in Chinook salmon (Ots01p). Additional information is necessary to validate this finding.

We mapped 1,101 potentially duplicated loci on the female linkage map using haploids. High concentrations of duplicated loci were found near the distal ends of 14 LGs (Fig. 5.2), and eight homeologous relationships were identified by mapping 94 duplicated loci that segregated at both paralogs (Table 5.3). Comparisons with Chinook salmon, coho salmon, rainbow trout, and Atlantic salmon revealed that orthologous suites of chromosome arms are involved in homeologous pairing across these species (Table 5.3).

Phenotypic data

Phenotypic data for thermotolerance, length, weight, and condition factor was obtained from three diploid families (Table 5.4). Distributions of length and condition factor were significantly different among the three families ($P < 0.05$), but distributions of weight were not. Thermotolerant individuals in families D3 and D4 were significantly larger than thermosusceptible individuals ($P < 0.05$). However, the opposite trend was present in family D5

(Fig. 5.3, Table 5.4). Condition factor was higher for thermotolerant individuals in all three families but was only significantly higher in families D3 and D4.

QTL analysis

We conducted QTL mapping for four phenotypic traits in three diploid families using 3,496 unique loci placed on the female linkage map. Family D3 contained 2,218 loci suitable for QTL mapping; family D4 contained 2,160 loci; and family D5 contained 2,212 loci. We identified 26 QTL with peaks at 22 unique genomic positions (Table 5.5). Of these QTL, two were identified as significant at the LG and experiment-wide level, and 24 were identified as significant at the LG level. The percentage of variation explained by each QTL ranged from 6.18% to 34.08%. No significant epistatic interactions were found among QTL ($P > 0.1$).

The number of QTL identified varied substantially by family, phenotypic trait, and LG. We identified four QTL in family D3, ten QTL in family D4, and 12 QTL in family D5. The phenotypic trait with the most QTL was weight (nine), followed by thermotolerance and condition factor (six), and length (five). Two LGs contained four QTL (So6, So11), So LG contained three QTL (So28), and the remaining LGs contained two or fewer QTL (Fig. 5.4, Fig. S5.2).

Genomic regions containing overlapping QTL from different families were generally uncommon, but we did see this pattern on LGs So6 and So11 (Fig. 5.4). So6 contained overlapping QTL for thermotolerance (family D5), weight (family D5), and condition factor (families D3 and D4), and So11 contained overlapping QTL for thermotolerance (family D3), length (family D5), and weight (families D3 and D5). We also found eight QTL that shared a peak marker with another QTL within the same family. Shared markers were most often associated with QTL for length and weight, but we did find one example of a shared peak QTL marker for thermotolerance and weight (locus 7896).

Paired-end assembly, alignments to genomic resources, and functional annotation

Construction of consensus sequences longer than 150 bp from PE data was possible for 7,146 of 7,367 loci (97%, average length 258 bp, Table S5.1). Consensus sequences from 480 loci were successfully aligned to the Atlantic salmon genome and used to anchor 97 unique scaffolds spanning approximately 25% of the total female map (Table S5.1, S5.2). Alignment to sockeye salmon ESTs from the cGRASP database was possible for 98 loci (Table S5.1).

Functional annotation from RAD sequence data was successful for 840 of 7,367 loci (11%, Table S5.1). Transposable elements comprised approximately 25% of these annotations; other common functional groups included DNA polymerases and genes involved in regulation of programmed cell death. Annotations were successful for eight QTL peak markers (nine total annotations, Table 5.5, Table S5.3). Of these annotations, five were derived from RAD sequence data, three were obtained using flanking sequence from the Atlantic salmon genome, and one was obtained from previous annotation of a 5'-nuclease assay. Notable annotations included a QTL for condition factor that aligned to a gene involved in fatty acid synthesis (FAS, locus 5975) and a QTL for length that aligned to a gene involved in metabolism and biosynthesis (RAG, locus One_RAG3-93).

Discussion

Linkage mapping and alignment to genomic resources

The first objective of this study was to create a dense linkage map for sockeye salmon that could be used for QTL analysis. Our linkage map was constructed from a combination of newly acquired data along with data from two previous mapping studies (Everett *et al.* 2012; Limborg *et al.* Submitted) and contained more than double the number of markers compared to those previous maps. Additionally, our map included data from both the northern and southern extremes of the species range in North America.

Comparisons between female and male maps revealed that the female map was approximately twice as long as the male map and that markers on the male map tended to group towards the centromeres. Similar results have been well-documented in salmonids and are thought to occur because of sex-specific differences in the distribution of recombination sites across chromosomes (Lien *et al.* 2011; Everett *et al.* 2012; Kodama *et al.* 2014). We did not merge sex-specific maps because of these differences and, instead, utilized the denser female map for QTL analysis and alignments to genomic resources. Although the male map was not heavily utilized in this study, the high levels of recombination found in the telomeric regions in males makes this map an important resource for future studies attempting to order telomeric markers or genome scaffolds (Lien *et al.* 2011).

Mapping of duplicated loci on the female map using haploids revealed similar patterns of homeology compared to previous studies (reviewed in Allendorf *et al.* 2015). High concentrations of duplicated loci were found in the telomeric regions of eight pairs of

homeologous chromosomes, and these chromosomes were orthologous with chromosomes involved in homeologous pairing in other species (Brieuc *et al.* 2014; Kodama *et al.* 2014; McKinney *et al.* Submitted). This finding provides further evidence for the existence of a conserved set of eight homeologous chromosome arms containing high concentrations of duplicated loci across all salmonids (reviewed in Allendorf *et al.* 2015).

Alignment of our mapped loci to existing genomic resources provided important functional annotations and allowed us to anchor our map in the context of four other salmonids. Functional annotations were similar to past RAD studies in salmonids and included a high proportion of transposable elements (e.g., Everett *et al.* 2012; Everett & Seeb 2014; Larson *et al.* 2014c). Notably, transposable elements comprised approximately 25% of annotations, but only 10% of annotations for duplicated loci (c.f., Waples *et al.* 2015; McKinney *et al.* Submitted). Transposable elements are hypothesized to serve an important role in rediploidzation, possibly explaining the reduced frequencies of transposable elements that we observed in regions that are not fully rediploidized (Waples *et al.* 2015; McKinney *et al.* Submitted). Alignment with RAD-derived linkage maps from other species revealed orthologous relationships for all 52 LG arms in sockeye salmon. The success of these alignments highlight the utility of comparing findings from RAD studies across species and demonstrate an advantage for the continued use of *SbfI* or other enzymes with restriction sites that overlap with *SbfI* to ensure the compatibility of future studies.

We were able to successfully anchor 97 scaffolds from the Atlantic salmon draft genome to our map, representing about 25% of the total map length. The continuous sequence provided by these scaffolds represents an excellent tool for functional annotation and exploration of genomic regions that are proximate to loci of interest (Allendorf *et al.* 2010). However, we were unable to anchor large portions of our linkage map to the genome, likely due to the draft nature of the genome assembly, the marker density of our linkage map, and sequence divergence between sockeye and Atlantic salmon.

Phenotypic variation in experimental families

Significant variation in size and condition factor existed among all three families in our study. This variation is likely a result of genetic rather than environmental effects because the families were raised in similar environments and these traits have been shown to have high heritability in other salmonids (reviewed in Garcia de Leaniz *et al.* 2007).

Thermotolerant individuals had higher condition factors in all three families and were larger in two of the three families. A positive correlation between condition factor and temperature tolerance was also demonstrated in cutthroat trout (*O. clarki*, Robinson *et al.* 2008). These results suggest that aspects of body composition that are correlated with condition factor, such as lipid and protein content, may be important components of temperature tolerance in salmonids (Robinson *et al.* 2008). Significant correlations between size and upper temperature tolerance have also been observed in rainbow trout (Perry *et al.* 2005), but these correlations were not consistent among experimental families and appeared to be related to parental effects. A more comprehensive study of size and thermotolerance across five species of Pacific salmon also found no consistent correlation between these traits (Brett 1952). Taken together, these results suggest that body size is unlikely to be an accurate predictor of thermotolerance, which may explain the inconsistent trend between these two traits in our families.

QTL analysis

We identified 26 QTL related to four phenotypic traits in three experimental families of anadromous sockeye salmon. Each trait displayed between zero and four QTL within each family. Across families, each trait contained at least one QTL that explained greater than 10% of the phenotypic variation. Prevailing theory suggests that most continuous phenotypic traits such as those that we examined are likely controlled by many genes of small effect (Roff 2007). The fact that we found relatively few QTL for each trait with generally large effect sizes appears to contrast this theory. However, it is important to note that the experimental design and sample sizes used in this study likely prevented us from finding the majority of small-effect QTL related to each trait. Classical QTL studies generally employ a multigenerational design and sample sizes of more than 300 individuals per family to maximize their power to detect QTL and accurately estimate QTL effect sizes (Beavis *et al.* 1994; Xu 2003; Erickson *et al.* 2004). Our study design employing ~100 individuals from F1 families derived from a wild population undoubtedly limited our power to detect QTL, but this design also provides a cost effective and practical template to discover large-effect QTL in wild populations of non-model organisms.

The distribution of QTL across the genome in our study was non-random and was characterized by a few regions containing high concentrations of QTL related to multiple phenotypic traits interspersed within large regions containing very few QTL. Past studies have suggested that regions containing large numbers of QTL (QTL hotspots) are likely involved in

the early stages of speciation as well as the evolution of different life history types (Via & West 2008; Hecht *et al.* 2012; Gagnaire *et al.* 2013a). Extensive life history diversity exists in populations of sockeye salmon from our study system, including the presence of three distinct ecotypes associated with environment used for spawning (Hilborn *et al.* 2003). These ecotypes are characterized by differences in a number of traits, such as size (Quinn *et al.* 2001) and condition factor (personal observation), and experience substantially different temperature regimes, likely leading to adaptive differences in thermotolerance. Although our families were derived from a single ecotype (stream type), the QTL hotspots that we discovered appear to control at least some of the variation in traits that distinguish ecotypes, providing evidence that these hotspots may be involved in local adaptation and the formation of distinct ecotypes. Future research should focus on investigating the co-location of QTL hotspots with loci displaying signatures of divergent selection among ecotypes to provide further support for this hypothesis (c.f., Via & West 2008; Gagnaire *et al.* 2013b).

We found four pairs of QTL that shared a peak marker with another QTL within the same family. All but one of these pairs contained QTL associated with length and weight, an anticipated result given the high degree of correlation between these two size-related traits. The remaining pair was associated with thermotolerance and weight. A QTL affecting thermotolerance and size was also discovered in rainbow trout and was hypothesized to be the result of either pleiotropy or linkage disequilibrium (Perry *et al.* 2005). Pleiotropy occurs when a single gene influences multiple seemingly unrelated phenotypic traits and may explain the results we observed. It is also possible that linkage disequilibrium between two proximate genes related to thermotolerance and size may be responsible for these results.

No QTL peak markers were replicated across families. This may be the result of different segregation patterns in these families and/or low power to detect small effect QTL due to sample size limitations. However, we did observe overlapping confidence intervals for QTL related to the same traits across families, providing strong evidence for the existence of the QTL.

Comparing the locations of single QTL and QTL hotspots across related species can provide important information about the genetic architecture of phenotypic traits (Reid *et al.* 2005). Comparisons of QTL discovered in this study to QTL found in rainbow trout and Chinook salmon revealed several orthologous regions of interest. For example, the QTL hotspot found on LG So6 in this study corresponded to a region of chromosome Omy14 in rainbow trout

that contained multiple QTL related to growth, condition factor, and morphology (Hecht *et al.* 2012). We also discovered that LG So7, which contained QTL for thermotolerance and weight in this study, was orthologous with regions containing QTL for length and thermotolerance in rainbow trout (Perry *et al.* 2005) and length in Chinook salmon (Everett & Seeb 2014). These regions represent ideal candidates for future research on the genetic basis of phenotypic traits in salmonids. Additionally, these results further illustrate the importance of a conserved RAD-seq protocol among salmonids facilitating accumulated evidence from orthologous replicates.

Functional annotation of QTL can potentially be used to identify the genes underlying phenotypic variation in traits of interest (Pavlidis *et al.* 2012). We were able to annotate about a third of our QTL, and several of these QTL annotated to genes with plausible connections to the phenotypic traits examined. For example, the QTL for condition factor on LG So28 annotated to a gene involved in fatty acid synthesis and the QTL for length on LG So9 annotated to a gene involved in metabolism and biosynthesis. However, it is important to note that the traits we examined are likely controlled by many genes with obscure roles and that "storytelling" using functional annotations should be approached with caution (Pavlidis *et al.* 2012). Nevertheless, information about genes found in genomic regions of interest is vital for increasing our understanding of the genetic architecture of phenotypic traits and guiding future research (Allendorf *et al.* 2010). It is also important to note that alignments to the Atlantic salmon genome proved helpful for annotating additional QTL. As this resource improves, it should be possible to annotate a much larger proportion of QTL discovered with RAD data.

Conclusions

We successfully constructed the densest linkage map to date for sockeye salmon and used this map to detect QTL for four phenotypic traits. QTL were distributed non-randomly across the genome and often colocalized within QTL hotspots. These hotspots may be important for adaptation and represent ideal candidates for future studies seeking to understand processes of local adaptation in salmonids. Moreover, comparison of our results with QTL studies in rainbow trout and Chinook salmon revealed several regions with overlapping QTL for similar phenotypic traits. These results provide evidence that the genetic basis of some phenotypic traits is similar across species and argue for additional interspecies comparisons.

Supplementary methods

Restriction site associated DNA (RAD) sequencing, SNP discovery, and genotyping

RAD sequencing of family GH1 was described in Limborg *et al.* (Submitted) and sequencing of family D5 was described in Everett *et al.* (2012). RAD libraries for crosses GH2, D3, and D4 were constructed with the enzyme *SbfI* following the methods of (Baird *et al.* 2008) and Everett *et al.* (2012). Sequencing of the offspring and parental female from family GH2 was conducted on an Illumina HiSeq2000 (single-end 100 bp reads (SE100), ~48 individuals per lane), and sequencing of the offspring from families D3-D4 was conducted on an Illumina Genome Analyzer II (single-end 80 bp reads (SE80), 16 individuals per lane). RAD sequence data from the parents of families D3-D5 was available from Everett *et al.* (2012, paired-end 2 x 80 bp reads (PE80)).

We used the *STACKS* software package (version 1.20, Catchen *et al.* 2011; Catchen *et al.* 2013) and methods developed by Waples *et al.* (2015) to identify and genotype SNPs from RAD sequence data. Our analysis pipeline consisted of quality filtering and demultiplexing raw sequences using *process_radtags*, identifying SNPs within individuals using *ustacks*, creating a catalog of loci with *cstacks*, and exporting and classifying individual genotypes with *sstacks* and a combination of *genotypes*, *populations*¸ and a maximum likelihood (ML) method developed by Waples *et al.* (2015). Analysis was conducted separately for the haploid and diploid families due to differences in sequence length and ploidy.

Identification and genotyping of SNPs in the haploid families followed the methods of Limborg *et al.* (Submitted) with one exception. Rather than building a catalog de novo, we incorporated the female parent from family GH2 into an existing catalog containing the female from family GH1 (catalog described in Limborg *et al.* Submitted). Individual genotypes were obtained by applying the ML genotyping algorithm developed by Waples *et al.* (2015) to a file of observed haplotypes generated by *populations*. This algorithm distinguishes duplicated and non-duplicated loci in haploid families and exports the resulting genotypes for use in downstream analyses.

SNP discovery and genotyping in the diploid families was conducted with parameters optimized for diploid individuals, non-duplicated loci, and SE80 data. The relevant flags and parameters used in each STACKS module were *process_radtags* (-c -r -q -E phred33 --filter_illumina -t 73), *ustacks* (-m 2 -M 2 -H -d -r --max_locus_stacks 3 --model_type bounded --bound_high 0.05), *cstacks* (-n 2), and *genotypes* (-c). We enabled the deleveraging and removal algorithms and employed smaller values of -M and -n compared to the haploid families to ensure

that duplicated loci were not merged into a single locus (Mastretta-Yanes *et al.* 2014; Waples *et al.* 2015). A single catalog was created with the parents from families D3-D5 and this catalog was used to genotype all offspring.

To create a single dataset for linkage mapping, we aligned the sequences from the haploid catalog (SE100) and the diploid catalog (SE80). Alignments were conducted with *BLASTN* (parameters: minimum alignment length of 70 bp, 95% identity, and no more than two mismatching bases). Loci found in both catalogs were renamed based on the SE100 catalog, loci found in only the SE100 catalog retained their name, and loci found in only the SE80 catalog were renamed sequentially after the last entry in the SE100 catalog. The SE100 contained 764,304 entries, and every locus name > 764,304 represents a locus found in only the SE80 catalog.

As a final step before linkage mapping, genotypes from all five families were filtered to remove individuals and SNPs with > 20% missing data and combined into a single dataset. Genotypes from the 96 5'-nuclease assays described above were also filtered using the same parameters and incorporated into this dataset.

Tables

Table 5.1. Sampling information for the five families used to place markers on the linkage map and conduct QTL analysis. The life history column denotes whether the family was constructed from resident sockeye salmon that remain in freshwater (kokanee) or from anadromous sockeye salmon. The construction of families D3-D5 was described in Everett *et al.* (2012); RAD sequencing of families D3 and D4 was conducted in the current study and sequencing of family D5 was conducted in Everett *et al.* (2012). Genotypes from gynogenetic diploid produced from family GH1 (family GD1) were used for centromere placement (see text and Limborg *et al.* Submitted). See Figure 5.1 for a visualization of the experimental design for this study.

Family	Source	Ploidy	Life history	Sampling location	Number of Individuals		Sequencing method[1]	Average no. reads/individual
					Mapping	QTL analysis		
GH1	(Limborg *et al.* Submitted)	Haploid	Resident	Puget Sound, Washington, USA	92	0	SE100	2,500,000
GD1	(Limborg *et al.* Submitted)	Diploid	Resident	Puget Sound, Washington, USA	NA[2]	NA[2]	NA[2]	NA[2]
GH2	This study	Haploid	Anadromous	Bristol Bay, Alaska, USA	86	0	SE100	1,117,053
D3	(Everett *et al.* 2012)	Diploid	Anadromous	Bristol Bay, Alaska, USA	79	79	SE80	1,387,758
D4	(Everett *et al.* 2012)	Diploid	Anadromous	Bristol Bay, Alaska, USA	88	87	SE80	1,323,787
D5	(Everett *et al.* 2012)	Diploid	Anadromous	Bristol Bay, Alaska, USA	138	96	SE80, SE100, PE80	4,671,087

[1] SE100: single-end 100 bp Illumina sequencing; SE80: single-end 80 bp sequencing; PE80: paired-end 80 bp sequencing.

[2] Genotypes from gynogenetic diploids were available from Limborg *et al.* (Submitted) and were used to place centromeres. See Limborg *et al.* (Submitted) for information on sample sizes and sequencing.

Table 5.2. Summary of male and female linkage maps for sockeye salmon. Arms for each LG were assigned based on centromere location. LG type denotes acrocentric (A) and metacentric (M) LGs. Orthology support is the number of homologous loci shared between sockeye salmon and Chinook salmon for each orthologous relationship. See Brieuc *et al.* (2014) and Kodama *et al.* (2014) for additional information on orthologous relationships between Chinook salmon, coho salmon, rainbow trout, and Atlantic salmon. LG arm designations and cM positions for markers in this map do not correspond to Limborg *et al.* (Submitted), but markers are named the same.

Sockeye	Length (cM)		# markers								
LG	Female	Male	Female	Male	LG type	Sockeye LG arm	Chinook chromosome	Coho LG	Rainbow trout chromosome	Atlantic Salmon chromosome	Orthology support
So1	82.88	42.57	170	44	A	So1a	Ots26	Co26	Omy22	Ssa21	12
So2	200.06	26.44	290	71	M	So2a	Ots07p	Co05a	Omy07p	Ssa17qb	6
					M	So2b	Ots07q	Co05b	Omy07q	Ssa22	11
So3	155.99	144.96	288	64	M	So3a	Ots03q	Co02b	Omy03q	Ssa25	9
					M	So3b	Ots03p	Co02a	Omy03p	Ssa02p	7
So4	129.91	214.42	272	110	M	So4a	Ots13p	Co17a	Omy18q	Ssa27	10
					M	So4b	Ots20	Co23	Omy05p	Ssa01qb	8
So5	150.6	84.79	246	67	M	So5a	Ots08q	Co15a	Omy25q(Omy29)	Ssa09qb	8
					M	So5b	Ots14p	Co16b	Omy18p	Ssa16qb	7
So6	137.72	101.15	192	79	M	So6a	Ots29	Co11b	Omy15p	Ssa29	5
					M	So6b	Ots21	Co19b	Omy14q	Ssa05p	5
So7	133.68	155.15	235	80	M	So7a	Ots31	Co14b	Omy14p	Ssa14qb	5
					M	So7b	Ots16q	Co17b	Omy09q	Ssa15qb	3
So8	167.79	92.96	186	60	M	So8a	Ots15q	Co09b	Omy21q	Ssa07q	3
					M	So8b	Ots15p	Co09a	Omy21p	Ssa07p	4
So9[1]	78.12		168		A	So9a	Ots19	Co22	Omy02q	Ssa10qb	13
So9.5[1]	71.26		116		A	So9.5a	Ots10q	Co30	Omy08q	Ssa14qa	9
So10	169.69	65.33	263	82	M	So10a	Ots06q	Co04b	Omy01q	Ssa18qa	4
					M	So10b	Ots30	Co28	Omy10p	Ssa04q	7

Sockeye LG	Length (cM)		# markers		LG type	Sockeye LG arm	Chinook chromosome	Coho LG	Rainbow trout chromosome	Atlantic Salmon chromosome	Orthology support
	Female	Male	Female	Male							
So11	167.61	33.75	226	67	M	So11a	Ots13q	Co15b	Omy27	Ssa20qb	3
					M	So11b	Ots34	Co12b	Omy10q	Ssa08q	5
So12	139.01	84.12	232	82	M	So12a	Ots08p	Co14a	Omy25p	Ssa09qa	12
					M	So12b	Ots10p	Co16a	Omy09p	Ssa18qb	8
So13	143.45	117.44	237	123	M	So13a	Ots18	Co21	Omy04q	Ssa06p	6
					M	So13b	Ots12p	Co08a	Omy11p&q	Ssa20qa	9
So14	161.94	52.96	271	91	M	So14a	Ots23	Co13b	Omy02p	Ssa05q	1
					M	So14b	Ots14q	Co18a	Omy24	Ssa09qc	8
So15	169.46	88.09	293	97	M	So15a	Ots02p	Co01a	Omy17p	Ssa02q	6
					M	So15b	Ots02q	Co01b	Omy17q	Ssa12qb	3
So16	68.61	80.06	126	53	A	So16a	Ots25	Co25	Omy20p+q	SSa08p&Ssa28	7
So17[2]	138.95	60.53	173	61	M	So17a	Ots01p	Co10a	Omy04p	Ssa23	2
					M	So17b	Ots01p	Co10a	Omy04p	Ssa23	15
So18[3]	150.19	31.24	239	60	M	So18a	Ots11q	Co07b	Omy19q	Ssa01p	14
					M	So18b	Ots11p	Co07a	Omy19p	Ssa04p	5
So19	107.94	41.62	156	57	M	So19a	Ots33p	Co29	OmySex	Ssa11qa	9
					M	So19b	Ots33q	Co29	OmySex	Ssa11qa	3
So20	144.99	105.9	227	108	M	So20a	Ots22	Co24	Omy16q	Ssa13qa	7
					M	So20b	Ots28	Co27	Omy28	Ssa03p	9
So21	130.01	25.88	189	64	M	So21a	Ots32	Co20b	Omy13p	Ssa12qa	4
					M	So21b	Ots27	Co10b	Omy13q	Ssa06q	3
So22	152.37	69.22	248	82	M	So22a	Ots16p	Co18b	Omy11p	Ssa19qa	6
					M	So22b	Ots09p	Co06a	Omy12p	Ssa13qb	9
So23	152.67	90.55	222	75	M	So23a	Ots17	Co19a	Omy15q	Ssa17qa	3
					M	So23b	Ots24	Co20a	Omy16p	Ssa19qb	5
So24	129.68	181.11	221	96	M	So24a	Ots05q	Co13a	Omy05q	Ssa10qa	9
					M	So24b	Ots05p	Co12a	Omy08p	Ssa15qa	4

Sockeye LG	Length (cM)		# markers		LG type	Sockeye LG arm	Chinook chromosome	Coho LG	Rainbow trout chromosome	Atlantic Salmon chromosome	Orthology support
	Female	Male	Female	Male							
So25	84.5	65.86	145	63	A	So25a	Ots06p	Co04a	Omy01p	Ssa16qa	10
So26	159.93	98.08	170	51	A	So26a	Ots09q	Co06b	Omy12q	Ssa03q	4
So27	225.35	12.00	281	77	M	So27a	Ots04q	Co03b	Omy06q	Ssa26	4
					M	So27b	Ots04p	Co03a	Omy06p	Ssa24	13
So28	177.74	14.96	240	108	M	So28a	Ots12q	Co11a	Omy23	Ssa01qa	3
					M	So28b	Ots01q	Co08b	Omy26	Ssa11qb	8
Total	4,082	2,291	6,322	2,179							353

[1] LG So9 is the sex chromosome for sockeye salmon and was represented by two acrocentric LGs in the female (So9a, So9.5a) and a single metacentric LG in the male (So9). LG So9 in the male contained 107 markers and was 109.5 cM long. LG So9a is denoted as So9A_(X_2) in (Limborg *et al.* Submitted), and So9.5a is denoted as 9B_(X_1).

[2] Designated as acrocentric in Limborg *et al.* (Submitted).

[3] Assembled as separate LGs in Limborg *et al.* (Submitted)

Table 5.3. Homeologous LG arms in sockeye salmon, the number of marker pairs supporting each relationship, and corresponding homeologous relationships in other salmonids (Brieuc *et al.* 2014; Kodama *et al.* 2014).

Homeology in sockeye	# marker pairs	Homeology			
		Chinook	coho	rainbow trout	Atlantic salmon
So2a-So5b	9	Ots07p-Ots14p	Co05a-Co16b	Omy07p-Omy18p	Ssa17qa-Ssa16qb
So3b-So14a	19	Ots03p-Ots23	Co02a-Co13b	Omy03p-Omy02p	Ssa02p-Ssa05q
So8b-So23a	9	Ots15p-Ots17	Co09a-Co19a	Omy21p-Omy15q	Ssa07p-Ssa17qa
So10a-So28b	6	Ots06q-Ots01q	Co04b-Co11a	Omy01q-Omy23	Ssa18qa-Ssa01qa
So11b-So18b	11	Ots11p-Ots34	Co07a-Co12b	Omy19p-Omy10q	Ssa04p-Ssa08q
So15b-So21a	17	Ots02q-Ots32	Co01b-Co20b	Omy17q-Omy13p	Ssa02q-Ssa12qa
So21b-So26	10	Ots09q-Ots27	Co06b-Co10b	Omy12q-Omy13q	Ssa03q-Ssa06p
So27a-So28a	13	Ots04q-Ots12q	Co03b-Co08b	Omy06q-Omy26	Ssa26-Ssa11qa

Table 5.4. Mean and standard deviation (SD) of length, weight, and condition factor in experimental families. Bold values indicate significant differences between thermotolerant and thermosusceptible groups based on a Student's t-Test ($P < 0.05$). Distributions of length and condition factor were significantly different ($P < 0.05$) among the three families, whereas distributions of weight were not. Combined statistics for both groups are also given.

Phenotype	Family	Thermotolerant		Thermosusceptible		Combined	
		Mean	SD	Mean	SD	Mean	SD
Length (mm)	D3	35.46	2.10	34.80	3.92	35.12	3.16
Length (mm)	D4	**37.79**	**2.55**	**34.80**	**4.14**	36.24	3.76
Length (mm)	D5	**35.94**	**4.16**	**39.56**	**5.52**	37.75	5.19
Weight (g)	D3	**0.37**	**0.07**	**0.32**	**0.12**	0.34	0.10
Weight (g)	D4	**0.42**	**0.08**	**0.29**	**0.11**	0.35	0.11
Weight (g)	D5	**0.31**	**0.12**	**0.41**	**0.17**	0.36	0.15
Condition factor	D3	**0.82**	**0.11**	**0.72**	**0.19**	0.77	0.17
Condition factor	D4	**0.77**	**0.10**	**0.66**	**0.10**	0.71	0.11
Condition factor	D5	0.64	0.11	0.62	0.07	0.63	0.10

Table 5.5. Description of 26 significant QTL for four phenotypes in three diploid families. QTL peak is the marker with the highest LOD score for each QTL, cM is the position of the QTL peak, 95% CI is the approximate 95% confidence interval for the position of the QTL, PVE is the percentage of variation in the phenotype explained by the QTL, p(F) is the P-value of the F-statistic in the multiple-QTL model, and Sig denotes whether the QTL was significant at the LG or experiment-wide level (Exp). Gene abbreviations are provided for loci that were annotated directly (**bold**) or annotated based on flanking sequence from the Atlantic salmon genome (*italics*). See Table S5.3 for more information on QTL annotations. Marker RAG3 is the 5'-nuclease assay One_RAG-93. Instances where the QTL peak location is not included in the 95% CI indicate a lack of confidence in the true location and should be interpreted with caution.

Phenotype	Family	QTL Peak	LG	cM	95% CI	LOD	PVE	p(F)	Sig	Annotation(s)
Thermotolerance	D3	63844	2	171.57	0.00-192.07	3.35	16.75	5.10E-04	LG	**KLD7A**, *MLVCB*
Thermotolerance	D3	36045	11	129.89	126.7-133.03	3.32	9.88	9.50E-03	LG	
Thermotolerance	D4	70808	25	48	41.28-50.74	3.57	12.37	4.00E-03	LG	*PABP*
Thermotolerance	D5	7896	6	31.28	17.53-90.35	3.45	7.58	8.22E-03	LG	**SMHD3**
Thermotolerance	D5	87489	13	51.86	6.13-138.3	4.01	16.35	5.68E-05	LG	*PANTR*
Thermotolerance	D5	85651	19	86.91	1.71-93.68	3.29	7.89	6.85E-03	LG	
Length	D4	RAG3	9	54.19	37.36-57.78	3.07	7.56	1.80E-02	LG	**RAG**
Length	D4	76581	26	68.59	14.44-158.81	3.49	8.49	1.10E-02	LG	
Length	D4	767479	28	21.56	19.73-23.02	3.59	10.63	4.00E-03	LG	
Length	D5	44010	7	63.04	3.61-111.8	3.40	6.97	1.53E-03	LG	
Length	D5	84879	11	164.78	164.48-164.9	4.13	13.06	1.03E-05	LG	
Weight	D3	82443	11	117.29	0.83-167.61	3.41	11.41	1.00E-02	LG	
Weight	D4	70636	15	116.61	112.53-117.01	3.67	10.85	2.20E-03	LG	
Weight	D4	2639	25	78.49	54.16-78.92	3.09	6.18	2.70E-02	LG	
Weight	D4	76581	26	68.59	21.35-158.81	3.28	8.34	8.00E-03	LG	**SIA7B**
Weight	D4	767479	28	21.56	19.73-23.02	4.64	13.90	4.60E-04	Exp	
Weight	D5	7896	6	31.28	3.60-126.87	3.60	6.67	4.70E-03	LG	**SMHD3**
Weight	D5	66074	7	94.75	3.56-111.8	3.11	13.99	2.53E-05	LG	**BSN**
Weight	D5	84879	11	164.78	164.48-164.9	4.60	19.75	6.22E-07	Exp	
Weight	D5	30636	19	11.35	7.24-58.94	3.40	9.96	4.13E-04	LG	
Condition factor	D3	86442	6	53.86	50.26-58.87	5.56	34.08	1.33E-07	LG	
Condition factor	D4	765637	6	4.82	1.52-97.22	3.45	13.72	1.60E-03	LG	
Condition factor	D4	5975	28	145.33	3.70-68.97	3.47	6.92	3.50E-02	LG	**FAS**
Condition factor	D5	53011	4	64.01	139.87-145.86	3.16	15.16	5.61E-06	LG	
Condition factor	D5	15419	10	27.56	9.69-82.72	4.92	33.89	5.86E-11	LG	
Condition factor	D5	7448	20	0	0.00-144.99	3.49	9.74	2.93E-04	LG	

Supplementary table legends

Table S5.1. Information for each marker on the female and male linkage maps. Tag is the RAD tag, and marker name is the tag followed by a designation used to differentiate duplicated loci. The “Sockeye EST NCBI” column describes alignments to sockeye salmon ESTs in the NCBI database, the “Sequence P1 column” is the sequence from the P1 read for each RAD tag, and the “Sequence PE” is the sequence obtained from paired-end assemblies. Arm designations and cM position of markers do not correspond with Limborg *et al.* (Submitted), but marker names are the same.

Table S5.2. Coverage of Atlantic salmon genome scaffolds across the female sockeye salmon linkage map.

Table S5.3. Annotation information for the eight QTL that were successfully annotated. Direct annotations were based on RAD tag sequence, flanking sequence annotations were obtained using flanking sequence from the Atlantic salmon genome, and 5'-nuclease assay annotations were obtained from previously published 5’ nuclease assays.

Figures

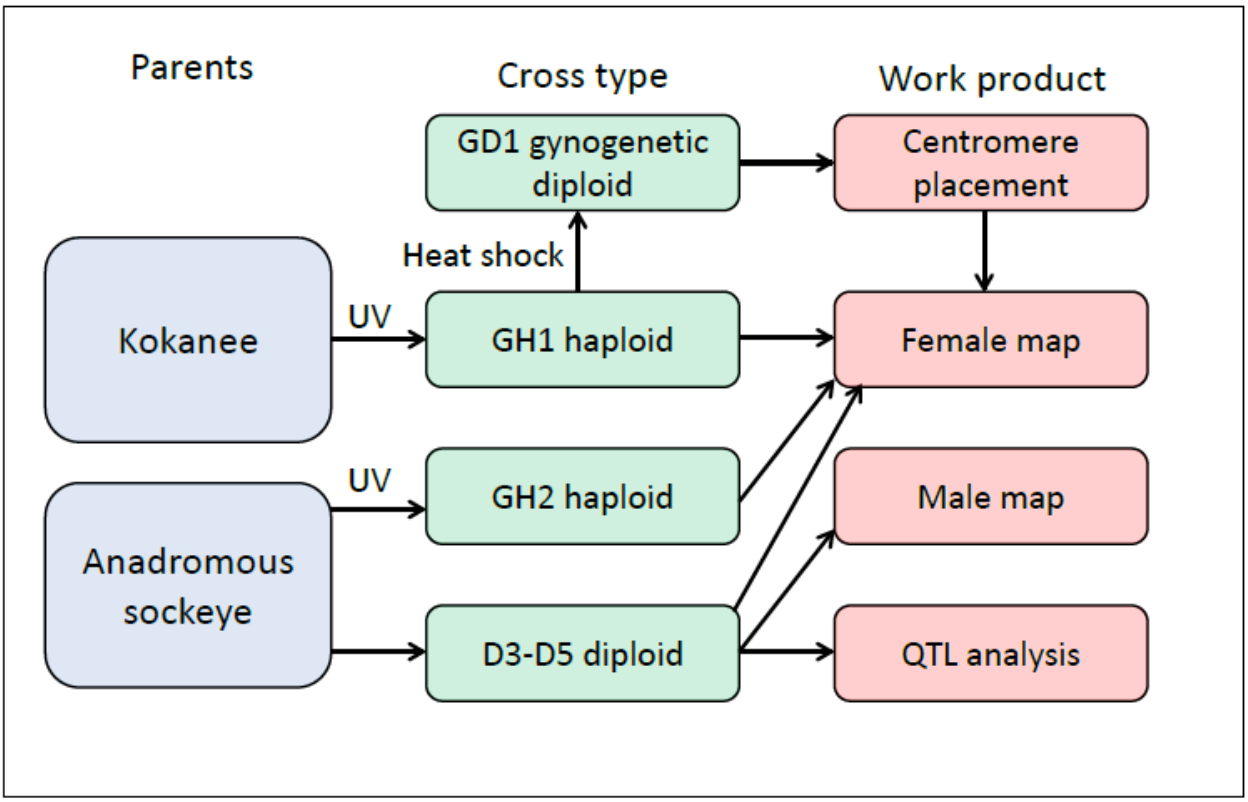

Fig. 5.1. Workflow for this study. The study included a haploid and gynogenetic diploid family of kokanee (landlocked sockeye salmon) sampled at the southern end of the species range and haploid and diploid families of anadromous sockeye salmon sampled at the northern end of the species range (GH2, D3-D5). Haploids (families GH1, GH2) were created by combining eggs and UV irradiated sperm, and gynogenetic diploids (family GD1) were created by heat shocking eggs and UV irradiated sperm after fertilization. Diploids (families D3-D5) were created by mating wild individuals from a single population. The female linkage map included data from all five families, and centromeres were placed on this map using knowledge of recombination events available from the gynogenetic diploids. The male map was constructed from the diploid families because haploid families do not include information about recombination events in males. QTL analysis for four phenotypic traits was conducted for each of the diploid families (D3-D5). Haploid families were not used for QTL analysis because haploid embryos do not survive past hatch. Additional information on each family including sample size can be found in Table 5.1.

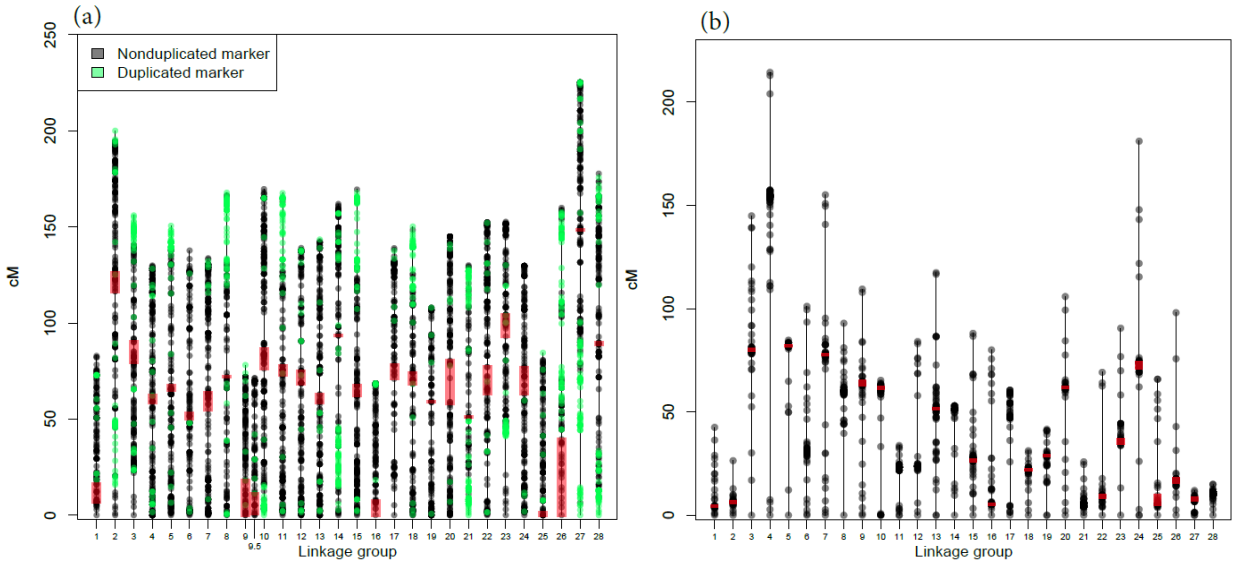

Fig. 5.2. (a) Female and (b) male linkage maps for sockeye salmon containing 6,322 and 2,179 loci respectively. Each dot represents a locus, and darker colors indicate higher marker density. Putative centromeres are denoted by rectangular boxes. Centromeres were successfully placed on all LGs in the female map and 19 of 28 LGs in the male map. LGs 9 and 9.5 are the sex chromosomes in sockeye salmon and are represented by two acrocentric LGs in the female map and a single metacentric LG in the male map (see text for additional information).

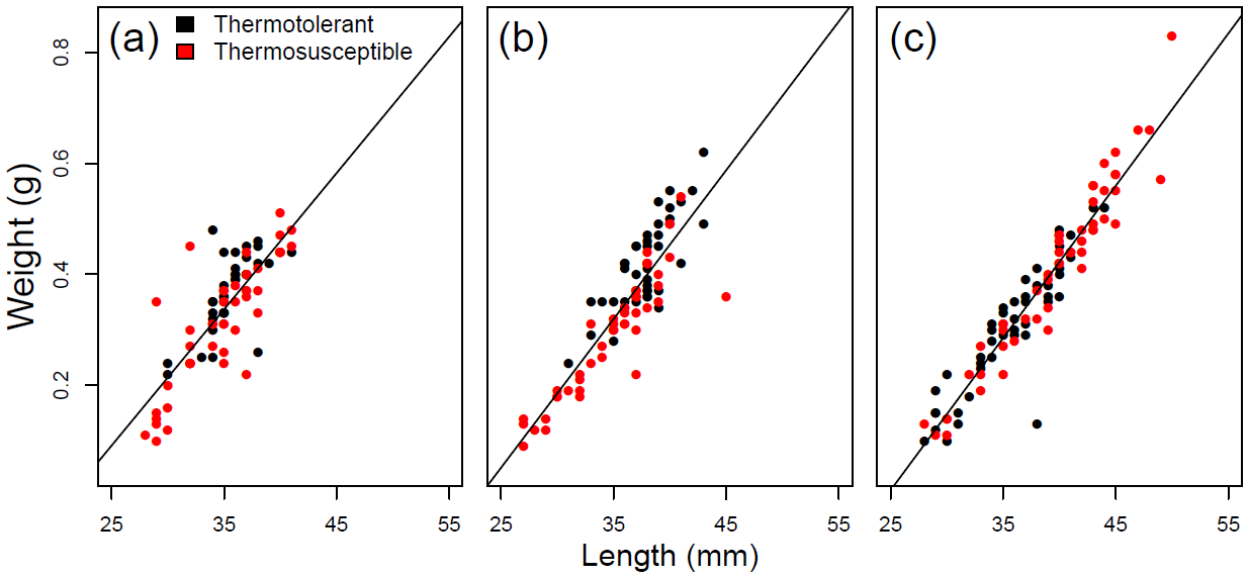

Fig. 5.3. Visualization of weight versus length relationships for each individual in family (a) D3, (b) D4, and (c) D5. Each dot represents an individual; dots are colored according to thermotolerance. A line of best fit is drawn through each distribution.

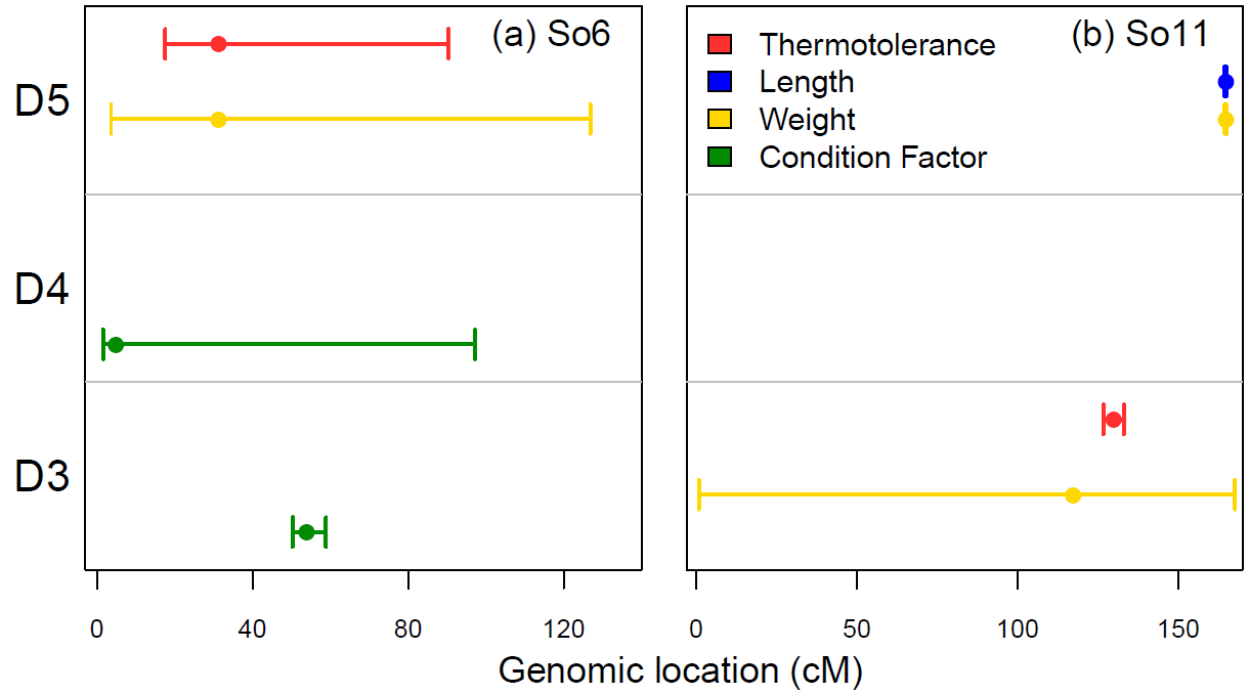

Fig. 5.4. Results from QTL analysis for the two LGs containing the most QTL, (a) LG So6 and (b) LG So11. Bracket lines represent 95% confidence intervals for the location of each QTL, and dots signify the QTL peak.

Supplementary figure legends

Fig. S5.1. Graphs of observed heterozygosity in gynogentic diploids vs map position for each LG in the female map. Each dot represents a locus, and loci with heterozygosity < 0.1 (denoted by the red line) were considered to be centromeric.

Fig. S5.2. Results from QTL analysis for all LGs, phenotypes, and families. Horizontal lines represent 95% confidence intervals for the location of each QTL, and dots signify the QTL peak. Gray vertical lines separate each LG and LGs are denoted at the top of the graph. Families (D3, D4, D5) and phenotypes are on the y-axis.

Chapter 6

Genomic islands of divergence are involved in parallel evolution of sockeye salmon ecotypes[6]

Abstract

Genomic islands of divergence (islands of divergence) are thought to play a significant role in the early stages of speciation and population differentiation. However, the frequency, size, and magnitude of islands of divergence in natural populations remains largely unknown. Here, we investigated genomic divergence in distinct ecotypes of sockeye salmon (*Oncorhynchus nerka*) from two drainages in Bristol Bay, Alaska. We found five islands of divergence that displayed high levels differentiation within a single drainage. We then screened loci from the three largest islands in a neighboring drainage and discovered strong signals of parallel evolution. We also conducted functional annotation and found that the peak of the largest divergence island contained a non-synonymous mutation in a putatively important gene involved in size and growth. Our results support a role of islands of divergence in the early stages of ecologically driven differentiation and represent a significant advance towards linking genotypes and phenotypes in non-model organisms.

Introduction

Understanding the genetic mechanisms that facilitate local adaptation is a central goal of evolutionary biology (Reeve & Sherman 1993). Studies that have successfully identified signals of adaptive divergence in wild populations have frequently focused on sympatric populations that are experiencing differential selective pressures but are not highly diverged at neutral markers (Andrew & Rieseberg 2013; Hemmer-Hansen *et al.* 2013; Soria-Carrasco *et al.* 2014). This divergence with gene flow scenario has been hypothesized to create genomic islands of divergence (islands of divergence) that display high levels of genetic differentiation, contrasting low levels of differentiation across the rest of the genome (Via & West 2008; Nosil *et al.* 2009). These islands of divergence can provide important insights into the genetic basis of local

[6] Full citation: Larson, W.A., M.T. Limborg, G.J. McKinney, T.H. Dann, J.E. Seeb, and L.W. Seeb. (In prep). Genomic islands of divergence linked to parallel evolution of sockeye salmon ecotypes. Supplementary tables and figures are available online from ProQuest LLC. Supplementary tables and figures are available online from ProQuest LLC (www.proquest.com) and legends for supplementary tables and figures are available at the end of this chapter.

adaptation (Hohenlohe *et al.* 2010; Nadeau *et al.* 2012; Yeaman 2013). However, the frequency, magnitude, and functional significance of islands of divergence remains poorly understood (Renaut *et al.* 2013; Cruickshank & Hahn 2014). Additionally, few studies have investigated the role of islands of divergence in parallel evolution of similar phenotypes (but see Pearse *et al.* 2014; Soria-Carrasco *et al.* 2014).

Salmon (*Oncorhynchus spp.*, *Salmo spp.*) represent an excellent model for investigating islands of divergence because they return to their natal streams with high fidelity, providing extensive opportunities for local adaptation on small spatial scales (Hendry *et al.* 2000; Quinn 2005; Pavey *et al.* 2010). However, this homing fidelity also promotes reproductive isolation, making it difficult to differentiate islands of divergence created by selection from neutral structure (Via & West 2008). Many studies investigating the genomic basis of local adaptation in salmon have concluded that signals of adaptive divergence are essentially randomly distributed across the genome and rarely co-locate in islands of divergence (e.g., Bourret *et al.* 2013; Hale *et al.* 2013; Larson *et al.* 2014c). It is important to note that most of these studies were conducted in systems with at least moderate levels of neutral structure ($F_{ST} >> 0.01$); in this case islands of divergence created by selection are likely to be difficult to distinguish from genomic regions that diverged due to neutral process (Via & West 2008; Cruickshank & Hahn 2014). These results suggest further research in systems with low neutral structure is necessary to gain a more complete understanding of the genomic basis of local adaption in salmon.

Here, we investigated the prevalence, size, and magnitude of islands of divergence in sockeye salmon (*Oncorhynchus nerka*) sampled across Bristol Bay, Alaska. Sockeye salmon from Bristol Bay comprise one of the most valuable fisheries in North America and produce remarkably consistent returns due to the high degree of population diversity found within this species and system (Hilborn *et al.* 2003; Wood *et al.* 2008; Schindler *et al.* 2010). A major component of this population diversity is associated with the habitat used for spawning and has resulted in the formation of distinct spawning ecotypes (Quinn *et al.* 1995; Lin *et al.* 2008a). For example, sockeye salmon that spawn in streams are much smaller than those that spawn on beaches or in rivers due to size selective predation by bears and the physical constraints of stream depth (Fig. 6.1, Quinn *et al.* 2001). The spawning environments of these ecotypes also vary in a number of other characteristics including temperature, gravel size, and spawning

density, leading to differences in egg morphology (Quinn *et al.* 1995), spawn timing (Schindler *et al.* 2010), and pathogen susceptibility (hypothesized in Larson *et al.* 2014a).

Distinct ecotypes often spawn less than one kilometer apart and generally show little, if any, neutral divergence within lakes (Lin *et al.* 2008a; Gomez-Uchida *et al.* 2011; McGlauflin *et al.* 2011; Larson *et al.* 2014a). However, Peterson *et al.* (2014) demonstrated that populations are locally adapted to their natal spawning habitats as individuals that dispersed from their natal beach habitats into a neighboring stream suffered reduced reproductive success compared to philopatric spawners. Additionally, high levels of genetic divergence among ecotypes has been observed in the genes of the major histocompatibility complex (MHC), indicating that adaptive genetic differences among ecotypes do exist (Gomez-Uchida *et al.* 2011; McGlauflin *et al.* 2011; Larson *et al.* 2014a). Distinct spawning ecotypes in Bristol Bay clearly experience vastly different selective pressures during the short spawning and early rearing phases of their life history but are hardly diverged at neutral markers; this makes Bristol Bay an ideal system to investigate the hypothesis that islands of divergence are involved in the earliest phases of ecologically-driven differentiation.

We investigated the genetic basis of ecotypic divergence in sockeye salmon from two drainages in Bristol Bay, the Wood River basin and Lake Iliamna. The Wood River basin is a series of five interconnected lakes that encompasses a drainage area of 3,590 km^2, and Lake Iliamna is a large lake with a drainage basin of approximately 20,000 km^2. Sockeye salmon from our study systems are composed of five distinct ecotypes associated with the habitat used for spawning: island beaches (Iliamna), mainland beaches (both drainages), rivers (Wood), streams (Wood), and tributaries (Iliamna) (Blair *et al.* 1993; Quinn *et al.* 2001; Hilborn *et al.* 2003; Stewart *et al.* 2003). A well-supported model of sockeye salmon recolonization suggests these ecotypes evolved independently within each drainage (recurrent evolution hypothesis, reviewed in Wood *et al.* 2008). This independent colonization history facilitates comparisons between the drainages to determine whether the same genes or genomic regions are involved in the recurrent evolution of distinct ecotypes. Additionally, the presence of the mainland beach ecotype (referred to hereafter as beach) in both drainages allowed us to investigate whether parallel evolution of this ecotype occurred through similar genetic mechanisms.

The primary goals of this study were to: (1) quantify the frequency and magnitude of islands of divergence linked to ecotypic differentiation in the Wood River basin, (2) investigate

the functional significance of genes in these islands, and (3) determine whether the same islands of divergence were also involved in ecotypic differentiation in an adjacent drainage, Lake Iliamna. Our study is one of the first to describe the functional significance of islands of divergence and provides novel evidence that the same islands of divergence may play an important role in the recurrent formation of similar phenotypes through parallel evolution.

Results

Sample collection and genotyping

We obtained tissue samples from five ecotypes of sockeye salmon (14 populations total) from the Wood River basin and Lake Iliamna (Table 6.1, Table S6.1, Fig. 6.1, Fig S6.1). RAD sequencing on the six populations from the Wood River basin initially produced genotypes at 16,528 putative SNPs. After removing SNPs with low minor allele frequencies (< 0.05), SNPs whose genotype frequencies deviated from Hardy-Weinberg equilibrium, and all but one SNP per tag, 6,254 SNPs remained (see Table S6.2 for summary statistics). Seven SNPs showing high levels of divergence in these populations were screened in populations from Lake Iliamna. Details on the genotyping results for Lake Iliamna populations are found below.

Demography, diversity, and selection

We estimated relatedness between individuals, effective population size (N_e), and heterozygosity to explore patterns of demography and genetic diversity within the Wood River basin. We found 18 individuals that were part of eight full sibling groups (Table S6.3). Most of the related individuals were found in the smaller stream populations. Estimates of N_e ranged from about 250 for the streams populations to infinite for the Agulowak River and were close to 1,000 for the other populations (Table 6.1). The ratio of effective to census size was highly variable, likely due to differences in metapopulations dynamics among ecotypes (Table 6.1, Lin *et al.* 2008a; Larson *et al.* 2014a). Observed and expected heterozygosities (H_O and H_E) calculated from all loci were indistinguishable among populations (H_O = 0.32, H_E = 0.33, Table 6.1).

Tests for loci displaying putative signals of divergent selection (outlier loci) were conducted with BayeScan v2.1 (Foll & Gaggiotti 2008) and revealed 37 loci that were candidates for selection with a false discovery rate of 0.01 (Fig. S6.2, Table S6.4). Loci that were not candidates for divergent selection were categorized as neutral for further analyses.

Islands of divergence

We obtained the genomic location of loci genotyped in our study from a linkage map constructed in Larson *et al.* (Submitted, Chapter 5) to investigate islands of divergence. Loci were classified as belonging to an island of divergence if they were putatively under selection according to an outlier test (outlier locus) and were found within 10 centimorgans (cM) of another outlier locus. We did not consider pairs of outlier loci that were within 10 cM of each other but statistically linked in every population as islands because these loci likely represent the same signal of population differentiation. Islands of divergence are named according to the linkage group (LG) they were found on, and islands found on the same LG are differentiated based on their location. For example, island LG13_1 is the first island on LG 13 and island LG13_2 is the second island.

We discovered five islands of divergence containing between two and six loci that contrasted with the low neutral structure found throughout the rest of the genome (Fig. 6.2; Table 6.2, Fig. S6.3). These islands of divergence contained 11 of the 12 loci showing the highest level of genetic differentiation in this study (Table S6.2). The width of islands of divergence ranged from 0 to 8.29 cM, and the average F_{ST} of loci in these islands ranged from 0.10 to 0.40 (Table 6.2). Significant linkage disequilibrium (LD) in more than half of populations was found for a number of locus pairs in islands of divergence, an expected result given the proximity of loci in these island (Table S6.5). However, no locus pairs in different islands were statistically linked in more than two populations suggesting long distance LD between islands of divergence is unlikely. Both MHC loci showed high levels of differentiation, similar to SNPs in islands of divergence, but were not part of an island of divergence identified here.

Alignments to Atlantic salmon genome scaffolds (ICSASB_v1 ; GenBank accession: GCA_000233375.3) were possible for 13 of the 15 loci found in islands of divergence, and we were able to align multiple loci from a single island to the same scaffold for three islands (Table 6.3, Table S4). Of these three islands, two were composed of pairs of loci that were approximately 100 kilobases (kb) apart, but the large island on LG 13 (LG13_1) contained six loci that spanned a 402 kb region (Table 6.2, Fig. 6.3). In order to further investigate the shape and boundaries of this large island, we aligned all available loci to the scaffold containing the island. We were able to align eight loci to this scaffold, six loci that were found within the island and two loci with low levels of differentiation that flanked the island (Fig. 6.3). The island

demonstrated a peak like shape with loci on the margins displaying lower genetic differentiation than loci in the center.

Population structure in the Wood River basin

We investigated patterns of population structure in the Wood River basin with three datasets: (1) 6,217 putatively neutral loci, (2) 22 outlier loci outside islands of divergence, and (3) 15 outlier loci found within islands of divergence. Overall F_{ST} was much lower for the neutral dataset compared to the datasets containing outlier loci, and the F_{ST} for outliers found outside of islands was lower than for outliers found within islands (neutral F_{ST} = 0.009, outliers outside islands F_{ST} = 0.088, outliers within islands F_{ST} = 0.207). Patterns of population structure visualized with principal coordinate analyses (PCoAs) were highly variable across datasets (Fig. 6.2, Fig. S6.4, Table S6.6). In general, populations did not cluster by ecotype based on the neutral dataset or the outliers outside of islands dataset (Fig. 6.2b, Fig. S6.4), but clustered tightly by ecotype based on the dataset comprised of outliers within islands (Fig. 6.2c). Results from an analysis of molecular variance (AMOVA) with populations grouped by ecotype also demonstrated a similar pattern. Specifically, the dataset containing loci within islands of divergence displayed a much higher level of variation partitioned among ecotypes than both the neutral dataset and the outliers outside of islands dataset (Table 6.4). H_O was not significantly different among ecotypes for any of the three datasets (Table S6.7).

Functional annotation and gene expression

We conducted functional annotations by aligning sequences from our loci as well as sequence from the Atlantic salmon genome flanking our loci to the Swiss-Prot database. No outlier loci were directly annotated based on RAD data (Table S6.2), but we were able to find annotations within 50 kb of 18 outlier loci and annotations within 200 kb of 21 outlier loci (Table 6.3, Table S6.4, Table S6.8). Common annotations flanking outlier loci included transcription factors, genes involved in cell adhesion, and cell membrane components. We also used transcriptome data from Everett *et al.* (2011) to investigate gene expression and found that 11 outlier loci belonged to genes that are putatively expressed in adult sockeye salmon from the Wood River basin (Table 6.3, Table S6.4, Fig S6.5).

Annotations for loci in islands of divergence included genes involved in transcription, cell adhesion, and cell membrane structure (Table 6.3). Additionally, seven of the 15 loci in islands of divergence are putatively expressed in adult sockeye salmon and may represent direct

targets of selection (Fig. S6.5). However most of the islands of divergence in this study only contained two or three loci, making it difficult to characterize the functional significance of these islands.

One exception was the divergence island LG13_1, which contained six loci spanning a 402 kb region. We annotated the full region and investigated expression levels at all annotated genes in order to characterize this island as thoroughly as possible (Fig. 6.3, Fig. S6.6). We found eight genes in this island, and six of these genes were putatively expressed in adult sockeye salmon according to transcriptome data from (Everett *et al.* 2011) (Fig. 6.3, Table S6.4). Additionally, we discovered that the two loci at the peak of the island (locus 1400 and 2462) are found within the *TULP4* gene, which is involved in transcription and appears to be expressed in sockeye salmon from our study system (Fig. 6.3). Comparison of our data with the ORF in rainbow trout suggested that locus 24362 represents a non-synonymous mutation with one variant coding for the amino acid glutamine and the other variant coding for histidine (Fig. 6.3). These amino acids differ in structure as well as side chain charge. The other polymorphism found in the *TULP4* gene, locus 1400, codes for a synonymous mutation.

Signals of parallel evolution

We designed seven 5'-nuclease assays from RAD loci found within islands of divergence (Table 6.3) and screened these loci in populations from Lake Iliamna (Fig. 6.4a, Fig. S6.1) to test for signals of parallel evolution between drainages. We also obtained genotypes from 87 putatively neutral SNPs (Elfstrom *et al.* 2006; Dann *et al.* 2012; Storer *et al.* 2012; Larson *et al.* 2014a) to investigate neutral population structure. None of the seven loci developed in this study displayed deviations from Hardy-Weinberg equilibrium in more than two of the eight Lake Iliamna populations, and patterns of linkage disequilibrium at these loci were similar between Lake Iliamna and Wood River populations (summary statistics and population information in tables S6.1 and S6.9.

Outlier tests conducted in *Bayescan* revealed that six of the seven loci developed from islands of divergence were putatively under selection in both the Wood River basin and Lake Iliamna (Fig. 6.4 b,c). Loci from the MHC were also putatively under selection in either one or both drainages. The remaining 87 putatively neutral SNPs did not display signatures of selection in either drainage.

In order to investigate patterns of population structure at neutral and putatively adaptive loci we conducted PCoAs using (1) 87 neutral SNPs and (2) seven putatively adaptive SNPs in islands of divergence. PCoAs based on these two datasets produced strikingly different patterns of population structure (Fig. 6.4 d,e, Table S6.10). The PCoA constructed from the neutral panel revealed strong differentiation between the two drainages and relatively little differentiation within the drainages (Fig. 6.4 d). Contrastingly, the PCoA based on adaptive loci from islands of divergence grouped beach populations from the Wood River and Lake Iliamna into a single cluster (Fig. 6.4 e). It is important to note that we did not include loci from the MHC in either PCoA because MHC differentiation is influenced by fine-scale differences in pathogen communities which may obscure signals of parallel evolution (Miller *et al.* 2001; Larson *et al.* 2014a).

Tests for differences in H_O among ecotypes across all populations revealed that H_O was similar at the 87 neutral SNPs for all ecotypes ($P > 0.05$). However, H_O was significantly lower in the beach ecotype compared to the other ecotypes at the seven SNPs developed from islands of divergence ($P < 0.01$, Table S6.7, Fig S6.7). Average H_O for the beach ecotype at the SNPs developed from islands of divergence was similar to the neutral SNPs (0.26 vs 0.29), while average H_O across the other ecotypes was higher (0.37 vs 0.29). None of these loci displayed significant deviations from Hardy-Weinberg equilibrium in more than half of the populations tested.

Discussion

Patterns of population differentiation across the genome

We discovered five islands of divergence that displayed high levels of differentiation and strong signals of adaptive divergence in sockeye salmon from the Wood River basin. Differentiation at loci in these islands was generally partitioned among ecotypes, contrasting patterns from neutral markers that were found throughout the rest of the genome (cf., Lemay & Russello 2015). For example, stream populations were highly diverged from beach and river populations and from each other in a PCoA constructed with neutral markers but clustered together with markers found within islands of divergence. These results suggest that genetic drift due to relatively low effective population sizes in stream populations is driving patterns of neutral differentiation, while patterns of differentiation in islands of divergence are influenced by adaptive processes.

Frequency and size of islands of divergence

Much uncertainty exists regarding the frequency and size of islands of divergence in natural populations (Nosil & Feder 2012). Some studies have found a low frequency of large islands that span over half a chromosome (Andrew & Rieseberg 2013; Hemmer-Hansen et al. 2013), but most studies document much smaller islands that range in frequency from less than ten islands across the genome (Hohenlohe et al. 2010), to over 7,000 (Soria-Carrasco et al. 2014). It is important to note that the number of loci genotyped and methods used to identify islands differ significantly across studies and that denser sequencing tends to reveal more and smaller islands (Nosil & Feder 2012; Soria-Carrasco et al. 2014). However, studies using different methodologies in the same study organisms have generally shown similar trends (e.g., Colosimo et al. 2004; Hohenlohe et al. 2010), indicating that results are likely to be fairly robust to methodological differences.

We found five highly differentiated but relatively small islands of divergence that ranged in size from 80 to 402 kb (0 – 8.29 cM). The frequency and magnitude of islands of divergence has been hypothesized to be a reflection of a number of processes including strength of selection and levels of gene flow. For example, strong selection and high levels of gene flow are hypothesized to create fewer high-magnitude islands, whereas lower levels of selection and low gene flow are hypothesized to create numerous islands showing relatively lower levels of increased differentiation (Via & West 2008; Nosil et al. 2009; Nosil & Feder 2012). The high levels of selection and gene flow in our study system (demonstrated by Lin *et al.* 2008a; Peterson *et al.* 2014) coupled with the small number and high divergence of islands that we observed provides support for this hypothesis. It is also important to note that all five islands of divergence we discovered occurred near the ends of linkage groups (subtelomeric regions), indicating that these regions may be especially important for adaptive divergence (c.f., Brown *et al.* 2010; Ellegren *et al.* 2012; Gagnaire *et al.* 2013a).

Many genomic mechanisms have been hypothesized to explain the presence of islands of divergence including (1) divergence hitchhiking (Via 2012), (2) chromosomal inversions (Kirkpatrick & Barton 2006), (3) co-location of genes involved in adaptation (Yeaman & Whitlock 2011), (4) reduced recombination rates (Renaut et al. 2013), and (5) reduced diversity (Cruickshank & Hahn 2014). Our study was not specifically designed to test these hypotheses, but we can provide some evidence that certain mechanisms are more plausible than others. The

hypothesis that is most strongly supported by our data is divergence hitchhiking. We observed moderate but not high levels of linkage disequilibrium within islands of divergence, and our islands spanned larger regions than would be expected if physical linkage was the only force maintaining them (Via 2012). These observations provide support for the hypothesis that divergent selection has reduced gene exchange over moderately sized genomic regions.

Support for the other hypotheses presented above is much weaker. For example, it is unlikely that chromosomal inversions were responsible for the formation of the islands of divergence we observed because no inversions were documented by Larson *et al.* (Submitted, Chapter 5), who constructed linkage maps using beach and stream populations from the same system. Additionally, the linkage map did not show reduced recombination in islands of divergence (data not shown) indicating that variation in recombination rates is unlikely to explain the presence of the islands we discovered. We did observe some evidence of reduced diversity in islands of divergence. However, this reduction in diversity was only observed in beach populations, indicating that selective sweeps rather than genomic regions characterized by low diversity likely explain this pattern (cf., Hemmer-Hansen *et al.* 2013). Unfortunately, we were unable to investigate the hypothesis that genes involved in adaptation are more prevalent in islands of divergence because this analysis would require a well annotated genome. Despite this limitation, we did find multiple genes in islands that are putatively involved in adaptive divergence.

Functional significance of islands of divergence

The functional significance of genes found within islands of divergence remains largely unexplored (Nosil & Feder 2012). Soria-Carrasco *et al.* (2014) found that genes in islands of divergence were involved in a variety of processes that were putatively important for adaptation including metabolism, signal transduction, and metal ion binding. Our annotation results are similar, with many loci in islands of divergence annotating to genes that are putatively involved in adaptive divergence such as transcription factors and genes involved in cell signaling.

The most compelling functional annotations in our study were found in the large island on LG13. We were able to annotate eight genes in this island, and two loci at the peak of the island annotated to the coding region of a single gene, *TULP4*. *TULP4* is a transcription factor in the tubby protein family, a set of genes that have been repeatedly linked to obesity and fat storage in mice (*Mus musculus*, reviewed in Mukhopadhyay & Jackson 2011) and may be

involved in height determination in humans (*Homo sapiens*; Allen *et al.* 2010; van Duyvenvoorde *et al.* 2014). We were able to align the two loci that annotated to *TULP4* to the open reading frame for an exon of this gene and found that one locus (locus 24362) encodes a non-synonymous mutation that changes glutamine to histidine. These two amino acids differ significantly in structure and have been linked to functional changes in multiple enzymes (e.g., Schroder & Schroder 1992; Gaume *et al.* 1995). Additionally, the *TULP4* gene appears to be expressed in adult sockeye salmon from our study system, indicating that the non-synonymous mutation we discovered may have a functional effect on adult sockeye salmon during the spawning phase, one of their most important life stages (Quinn 2005). We also found two other genes in the LG13_1 island with putative functional significance and possible connections to size and growth, *ESR1* and *MSGN1*.

Taken together, the functional annotation results from the LG13_1 island provide evidence that genes involved in size and growth could be important in promoting adaptive divergence between ecotypes of sockeye salmon. These results are unsurprising given the high degree of phenotypic differentiation among ecotypes, especially in depth and overall size (Quinn *et al.* 2001). However, we cannot definitively conclude that the genes we annotated are the true targets of selection without directly investigating the functional significance of these transcripts in sockeye salmon.

Genetic signals of parallel evolution between drainages

We found strong evidence for parallel evolution in sockeye salmon from the Wood River basin and Lake Iliamna suggesting that similar genetic mechanisms are likely involved in life-history diversification across these drainages. The probability of gene reuse during parallel evolution is relatively high (30-50%, Conte *et al.* 2012); however, the majority of studies in salmonids have not found strong evidence to support this hypothesis (e.g., Frazer & Russello 2013; Gagnaire et al. 2013b; Hale et al. 2013; Brieuc et al. 2015), but see Pearse *et al.* (2014). The contrasting results between our study and previous research in salmonids suggest that colonization history and shared ancestry may influence the frequency of gene reuse during parallel evolution.

The studies mentioned above investigated parallel evolution between highly divergent lineages (Frazer & Russello 2013; Hale et al. 2013; Brieuc et al. 2015) or in systems with complicated colonization histories including multiple lineages (Gagnaire et al. 2013b). The

drainages in our study system, on the other hand, were likely colonized by genetically similar populations of sea/river-type sockeye salmon in concordance with the recurrent evolution hypothesis (Wood et al. 2008). The recurrent evolution hypothesis postulates that sea/river-type sockeye salmon, which are characterized by high stray rates and low levels of population structure, have repeatedly colonized lake environments and given rise to distinct ecotypes of lake-type sockeye salmon. Evidence for this hypothesis has been found in a drainage close to our study system (the Kuskokwim River, McPhee et al. 2009), as well many other areas across the species range (reviewed in Wood et al. 2008).

According to a recent review by Conte et al. (2012), parallel evolution through similar genetic mechanisms occurs more frequently when populations share similar ancestry, as in our study system. This pattern of parallel evolution from similar standing genetic variation may explain the increased levels of genetic parallelism observed in our study compared to previous research in salmonids. It is important to note that we cannot completely rule out a colonization scenario where ecotypes evolved in sympatry prior to colonization (e.g., McKeown et al. 2010), however results from neutral loci clustered populations within drainages indicating that this hypothesis is highly unlikely.

Our results suggest that islands of divergence are involved in the parallel evolution of similar phenotypes across independently colonized systems. In fact, all three of the islands of divergence that we investigated in the Wood River basin and Lake Iliamna contained adaptively important loci in both systems. These results indicate that islands of divergence may play a larger role in parallel evolution than previously thought (e.g., Gagnaire et al. 2013b; Soria-Carrasco et al. 2014). However, it is still unclear whether genetic signals of parallel evolution are more likely to occur in islands of divergence compared to the rest of the genome, advocating for future research on this topic.

Conclusions

In conclusion**,** we demonstrated that islands of divergence containing putatively important genes are involved in the parallel evolution of distinct spawning ecotypes of sockeye salmon. Many studies in salmonids have suggested that large islands of divergence are relatively uncommon and are generally not involved in parallel evolution. Our results are not concordant with these previous studies and illustrate that islands of divergence can be important during adaptive differentiation in salmonids. Additionally, we investigated the functional significance

of islands of divergence and discovered that these islands harbor genes that are likely to be adaptively important, something that has not been previously described in salmonids. This study provides some of the first evidence that genomic islands of divergence are an important component of adaptive differentiation and parallel evolution in salmonids and represents a significant advance towards linking genotypes to phenotypes in this important non-model organism.

Materials and methods

Sample collection, RAD genotyping, and summary statistics

We obtained tissue samples from 14 populations of sockeye salmon, six populations from the Wood River basin and eight populations from Lake Iliamna (Fig. 6.1, Fig. 6.4a, Fig. S6.1). The six populations from the Wood River basin were composed of two replicate beach, river, and stream populations, each from a different lake (Table 6.1, Fig. 6.1, see Larson *et al.* 2014a for more information on ecotypes). RAD sequencing, SNP genotyping, and SNP filtering was conducted in these populations following the methods of Larson *et al.* (2014c) and Limborg *et al.* (Submitted) (see supplementary methods for more information). Summary statistics including F_{ST} (Weir & Cockerham 1984) and F_{IS} were calculated for each locus in *Genepop version 4* (Rousset 2008).

Samples from Lake Iliamna were composed of two island beach populations, two mainland beach (beach) populations, and four tributary populations (Table S6.1, Fig. S6.1), and were available from Alaska Department of Fish and Game (ADFG) archives. Ecotypes were classified according to Gomez-Uchida *et al.* (2011). These populations were genotyped with 87 neutral SNPs, two SNPs from the MHC, and seven SNPs designed from RAD data that showed high levels of divergence in populations from the Wood River basin (see below).

Demography, diversity, and selection in the Wood River basin

We estimated relationships between individuals, N_e, and heterozygosity to explore patterns of demography and genetic diversity within the Wood River basin. Relationship analysis was conducted with the program *ML-RELATE* (Kalinowski *et al.* 2006) using the default parameters and all loci in the RAD dataset. The most likely relationship for each pair of individuals was retained. Estimates of N_e for each population were obtained with the linkage disequilibrium method implemented in NeEstimator (Do *et al.* 2014). These estimates were then corrected for the bias introduced by physical linkage by removing comparisons between loci that

were found on the same LG (Larson *et al.* 2014c). Loci that were candidates for natural selection (see below) were also removed from this analysis. Locus and population-specific estimates of H_O and H_E were calculated in ARLEQUIN 3.5 (Excoffier & Lischer 2010).

Tests for loci displaying putative signals of divergent selection (outlier loci) were conducted with BayeScan v2.1 (Foll & Gaggiotti 2008) using the default parameters and a conservative false discovery rate of 0.01. Loci that were candidates for divergent selection were categorized as outliers in further analyses, and loci that were not candidates for selection were classified as putatively neutral.

Islands of divergence

We investigated the distribution of outlier loci across a linkage map for sockeye salmon (Larson *et al.* Submitted) to identify islands of divergence displaying high levels of differentiation among populations (see supplementary methods). Outlier loci were classified as belonging to a divergence island if they were found within 10 cM of another outlier. It is important to note that we did not define islands of divergence using a weighted average or kernel smoothing method (e.g., Gagnaire *et al.* 2013b; Larson *et al.* 2014c; Brieuc *et al.* 2015) because single positions on the map used here can contain dozens of loci that are over a megabase (Mb) apart (Larson *et al.* Submitted). Additionally, loci that are only a few kb away from each other can be separated by multiple cMs on the map due to differences in segregation patterns among loci (Larson *et al.* Submitted). We therefore decided to leverage the linkage map to loosely define islands of divergence and investigate each island separately using scaffolds from the Atlantic salmon genome (see supplementary methods).

Tests for linkage disequilibrium (LD) among outlier loci were conducted in order to investigate levels of LD within islands of divergence and patterns of long range LD. Tests were conducted in *Genepop* with the default parameters and a significance level of 0.05. Locus pairs displaying significant linkage disequilibrium in all six populations were considered highly linked and were not classified as belonging to a genomic island of divergence unless other proximate outliers were found.

Population structure in the Wood River basin

Genetic differentiation among populations in the Wood River basin was investigated separately for three sets of loci: (1) all neutral loci; (2) outlier loci that were found outside of islands of divergence; and (3) outlier loci found within genomic islands. First, we estimated

overall and pairwise-F_{ST} values for each dataset in *Genepop*. We then visualized patterns of population structure based on pairwise-F_{ST} with a principal coordinate analysis (PCoA) and tested the significance of each population comparison with a test for genic differentiation conducted in *Genepop* (alpha = 0.01). We also conducted an analysis of molecular variance (AMOVA) for each set of loci in Arlequin 3.5. Finally, we conducted permutation tests in FSTAT (Goudet 1995) to test for differences in H_O among populations (alpha = 0.05). This analysis did not include neutral loci because no variation in H_O was present among populations with this marker set. Populations were grouped by ecotype for the AMOVA and permutation tests.

Functional annotation and gene expression

Functional annotation was conducted by aligning consensus sequences for each locus to the Swiss-Prot database. Additional annotations were attempted for outlier loci that were placed on the Atlantic salmon genome by aligning 200 kb of 3' and 5' flanking sequence for each locus to the Swiss-Prot database. We also used transcriptome data from Everett *et al.* (2011) to investigate whether genes of interest that co-located with outlier loci are expressed in adult sockeye salmon from the Wood River basin. See supplementary methods and results for more information on functional annotations and gene expression.

Signals of parallel evolution

We developed seven high-throughput 5'-nuclease assays from RAD loci found within islands of divergence in the Wood River basin and screened these loci in eight populations from Lake Iliamna to investigate signals of parallel evolution between drainages. We also obtained genotypes from 87 putatively neutral SNPs for both systems to investigate neutral population structure, and obtained genotypes from two SNPs in the MHC to compare patterns of adaptive divergence at the MHC with patterns from RAD loci. See supplementary methods and results for more information on assay development and genotyping.

We jointly analyzed data from SNPs common to both the Wood River basin and Lake Iliamna to test for signals of parallel evolution between drainages. First, we calculated summary statistics for each locus including allele frequencies, overall and pairwise-F_{ST}, H_O, H_E, and F_{IS} in *Genepop* and tested for significant differences in H_O among ecotypes in FSTAT. We also conducted tests for deviations from Hardy-Weinberg and linkage disequilibrium for the seven loci developed in this study in *Genepop* (see supplementary methods). We then conducted

separate outlier tests for each drainage in *Bayescan* using the parameters described above. A SNP was considered neutral if it was not an outlier in either drainage and was considered adaptive if it was an outlier in either drainage. Pairwise-F_{ST} values were calculated separately for the datasets containing (1) neutral SNPs and (2) putatively adaptive SNPs, and PCoAs were constructed based on these values. Loci from the MHC were not included in the adaptive panel even though they were found to be under selection because patterns of differentiation at MHC loci have been show to reflect extremely local processes and may obscure signals of parallel evolution (Miller *et al.* 2001; Larson *et al.* 2014a). Finally, we estimated the N_e of populations in Lake Iliamna using genotypes from neutral, unlinked loci and the program NeEstimator.

Supplementary methods

RAD sequencing, SNP discovery, and genotyping

RAD libraries were prepared for 24 males and 24 females from each sample population from the Wood River basin with the restriction enzyme *SbfI* following the methods of Baird *et al.* (2008) and Everett *et al.* (2012). Sequencing was conducted on an Illumina HiSeq2000 (single-end, 100 bp target (SE100)), and 48 individuals were pooled in each sequencing lane.

Identification and genotyping of SNPs from RAD data was conducted using the *STACKS* software package (version 1.20, Catchen *et al.* 2011; Catchen *et al.* 2013) following the methods of Limborg *et al.* (Submitted). Our analysis pipeline consisted of quality filtering and demultiplexing raw sequences using *process_radtags*, identifying SNPs within individuals using *ustacks*, creating a catalog of loci with *cstacks*, and exporting and classifying individual genotypes with *sstacks* and *populations*. The locus catalog used to genotype individuals in this study was constructed by adding the two individuals from each sample population with the most data to the SE100 catalog described in Larson *et al.* (Submitted). We also used the *rxstacks* module to correct genotype calls based on population information, a step that was not included in Limborg *et al.* (Submitted).

Putative SNPs discovered with *STACKS* were filtered using methods similar to Larson *et al.* (2014c) to remove redundant data, possible sequencing errors, loci containing null alleles, and uninformative polymorphisms. SNPs were excluded from the dataset if they were genotyped in less than 80% of individuals, had a minor allele frequency less than 0.05 in all sample populations, or were found to deviate significantly from Hardy-Weinberg equilibrium in more than half of the study populations (alpha = 0.05). Tests for deviations from Hardy-Weinberg

equilibrium were conducted in *Genepop* version 4 (Rousset 2008). If a RAD tag contained more than one SNP, the SNP with the highest minor allele frequency was retained. As a final filtration step, we removed individuals that were genotyped in less than 80% of the SNPs that passed the filters discussed above.

Paired-end assembly, placement of loci on a linkage map, and alignment to the Atlantic salmon genome

We conducted paired-end assemblies for each locus to increase query length for functional annotation and alignment to genomic resources. Paired-end sequence (80 x 2 bp target (PE80)) from six individuals sampled in Wood River basin was available from Everett *et al.* (2012). These PE80 sequences were assembled with SE100 reads from this study using the alignment program *CAP3* (150 bp minimum alignment length, Huang & Madan 1999) following the methods of Etter *et al.* (2011) and Waples *et al.* (2015).

We used the linkage map for sockeye salmon described in Larson *et al.* (Submitted) to infer the genomic location (linkage group (LG) and position within LG) of loci genotyped in this study. This map also contains additional information including alignments to the Atlantic salmon genome. No alignment step was necessary because the locus names were identical between studies.

Outlier loci that did not align to the linkage map or Atlantic salmon genome based on the results of Larson *et al.* (Submitted) were aligned to the genome using BLASTN and the following parameters: > 85% identity, alignment length > 150 bp. These alignments were conducted with PE data when possible, and the best alignment for each locus was retained. Outlier loci that aligned to the genome but were not placed on the linkage map were added to the map based on scaffold-specific relationships between recombination and physical distance.

Functional annotation and gene expression

Functional annotation was conducted by aligning consensus sequences for each locus to the Swiss-Prot database using BLASTX. The alignment with the lowest e-value < 10^{-4} for each locus was accepted as the annotation. Additional annotations were attempted for outlier loci that were placed on the Atlantic salmon genome by aligning 200 kb of 3' and 5' flanking sequence for each locus to the Swiss-Prot database using BLASTX and the following parameters: alignment length > 100 bp, < 30 mismatches per 100 bp, < 3 gaps per 100 bp. If multiple alignments met these criteria for a given scaffold position, the best alignment was retained. We

also aligned all loci to the expressed sequence tags (ESTs) for sockeye salmon in the cGrasp database (http://web.uvic.ca/grasp/) using BLASTN (parameters: >90% identity, < 4 mismatches per 100 bp, < 1 gap per 100 bp, and alignment length > 50% of query sequence). If multiple alignments met these parameters for a single locus, the alignment with the lowest e-value was retained.

We used transcriptome data from Everett *et al.* (2011) to investigate whether genes of interest that co-located with outlier loci are expressed in adult sockeye salmon from the Wood River basin. Everett *et al.* (2011) obtained SOLiD data (50 bp reads) from the transcriptomes of six adult sockeye salmon from the Wood River basin. We aligned those sequences to scaffolds from the Atlantic salmon genome containing outlier loci with the program *Bowtie* V0.12.9 (Langmead *et al.* 2009, -v 3). We then constructed histograms of read counts across each scaffold to investigate patterns of gene expression. We classified regions that aligned to multiple SOLiD sequences as "expressed" (likely containing expressed genes), and regions where few or no sequences aligned as "unexpressed" (unlikely to contain any expressed genes).

Development and genotyping of high-throughput assays

5'-nuclease assays from divergence islands were developed with TaqMan chemistry (Life Technologies, Grand Island, New York) following the methods of Larson *et al.* (2014b). Assays were screened on 96 individuals from the Wood River basin to confirm high concordance between RAD and 5'-nuclease genotypes. We then genotyped these assays in eight populations from Lake Iliamna (48 individuals per population, sample populations described above and in Fig. S6.2 and Table S6.2).

Genotypes for 87 neutral SNPs and two MHC SNPs were available for populations from both drainages from Dann *et al.* (2012) and Larson *et al.* (2014a). These SNPs are a subset of the 96-SNP panel described in Elfstrom *et al.* (2006) and Storer *et al.* (2012). We removed seven SNPs from the panel that were either in linkage disequilibrium with another SNP according to Dann *et al.* (2012) (*One_GPDH2-187, One_Tf_ex11-750*), produced inconsistent separation among genotypes (*One_SUMO1-6, One_U1016-115*), or were designed from mtDNA and were not comparable to diploid SNPs (*One_CO1, One_Cytb_17, One_Cytb_26*).

SNPs developed from RAD data were not necessarily genotyped in the same individuals or collections as SNPs from the 96-panel (Table 6.1). In order to facilitate further analyses with these SNPs, we combined data from the two panels to form composite individuals. These

composite individuals were not used for any individual based analyses. Combining genotypes from individuals samples in different years should not bias our estimates of population structure because Dann *et al.* (2012) showed that structure in the Bristol Bay region is generally very temporally stable.

Supplementary results

Sequencing, SNP discovery, and genotyping

RAD sequence data was obtained from 286 individuals, and the average number of reads per individual was 1.7 million. We excluded two individuals from the Anvil Beach population prior to analysis because they produced < 10,000 reads (see Table 6.1 for sample sizes). SNP genotyping with *STACKS* revealed 16,538 putative SNPs genotyped in at least 80% of individuals across the dataset, and 6,254 SNPs were retained after filtering. No individuals were genotyped in < 80% of the filtered SNPs.

Paired-end assembly and placement of loci on a linkage map

Consensus sequences longer than 150 bp were assembled for 6,121 of the 6,254 loci in the dataset (98%), and the average length of each sequence was 259 bp (Table S6.2). We were able to place 3,536 (57%) of the loci in our study on the linkage map from Larson *et al.* (Submitted, Table S6.1).

Functional annotation

Functional annotation from RAD sequence data was successful for 613 of 6,255 loci (10%, Table S6.2). Transposable elements comprised approximately 21% of these annotations. Other common functional groups included DNA polymerases and membrane proteins. Alignment to sockeye salmon ESTs from the cGRASP database was possible for 99 loci, but none of these were outliers (Table S6.2). We were able to classify 11 outlier loci as belonging to a gene that is putatively expressed in adult sockeye salmon from the Wood River basin (i.e. belonging to an exon, Table 6.3, Fig S6.5).

Development and genotyping of high-throughput assays

We designed seven 5'-nuclease assays to explore signals of parallel evolution between populations from the Wood River basin and Lake Iliamna. These assays were developed from loci in the three most prolific divergence islands: island LG13_1 (five assays), island LG13_2 (one assay), and island LG12_1 (one assay). Assays were initially screened on 96 individuals from the Wood River basin that had also been RAD sequenced, and the concordance rate

between RAD and 5’-nuclease genotypes was 99.5%. Assays were then screened in 383 individuals from Lake Iliamna, and we were able to genotype at least five of the seven loci for all individuals.

Tables

Table 6.1. Collection data and summary statistics for six populations of sockeye salmon from the Wood River basin in southwestern Alaska. N is the number of individuals that were successfully RAD sequenced and genotyped with 5'-nuclease assays, N_e is the effective population size of each population (95% confidence intervals in parentheses), and census is the average census size of each population from 2009-2013 (Alaska Salmon Program, unpublished data). The first year denotes collections that were RAD sequenced; the second year denotes collections that were genotyped with 5'-nuclease assays. The number of individuals that were part of full sibling groups in each population is also given.

Population	Ecotype	Lake	N	Year	N_e	Census	No. full siblings	H_O	H_E
Anvil Beach	Beach	Nerka	46	2011, 2006	780 (730-838)	13,700	2	0.32	0.33
Yako Beach	Beach	Aleknagik	48	2011, 2006	1,426 (1,276-1,615)	2,900	2	0.32	0.33
Little Togiak River	River	Nerka	48	2011, 2008	1,161 (1,058-1,285)	8,000	2	0.32	0.33
Agulowak River	River	Aleknagik	48	2013, 2001	Inf (18,384-Inf)	115,520	0	0.32	0.33
Teal Creek	Stream	Nerka	48	2013, 2013	231 (226-236)	2,865	6	0.32	0.33
Hansen Creek	Stream	Aleknagik	48	2013, 2004	298 (290-306)	4,687	6	0.32	0.33

Table 6.2. Summary of islands of divergence. Islands are named according to the linkage group they were found on, and the number after the "_" is used to differentiate islands found on the same linkage groups. The "# SNPs" column denotes the number of outlier SNPs in each island. The physical size of each island was only calculated if all SNPs in the island could be mapped to the genome. See Table 6.3 for additional information about loci found in islands.

Island	# SNPs	Average F_{ST}	LG	cM start	cM end	size (cM)	size (Kb)
LG7_1	3	0.13	7	9.79	12.91	3.12	NA
LG12_1	2	0.40	12	12.72	18.34	5.62	NA
LG12_2	2	0.10	12	113.02	119.43	6.41	150.61
LG13_1	6	0.22	13	0.00	8.29	8.29	402.47
LG13_2	2	0.14	13	19.58	19.58	0.00	79.78

Table 6.3. Description of 15 outlier loci found within islands of divergence. The "Assay" column indicates whether a 5'-nuclease assay was designed for the locus, the "Expressed" column indicates whether a locus was putatively expressed in adult sockeye salmon from the Wood River basin (see text), and the "annotations" column denotes any genes that were found within 50 kb of the outlier locus (see table S8 for additional annotation information). Bold values in the "cM" column indicate that map positions were inferred based on genome alignments.

Tag	F_{ST}	LG	cM	Assay	Genome scaffold	Expressed	Annotation	Function
97866	0.14	7	9.79	no	NA	NA	NA	NA
39512	0.11	7	12.91	no	jcf1001126827_0-0	no	*CPSF1*	mRNA binding
24805	0.14	7	12.91	no	jcf1000486106_0-0	no	*FA46A*	regulation of gene expression
27165	0.38	12	**12.72**	yes	jcf1000217495_0-0	yes	NA	NA
41305	0.42	12	18.34	no	NA	NA	NA	NA
41402	0.08	12	113.02	no	jcf1001006045_0-0	yes	*PTPRS*, *PTPRD*	protein tyrosine phosphatase activity
74619	0.11	12	119.43	no	jcf1001006045_0-0	yes	*PTPRF*	cell adhesion
72071	0.14	13	**0.00**	yes	jcf1000459108_0-0	yes	*CF211*, *ESR1*	DNA binding,
90464	0.21	13	**3.08**	yes	jcf1000459108_0-0	no	NA	NA
24362	0.30	13	5.52	yes	jcf1000459108_0-0	yes	*TULP4*	regulation of transcription
1400	0.29	13	5.52	no	jcf1000459108_0-0	yes	*TULP4*	regulation of transcription
56448	0.20	13	7.28	yes	jcf1000459108_0-0	no	*MSGN1*	chromatin binding
18507	0.23	13	8.29	yes	jcf1000459108_0-0	no	*SMC6*, *TMM18*	telomere maintenance, membrane component
82702	0.18	13	19.58	yes	jcf1000599557_0-0	no	*PTPRK*	cell adhesion
65253	0.10	13	19.58	no	jcf1000599557_0-0	yes	*PTPRU*	cell adhesion

Table 6.4. Results from three AMOVA using different sets of loci. Populations are grouped by ecotype. SSQ is sum of squares. Italicized values indicate significance ($P < 0.05$). SSQ, sum of squares.

Source of variation	d.f.	SSQ	Var	% of variation
6,217 neutral loci				
Among ecotypes	2	4301.78	2.96	0.3
Among populations within ecotypes	3	4758.00	6.53	*0.67*
Within populations	566	545298.28	963.42	*99.02*
22 outlier loci outside islands				
Among ecotypes	2	109.58	0.18	4.94
Among populations within ecotypes	3	59.87	0.17	*4.72*
Within populations	566	1889.31	3.34	*90.34*
15 outlier loci within islands				
Among ecotypes	2	272.44	0.62	19.55
Among populations within ecotypes	3	48.17	0.14	*4.43*
Within populations	566	1387.60	2.45	*76.03*

Supplementary table legends

Table S6.1. Collection data and summary statistics for eight populations of sockeye salmon from Lake Iliamna in southwestern Alaska. N is the number of individuals that were successfully genotyped with 5'-nuclease assays, year is year sampled, and N_e is the effective population size of each population (95% confidence intervals in parentheses).

Table S6.2. Summary statistics for 6,257 loci genotyped in the Wood River basin. The "LG" and "cM" columns denote map location, the "in island" column denotes loci that are part of islands of divergence, the "num pops out of HW" column denotes the number of population that deviated from Hardy-Weinberg equilibrium ($P< 0.05$), columns ending in "AF" denote population allele frequencies, the "Sockeye EST NCBI" column describes alignments to sockeye salmon ESTs in the NCBI database, the "Sequence P1 column" is the sequence from the P1 read for each RAD tag, and the "Sequence PE" is the sequence obtained from paired-end assemblies.

Population abbreviations are Anvil Beach (AnvilB), Yako Beach (YakoB), Little Togiak River (LTogR), Agulowak River (AgulR), Teal Creek (TealCk), and Hansen Creek (HansCk).

Table S6.3. Relatedness coefficients for pairwise comparisons of full siblings found in the Wood River basin as inferred from the program ML-relate.

Table S6.4. Description of 37 outlier loci discovered in the Wood River basin. The "in island" column denotes loci that are part of islands of divergence, the "possibly expressed" column indicates whether a locus appeared to be expressed in adult sockeye salmon from the Wood River basin (see text), and the "annotations within 50 kB" column denotes any genes that were found within 50 kb of the outlier locus (see table S8 for additional annotation information). Bold values in the "cM" column indicate that map positions for those loci were inferred based on genome alignments. No outlier loci were directly annotated using only RAD sequence.

Table S6.5. The number of populations that were out of linkage disequilibrium for each pair of outlier loci genotyped in the Wood River basin ($P< 0.05$). Locus names are found on the x and y axes and are colored coded by island (rows 2-16, cols B-P). Loci colored with light green outside of this range were outliers but were not found in islands.

Table S6.6. Pairwise F_{ST} values for each population pair from the Wood River basin calculated with three marker sets: (1) 6,217 neutral loci, (2) 22 outlier loci located outside islands of divergence, and (3) 15 outlier loci located within islands of divergence. Genetic differentiation between all population pairs was highly significant for all three marker sets ($P << 0.01$). Population abbreviations are identical to those in Table S2.

Table S6.7. Observed heterozygosity (H_O) for each populations and marker set included in this project. Abbreviations for Wood River populations are identical to those in Table S2. Abbreviations for populations from Lake Iliamna are Triangle Island (TriaIs), Woody Island (WoodIs), Finger Beach (FingB), Knutson Beach (KnutB), Iliamna River (IliaR), Copper River (CoppR), Dream Creek (DreamCk), Lower Talarik Creek (LTalCk).

Table S6.8. BLAST results for genes found within 200 kb of outlier loci.

Table S6.9. Summary statistics for 96 SNPs genotyped in Wood River basin and Lake Iliamna populations. Columns ending in "AF" denote population allele frequencies. Abbreviations for Wood River populations are identical to those in Table S2 and abbreviations for Iliamna populations are found in Table S7. SNPs with the prefix "One_RAD" were developed from RAD loci, and numbers following the prefix correspond to RAD tag IDs from Table S2.

Table S6.10. Pairwise F_{ST} values for populations from the Wood River basin and Lake Iliamna genotyped with (a) 87 putatively neutral SNPs and (b) seven SNPs in genomic islands. Values in bold were significantly differentiated ($P < 0.05$). Population abbreviations for the Wood River basin populations are identical to those in Table S2, and abbreviations for Iliamna populations are found in Table S7.

Figures

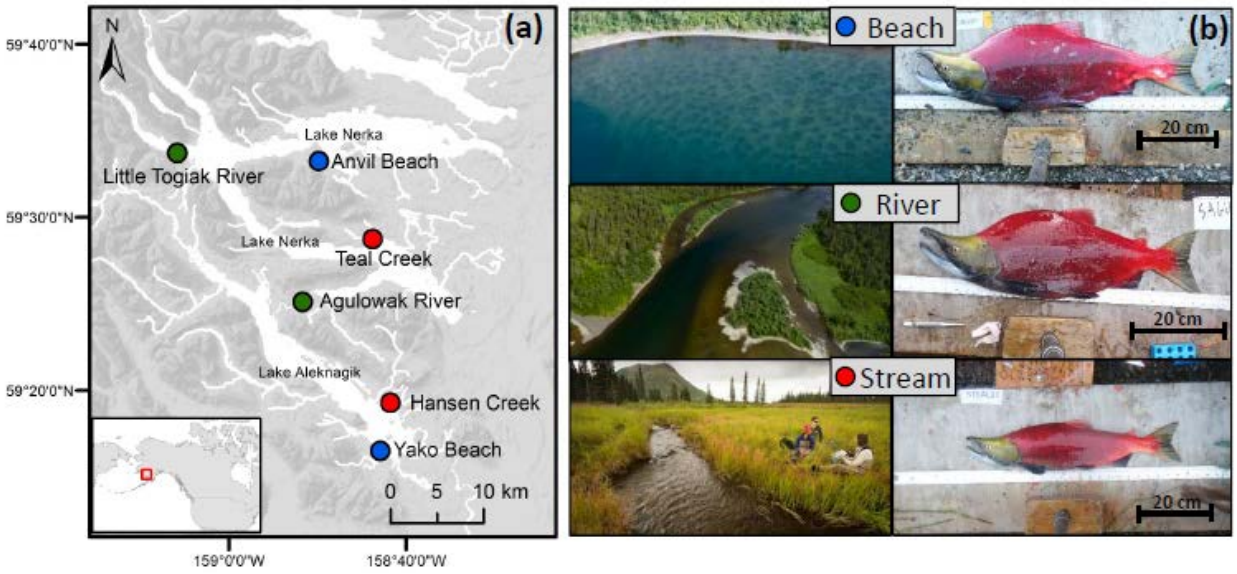

Fig. 6.1. Study system and characterization of ecotypes. (a) Map of Wood River basin. The six sample populations are color coded by ecotype, and colors correspond to those in panel b. See Table 6.1 for more information on sample populations. (b) Photos of typical spawning habitat and representative males from each ecotype. Photos of habitats courtesy of J. Armstrong and J. Ching.

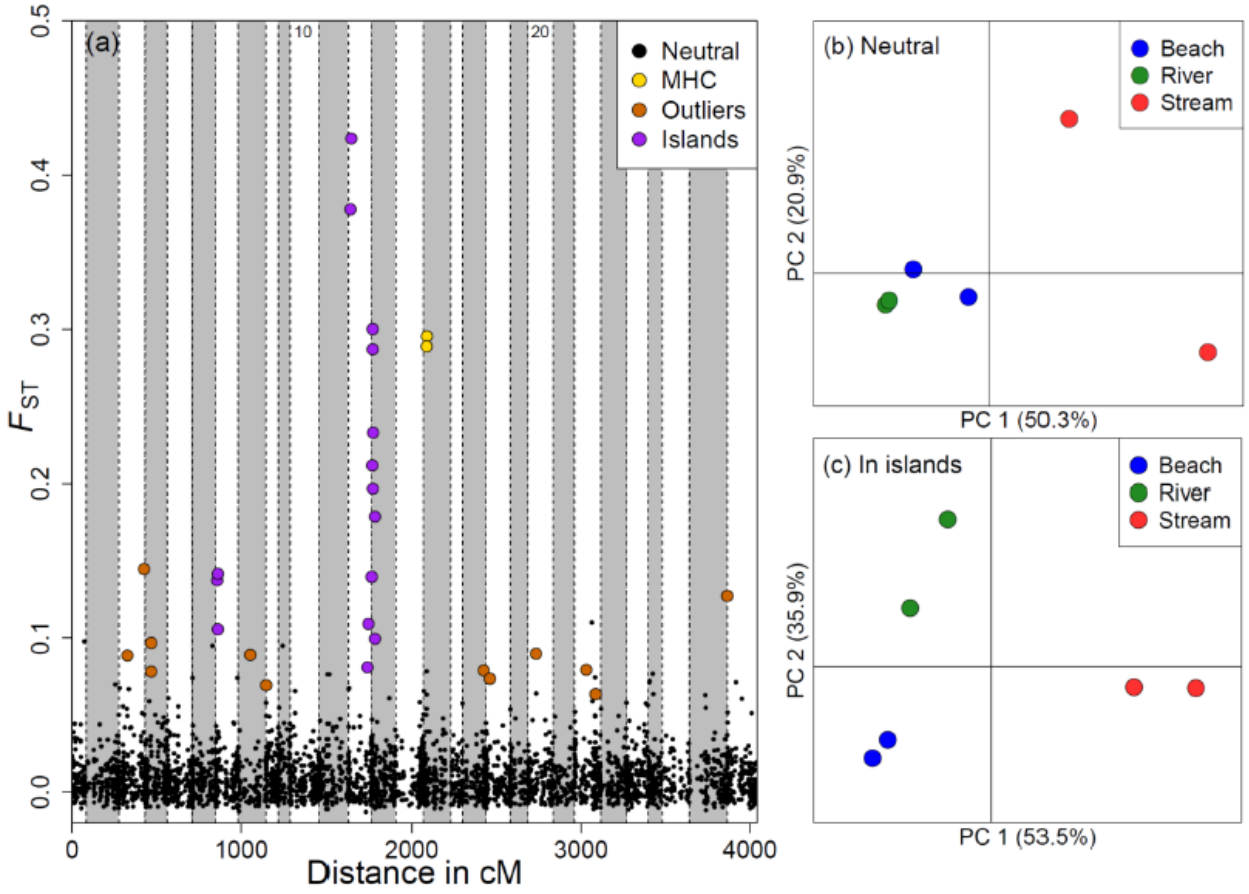

Fig. 6.2. Genomic differentiation and population structure. (a) Genetic differentiation (F_{ST}) across the genome for 3,537 loci that were placed on the genetic linkage map. Linkage groups (LGs) are separated by dashed lines, and LG 10 and 20 are denoted at the top of the figure. LG 9 is split into two LGs (see Larson *et al.* Submitted, Chapter 5). The "outliers" designation indicates outlier loci that are found outside of islands of divergence, and the "islands" designation indicates loci that are found within islands of divergence. Two loci from the MHC are plotted for comparison but were not included in any other within Wood River analyses. (b,c) Principal coordinate analysis (PCoA) of population differentiation based on pairwise-F_{ST} for six sample populations from the Wood River basin. The PCoAs were constructed using (a) 6,217 neutral loci, and (b) 15 outlier loci located within islands of divergence.

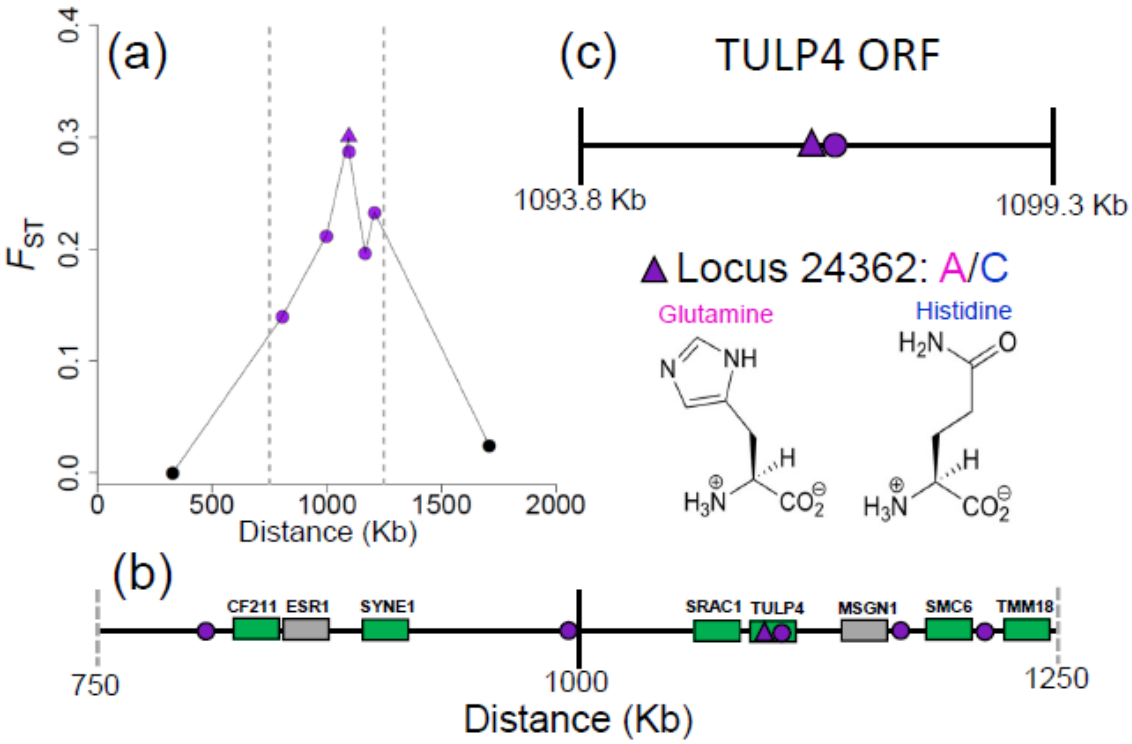

Fig. 6.3. Dissection of the largest genomic island in this study (LG13_1). (a) Magnified view of genetic differentiation for the LG13_1 island. The boundaries of the island are denoted with dashed vertical lines. Purple dots are loci within the island, and black dots are loci that aligned to the same scaffold but did not show high levels of divergence and were considered outside of the island. The triangle represents locus 24362 throughout this figure. (b) Conceptual representation of the genes found within the genomic island of divergence described above. Loci are represented with purple dots, and green and gray boxes indicate the locations of genes (boxes not to scale). Gene abbreviations are located above each box. Green boxes denote genes that appear to be expressed in adult sockeye salmon from the Wood River basin, and gray boxes denote genes that are unlikely to be expressed. (c) Visualization of the opening reading frame (ORF) for the *TULP4* gene. Locus 24362 (purple triangle) codes for a non-synonymous mutation in this gene. The A allele codes for glutamine and the C allele codes for histidine.

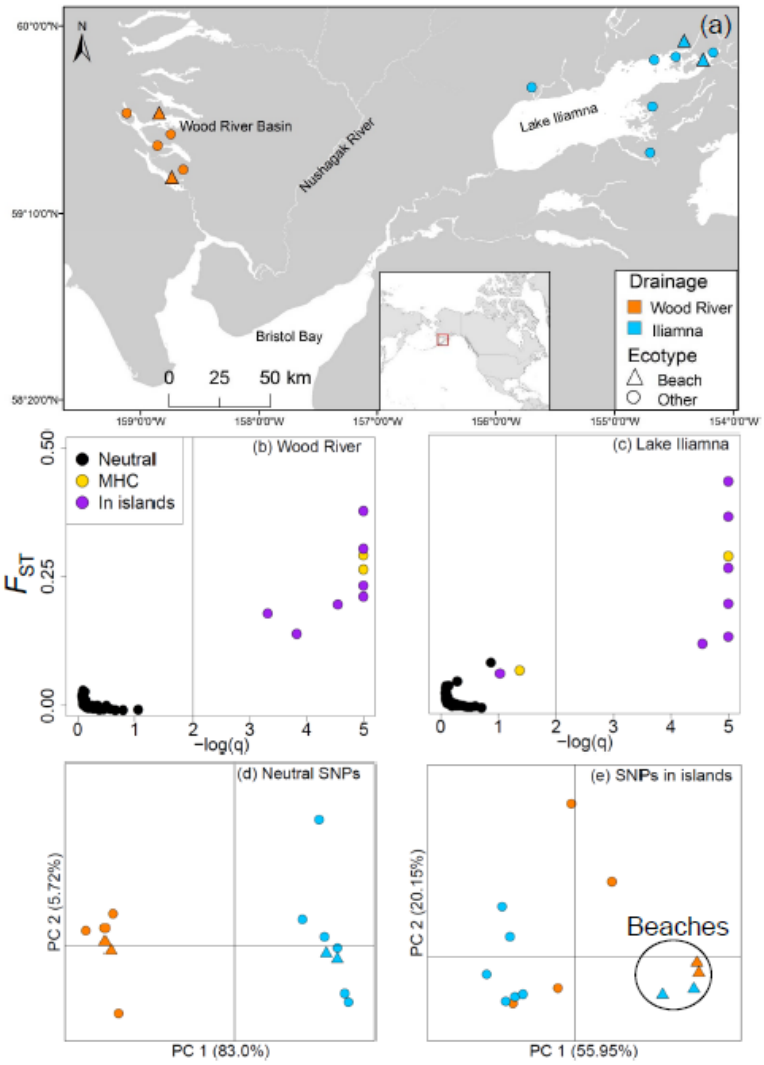

Fig. 6.4. Signatures of parallel selection between the Wood River basin and Lake Iliamna. (a) Map of northwestern Bristol Bay denoting sampling sites from both study systems. Sites are characterized as beach ecotype or other (one of the other four ecotypes). (b) Results from an outlier test for populations from the Wood River. The test was conducted in *BayeScan* using 87 putatively neutral SNPs, two SNPs in the MHC that have been shown to be under selection (Gomez-Uchida *et al.* 2011; McGlauflin *et al.* 2011), and seven SNPs found in islands of divergence. (c) Results from an outlier test for populations from Lake Iliamna. The loci used were identical to (b). (d) Principal coordinate analysis (PCoA) of population differentiation based on pairwise-F_{ST} from 87 putatively neutral SNPs. (e) PCoA based on seven SNPs developed from RAD data that were found in islands of divergence. Beach ecotypes are circled for emphasis.

Supplementary table legends

Fig. S6.1. Map of sample populations collected from Lake Iliamna. Populations are color coded by ecotype.

Fig. S6.2. Results from a test for outlier test conducted in *BayeScan*. Each dot represents a locus, and loci to the right of the black vertical line are candidates for divergent selection with an FDR threshold of 0.01. Outlier loci are colored based on whether they are located within or outside a genomic island of divergence. F_{ST} was calculated using the Weir and Cockerham method (Weir & Cockerham 1984).

Fig. S6.3. Genetic differentiation for five islands of divergence on three LGs: (a) LG 7, (b) LG 12, and (c) LG 13. Each dot represents a locus and loci that are part of islands of divergence are denoted with different colors. Linkage groups 7, 12, and 13 are metacentric with centromeres between 50 and 90 cM.

Fig. S6.4. Principal coordinate analysis (PCoA) of population differentiation for Wood River basin populations. The PCoA was constructed using pairwise-F_{ST} values calculated from 22 outlier loci identified with *Bayescan* that were found outside islands of divergence. Populations are color coded by ecotype.

Fig. S6.5. Histograms of the number of SOLiD transcriptome reads from Everett *et al.* (2011) that aligned to genome scaffolds containing loci that were candidates for divergent selection. The locations of RAD loci are denoted with red vertical lines. Only reads from adult sockeye salmon sampled in the Wood River basin were used.

Fig. S6.6. Histograms of the number of SOLiD transcriptome reads from Everett *et al.* (2011) that aligned to genes found within the genomic island on LG13 described in Fig. 3. The starting alignment location for each gene is denoted with a red vertical line. Only reads from adult sockeye salmon sampled in the Wood River basin were used.

Fig. S6.7. Boxplots of observed heterozygosity (H_O) by ecotype for 87 neutral SNPs and 7 SNPs in islands of divergence.

References

Abdul-Aziz OI, Mantua NJ, Myers KW (2011) Potential climate change impacts on thermal habitats of Pacific salmon (*Oncorhynchus* spp.) in the North Pacific Ocean and adjacent seas. *Canadian Journal of Fisheries and Aquatic Sciences* **68**, 1660-1680.

Ackerman MW, Habicht C, Seeb LW (2011) Single-nucleotide polymorphisms (SNPs) under diversifying selection provide increased accuracy and precision in mixed-stock analyses of sockeye salmon from the Copper River, Alaska. *Transactions of the American Fisheries Society* **140**, 865-881.

Ackerman MW, Templin WD, Seeb JE, Seeb LW (2013) Landscape heterogeneity and local adaptation define the spatial genetic structure of Pacific salmon in a pristine environment. *Conservation Genetics* **14**, 483-498.

ADF&G (2013) Chinook salmon stock assessment and research plan [online]. Alaska Department of Fish and Game, Special Publication No. 13-01. Available from http://www.adfg.alaska.gov/fedaidpdfs/sp13-01.pdf [accessed 7 March 2013].

Albertson RC, Powder KE, Hu YA, Coyle KP, Roberts RB, Parsons KJ (2014) Genetic basis of continuous variation in the levels and modular inheritance of pigmentation in cichlid fishes. *Molecular Ecology* **23**, 5135-5150.

Alcaide M, Edwards SV, Negro JJ, Serrano D, Tella JL (2008) Extensive polymorphism and geographical variation at a positively selected MHC class IIB gene of the lesser kestrel (*Falco naumanni*). *Molecular Ecology* **17**, 2652-2665.

Allen HL, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838.

Allendorf F, Thorgaard GH (1984) Polyploidy and the evolution of salmonid fishes *In The Evolutionary Genetics of Fishes, Edited by B. J. Turner. Plenum Press, New York*, 1-53.

Allendorf FW, Bassham S, Cresko W, Limborg MT, Seeb LW, Seeb JE (2015) Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *Journal of Heredity* **106**, 217-227.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**, 697-709.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Amish SJ, Hohenlohe PA, Painter S, Leary RF, Muhlfeld C, Allendorf FW, *et al.* (2012) RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. *Molecular Ecology Resources* **12**, 653-660.

Anderson EC (2010) Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources* **10**, 701-710.

Anderson EC, Waples RS, Kalinowski ST (2008) An improved method for predicting the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences* **65**, 1475-1486.

Andrew RL, Rieseberg LH (2013) Divergence is focused on few genomic regions early in speciation: incipent speciation of sunflower ecotypes. *Evolution* **67**, 2468-2482.

Angeloni F, Wagemaker N, Vergeer P, Ouborg J (2012) Genomic toolboxes for conservation biologists. *Evolutionary Applications* **5**, 130-143.

Armstrong JB, Schindler DE, Omori KL, Ruff CP, Quinn TP (2010) Thermal heterogeneity mediates the effects of pulsed subsidies across a landscape. *Ecology* **91**, 1445-1454.

Avise JC (2010) Perspective: conservation genetics enters the genomics era. *Conservation Genetics* **11**, 665-669.

Bagenal TB, Tesch FW (1978) Age and growth. In: T.B. Bagenal, (ed) Methods for assessment of fish production in freshwater, 3rd edition. Blackwell Scientific Publication, Oxford, UK. Pages 101–136.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376.

Baker TT, Fair LF, Clark RA, Hasbrouck JJ (2006) Review of salmon escapement goals in Bristol Bay, Alaska, 2006. Alaska Department of Fish and Game, Fishery Manuscript No. 06-05. Available from http://www.sf.adfg.state.ak.us/FedAidPDFs/fms06-05.pdf [accessed 6 May 2012].

Banducci A, Kohler T, Soong J, Menard J (2007) 2005 Annual management report for Norton Sound, Port Clarence, and Kotzebue [online]. Alaska Department of Fish and Game, Fishery Management Report No. 07-32. http://www.adfg.alaska.gov/FedAidPDFs/fmr07-32.pdf [accessed January 30, 2013].

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD, *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* **6**.

Beacham TD, Candy JR, Jonsen KL, Supernault J, Wetklo M, Deng L, *et al.* (2006) Estimation of Stock Composition and Individual Identification of Chinook Salmon across the Pacific Rim by Use of Microsatellite Variation. *Transactions of the American Fisheries Society* **135**, 861-888.

Beacham TD, Candy JR, Porszt E, Sato S, Urawa S (2011) Microsatellite identification of Canadian sockeye salmon rearing in the Bering Sea. *Transactions of the American Fisheries Society* **140**, 296-306.

Beacham TD, Candy JR, Wallace C, Wetklo M, Deng L, MacConnachie C (2012) Microsatellite mixed-stock identification of coho salmon in British Columbia. *Marine and Coastal Fisheries* **4**, 85-100.

Beacham TD, Murray CB, Withler RE (1989) Age, morphology, and biochemical genetic-variation of Yukon River Chinook salmon. *Trans. Am. Fish. Soc.* **118**, 46-63.

Beacham TD, Wetklo M, Wallace C, Olsen JB, Flannery BG, Wenburg JK, *et al.* (2008a) The application of microsatellites for stock identification of Yukon River Chinook salmon. *North American Journal of Fisheries Management* **28**, 283-295.

Beacham TD, Winter I, Jonsen KL, Wetklo M, Deng L, Candy JR (2008b) The application of rapid microsatellite-based stock identification to management of a Chinook salmon troll fishery off the Queen Charlotte Islands, British Columbia. *North American Journal of Fisheries Management* **28**, 849-855.

Beacham TD, Withler RE, Gould AP (1985) Biochemical genetic stock identification of chum salmon (*Oncorhynchus keta*) in southern British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences* **42**, 437-448.

Beavis WD, Smith OS, Grant D, Fincher R (1994) Identification of quantitative trait loci using a small sample of topcrossed and F4 progeny. *Crop Science* **34**, 882-896.

Begg GA, Friedland KD, Pearce JB (1999) Stock identification and its role in stock assessment and fisheries management: an overview. Fish. Res. **43**, 1-8.

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology* **16**, 363-377.
Blair GR, Rogers DE, Quinn TP (1993) Variation in life-history characteristics and morphology of sockeye salmon in the Kvichak River System, Bristol Bay, Alaska. *Transactions of the American Fisheries Society* **122**, 550-559.
Bourret V, Kent MP, Primmer CR, Vasemägi A, Karlsson S, Hindar K, *et al.* (2013) SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Molecular Ecology* **22**, 532-551.
Bradbury IR, Hubert S, Higgins B, Bowman S, Borza T, Paterson IG, *et al.* (2013) Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications* **6**, 450-461.
Bradbury IR, Hubert S, Higgins B, Bowman S, Paterson IG, Snelgrove PVR, *et al.* (2011) Evaluating SNP ascertainment bias and its impact on population assignment in Atlantic cod, *Gadus morhua*. *Molecular Ecology Resources* **11**, 218-225.
Bradford MJ (1995) Comparative review of Pacific salmon survival rates. *Canadian Journal of Fisheries and Aquatic Sciences* **52**, 1327-1338.
Braun DC, Reynolds JD (2011) Relationships between habitat characteristics and breeding population densities in sockeye salmon (*Oncorhynchus nerka*). *Canadian Journal of Fisheries and Aquatic Sciences* **68**, 758-767.
Brett JR (1952) Temperature tolerance in young Pacific salmon, genus *Oncorhynchus*. *Journal of the Fisheries Research Board of Canada* **9**, 265-323.
Brieuc MSO, Ono K, Drinan D, Naish KA (2015) Integration of Random Forest with population-based outlier analyses provides insight on the genomic basis and evolution of run timing in Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology* **24**, 2729-2746.
Brieuc MSO, Waters CD, Seeb JE, Naish KA (2014) 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**, 447-460.
Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889-890.
Brown CA, Murray AW, Verstrepen KJ (2010) Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Current Biology* **20**, 895-903.
Bugaev AV, Myers KW (2009) Stock-specific distribution and abundance of immature Chinook salmon in the Western Bering Sea in summer and fall 2002–2004. *North Pacific Anadromous Fish Commission Bulletin* **5**, 87-97.
Campbell NR, Narum SR (2008) Identification of novel single-nucleotide polymorphisms in Chinook salmon and variation among life history types. *Transactions of the American Fisheries Society* **137**, 96-106.
Campbell NR, Narum SR (2009) Identification and characterization of heat shock response related single nucleotide polymorphisms in *O. mykiss* and *O. tshawytscha*. *Molecular Ecology Resources* **9**, 1450-1559.
Campos JL, Posada D, Moran P (2006) Genetic variation at MHC, mitochondrial and microsatellite loci in isolated populations of Brown trout (*Salmo trutta*). *Conservation Genetics* **7**, 515-530.
Casacuberta E, González J (2013) The impact of transposable elements in environmental adaptation. *Molecular Ecology* **22**, 1503-1517.

Castano-Sanchez C, Fuji K, Ozaki A, Hasegawa O, Sakamoto T, Morishima K, *et al.* (2010) A second generation genetic linkage map of Japanese flounder (*Paralichthys olivaceus*). *BMC Genomics* **11**.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci *de novo* from short-read sequences. *G3: Genes Genomes Genetics*, 171-182.

Clemento AJ, Abadia-Cardoso A, Starks HA, Garza JC (2011) Discovery and characterization of single nucleotide polymorphisms in Chinook salmon, *Oncorhynchus tshawytscha*. *Molecular Ecology Resources* **11**, 50-66.

Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, Schluter D, *et al.* (2004) The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *Plos Biology* **2**, 635-641.

Consuegra S, Garcia de Leaniz C (2008) MHC-mediated mate choice increases parasite resistance in salmon. *Proceedings of the Royal Society B: Biological Sciences* **275**, 1397-1403.

Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B-Biological Sciences* **279**, 5039-5047.

Creelman EK, Hauser L, Simmons RK, Templin WD, Seeb LW (2011) Temporal and geographic genetic divergence: characterizing sockeye salmon populations in the Chignik watershed, Alaska, ising single-nucleotide polymorphisms. *Transactions of the American Fisheries Society* **140**, 749-762.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* **23**, 3133-3157.

Dann T, Habicht C, Jasper JR, Fox EKC, Hoyt H, Liller HL, *et al.* (2012) Sockeye salmon baseline for the western Alaska salmon stock identification project. Special Publication No. 12-12. Available from http://www.adfg.alaska.gov/FedAidPDFs/sp12-12 [accessed 11 June 2013].

Dann TH, Habicht C, Baker TT, Seeb JE (2013) Exploiting genetic diversity to balance conservation and harvest of migratory salmon. *Canadian Journal of Fisheries and Aquatic Sciences.* **70**, 785-793.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.

Davis ND, Myers KW, Walker RV, Harris CK (1990) The Fisheries Research Institute's high-seas salmonid tagging program and methodology for scale pattern analysis. *In* Fish marking techniques. *Edited by* N.C. Parker, A.E. Giorgi, R.C. Heidinger, D.B. Jester, Jr., E.D. Prince, and G.A. Winans, Am. Fish. Soc. Symp. 7. pp. 863-879.

Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics* **3**, 22-32.

Dionne M, Miller KM, Dodson JJ, Caron F, Bernatchez L (2007) Clinal variation in MHC diversity with temperature: evidence for the role of host-pathogen interaction on local adaptation in Atlantic salmon. *Evolution* **61**, 2154-2164.

Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR (2014) NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources* **14**, 209-214.

Du ZQ, Ciobanu DC, Onteru SK, Gorbach D, Mileham AJ, Jaramillo G, *et al.* (2010) A gene-based SNP linkage map for Pacific white shrimp, *Litopenaeus vannamei*. *Animal Genetics* **41**, 286-294.

Dupuis J, Siegmund D (1999) Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151**, 373-386.

Ekblom R, Saether SA, Jacobsson P, Fiske P, Sahlman T, Grahn M, *et al.* (2007) Spatial pattern of MHC class II variation in the great snipe (*Gallinago media*). *Molecular Ecology* **16**, 1439-1451.

Elfstrom CM, Smith CT, Seeb JE (2006) Thirty-two single nucleotide polymorphism markers for high-throughput genotyping of sockeye salmon. *Molecular Ecology Notes* **6**, 1255-1259.

Eliason EJ, Clark TD, Hague MJ, Hanson LM, Gallagher ZS, Jeffries KM, *et al.* (2011) Differences in thermal tolerance among sockeye salmon populations. *Science* **332**, 109-112.

Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**, 756-760.

Erickson DL, Fenster CB, Stenoien HK, Price D (2004) Quantitative trait locus analyses and the study of evolutionary process. *Molecular Ecology* **13**, 2505-2522.

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. *PLoS ONE* **6**, e18561.

Evans ML, Neff BD (2009) Major histocompatibility complex heterozygote advantage and widespread bacterial infections in populations of Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology* **18**, 4716-4729.

Evans ML, Neff BD, Heath DD (2010) MHC genetic structure and divergence across populations of Chinook salmon (*Oncorhynchus tshawytscha*). *Heredity* **104**, 449-459.

Everett MV, Grau ED, Seeb JE (2011) Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources* **11**, 93-108.

Everett MV, Miller MR, Seeb JE (2012) Meiotic maps of sockeye salmon derived from massively parallel DNA sequencing. *BMC Genomics* **13**, 521.

Everett MV, Seeb JE (2014) Detection and mapping of QTL for temperature tolerance and body size in Chinook salmon (*Oncorhynchus tshawytscha*) using genotyping by sequencing. *Evolutionary Applications* **7**, 480-492.

Ewens WJ (1972) Sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87-112.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285-298.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Faber-Hammond J, Phillips RB, Park LK (2012) The sockeye salmon neo-Y chromosome Is a fusion between linkage groups orthologous to the coho Y chromosome and the long arm of rainbow trout chromosome 2. *Cytogenetic and Genome Research* **136**, 69-74.

Fair LF, Moffitt SD, Evenson MJ, Erickson J (2008) Escapement goal review of Copper and Bering Rivers, and Prince William Sound Pacific salmon stocks, 2008. Alaska Department of Fish and Game, Fishery Manuscript No. 08-02. Available from http://www.sf.adfg.state.ak.us/FedAidpdfs/fms08-02.pdf [accessed 6 May 2012].

Fair LF, Wilette TM, Erickson JW, Yanusz RJ, McKinley TR (2010) Review of salmon escapement goals in upper Cook Inlet, Alaska, 2011. Alaska Department of Fish and Game, Fishery Manuscript Series No. 10-06. Available from http://www.adfg.alaska.gov/FedAidPDFs/FMS10-06.pdf [accessed 6 May 2012].

Farley EV, Murphy JM, Wing BW, Moss JH, Middleton A (2005) Distribution, migration pathways, and size of western Alaska juvenile salmon along the eastern Bering Sea Shelf. *Alaska Fishery Research Bulletin* **11**, 15-26.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977-993.

Frankham R (2005) Genetics and extinction. *Biological Conservation* **126**, 131-140.

Fraser BA, Ramnarine IW, Neff BD (2010) Selection at the MHC class IIB locus across guppy (*Poecilia reticulata*) populations. *Heredity* **104**, 155-167.

Frazer KK, Russello MA (2013) Lack of parallel genetic patterns underlying the repeated ecological divergence of beach and stream-spawning kokanee salmon. *Journal of Evolutionary Biology* **26**, 2606-2621.

Fry FEJ (1971) The effect of environmental factors on the physiology of fish. *In* Fish physiology. *Edited* by W.S. Hoar and D.J. Randall. Academic Press, New York. pp. 1–98.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution* **27**, 489-496.

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

Gagnaire P-A, Pavey SA, Normandeau E, Bernatchez L (2013b) The genetic architecture of reproductive isolation during speciation with gene flow in lake whitefish assessed by RAD sequencing. *Evolution* **67**, 2483-2497.

Garcia de Leaniz C, Fleming IA, Einum S, Verspoor E, Jordan WC, Consuegra S, *et al.* (2007) A critical review of adaptive genetic variation in Atlantic salmon: implications for conservation. *Biological Reviews of the Cambridge Philosophical Society (London)* **82**, 173-211.

Garvin MR, Kondzela CM, Martin P, Finney B, Guyon JR, Templin WD, *et al.* (2013) Recent physical connections may explain weak genetic structure in western Alaskan chum salmon (*Oncorhynchus keta*) populations. *Ecology and Evolution* **3**, 2362-2377.

Gaume B, Sharp RE, Manson FDC, Chapman SK, Reid GA, Lederer F (1995) Mutation to glutamine of histidine-372, the catalytic base of flavocytochrome-B(2) (L-lactate dehydrogenase). *Biochimie* **77**, 621-630.

Gauthier-Ouellet M, Dionne M, Caron F, King TL, Bernatchez L (2009) Spatiotemporal dynamics of the Atlantic salmon (*Salmo salar*) Greenland fishery inferred from mixed-stock analysis. *Canadian Journal of Fisheries and Aquatic Sciences* **66**, 2040-2051.

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statististical Science* **7**, 457-511.

Gharrett AJ, Shirley SM, Tromble GR (1987) Genetic-relationships among populations of Alaskan Chinook salmon (*Oncorhynchus tshawytscha*). *Canadian Journal of Fisheries and Aquatic Sciences* **44**, 765-774.
Gidskehaug L, Kent M, Hayes BJ, Lien S (2011) Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array. *Bioinformatics* **27**, 303-310.
Gisclair BR (2009) Salmon bycatch management in the Bering Sea walleye pollock fishery: threats and opportunities for western Alaska. American Fisheries Society Symposium 70:799-816.
Gomez-Uchida D, Seeb James E, Habicht C, Seeb Lisa W (2012) Allele frequency stability in large, wild exploited populations over multiple generations: insights from Alaska sockeye salmon (*Oncorhynchus nerka*). *Canadian Journal of Fisheries and Aquatic Sciences* **69**, 916-929.
Gomez-Uchida D, Seeb JE, Smith MJ, 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**.
Goudet J (1995) FSTAT (Version 1.2): A computer program to calculate F-statistics. *Journal of Heredity* **86**, 485-486.
Goudet J, Raymond M, deMeeus T, Rousset F (1996) Testing differentiation in diploid populations. *Genetics* **144**, 1933-1940.
Groot AT, Staudacher H, Barthel A, Inglis O, Schofl G, Santangelo RG, *et al.* (2013) One quantitative trait locus for intra- and interspecific variation in a sex pheromone. *Molecular Ecology* **22**, 1065-1080.
Gruenthal KM, Witting DA, Ford T, Neuman MJ, Williams JP, Pondella DJ, II, *et al.* (2013) Development and application of genomic tools to the restoration of green abalone in southern California. *Conservation Genetics*, 1-13.
Gunderson D (2011) The rockfish's warning. *University Bookstore Press, Seattle*.
Guthrie CM, III., Nguyen HT, Guyon JR (2012) Genetic stock composition analysis of Chinook salmon bycatch samples from the 2010 Bering Sea trawl fisheries. U.S. Dep. Commer., NOAA Tech. Memo. NMFS-AFSC-232, 22 p. Available from http://www.afsc.noaa.gov/Publications/AFSC-TM/NOAA-TM-AFSC-232.pdf [accessed 6 May 2012].
Guthrie CM, Nguyen HT, Guyon JR (2013) Genetic stock composition analysis of Chinook salmon bycatch samples from the 2011 Bering Sea and Gulf of Alaska trawl fisheries. U.S. Dep. Commer., NOAA Tech. Memo. NMFS-AFSC-244, 28 p.
Gutierrez AP, Lubieniecki KP, Fukui S, Withler RE, Swift B, Davidson WS (2014) Detection of Quantitative Trait Loci (QTL) Related to Grilsing and Late Sexual Maturation in Atlantic Salmon (*Salmo salar*). *Marine Biotechnology* **16**, 103-110.
Habicht C, Seeb LW, Myers KW, Farley EV, Seeb JE (2010) Summer–fall distribution of stocks of immature sockeye salmon in the Bering Sea as revealed by single-nucleotide polymorphisms. *Transactions of the American Fisheries Society* **139**, 1171-1191.
Habicht C, Seeb LW, Seeb JE (2007) Genetic and ecological divergence defines population structure of sockeye salmon populations returning to Bristol Bay, Alaska, and provides a tool for admixture analysis. *Trans. Am. Fish. Soc.* **136**, 82-94.

Hale MC, Thrower FP, Berntson EA, Miller MR, Nichols KM (2013) Evaluating adaptive divergence between migratory and nonmigratory ecotypes of a salmonid fish, *Oncorhynchus mykiss*. *G3: Genes Genomes Genetics* **3**, 1273-1285.
Hare MP, Nunney L, Schwartz MK, Ruzzante DE, Burford M, Waples RS, *et al.* (2011) Understanding and estimating effective population size for practical application in marine species management. *Conservation Biology* **25**, 438-449.
Hauser L, Seeb JE (2008) Advances in molecular technology and their impact on fisheries genetics. *Fish and Fisheries* **9**, 473-486.
Healey MC (1991) Life history of Chinook salmon (*Oncorhynchus tshawytscha*). *In* Pacific salmon life histories. *Edited by* C. Groot and L. Margolis. University of British Columbia Press, Vancouver. pp. 311-393. .
Heard WR, Shevlyakov E, Zikunova OV, McNicol RE (2007) Chinook salmon – trends in abundance and biological characteristics. N. Pac. Anadr. Fish Comm. Bull. **4**: 77–91.
Hecht BC, Campbell NR, Holecek DE, Narum SR (2013) Genome-wide association reveals genetic basis for the propensity to migrate in wild populations of rainbow and steelhead trout. *Molecular Ecology* **22**, 3061-3076.
Hecht BC, Thrower FP, Hale MC, Miller MR, Nichols KM (2012) Genetic architecture of migration-related traits in rainbow and steelhead trout, *Oncorhynchus mykiss*. *G3: Genes Genomes Genetics* **2**, 1113-1127.
Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution* **59**, 1633-1638.
Hemmer-Hansen J, Nielsen EE, Therkildsen NO, Taylor MI, Ogden R, Geffen AJ, *et al.* (2013) A genomic island linked to ecotype divergence in Atlantic cod. *Molecular Ecology* **22**, 2653-2667.
Henderson MA, Graham CC (1998) History and status of Pacific salmon in British Columbia. N. Pac. Anadr. Fish Comm. Bull. **1**: 13-22.
Hendry AP, Wenburg JK, Bentzen P, Volk EC, Quinn TP (2000) Rapid evolution of reproductive isolation in the wild: Evidence from introduced salmon. *Science* **290**, 516-518.
Hess JE, Campbell NR, Close DA, Docker MF, Narum SR (2013) Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. *Molecular Ecology* **22**, 2898-2916.
Hilborn R, Quinn TP, Schindler DE, Rogers DE (2003) Biocomplexity and fisheries sustainability. *Proceedings of the National Academy of Sciences* **100**, 6564-6568.
Hill AVS, Allsopp CEM, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, *et al.* (1991) Common west African HLA antigens are associated with protection from severe malaria. *Nature* **352**, 595-600.
Hill WG (1981) Estimation of effective population-size from data on linkage disequilibrium. *Genetical Research* **38**, 209-216.
Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* **11**, 117-122.
Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics* **6**, e1000862.

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC, *et al.* (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology* **22**, 3002-3013.

Holland PM, Abramson RD, Watson R, Gelfand DH (1991) Detection of specific polymerase chain-reaction product by the 5'- 3' exonuclease activity of *Thermus-aquaticus* DNA-polymerase. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 7276-7280.

Howard KG, Hayes SJ, Evenson DF (2009) Yukon River Chinook salmon stock status and action plan 2010; a report to the Alaska Board of Fisheries. Alaska Department of Fish and Game, Special Publication No. 09-26. Available from http://www.sf.adfg.state.ak.us/FedAidpdfs/Sp09-26.pdf [accessed 6 May 2012].

Huang SW, Yu HT (2003) Genetic variation of microsatellite loci in the major histocompatibility complex (MHC) region in the southeast Asian house mouse (*Mus musculus castaneus*). *Genetica* **119**, 201-218.

Huang XQ, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Research* **9**, 868-877.

Hubert S, Higgins B, Borza T, Bowman S (2010) Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). *BMC Genomics* **11**, 191.

Hutchinson WF (2008) The dangers of ignoring stock complexity in fishery management: the case of the North Sea cod. *Biology Letters* **4**, 693-695.

Iwamoto EM, Myers JM, Gustafson RG (2012) Resurrecting an extinct salmon evolutionarily significant unit: archived scales, historical DNA and implications for restoration. *Molecular Ecology* **21**, 1567-1582.

Jarvi SI, Tarr CL, McIntosh CE, Atkinson CT, Fleischer RC (2004) Natural selection of the major histocompatibility complex (MHC) in Hawaiian honeycreepers (*Drepanidinae*). *Molecular Ecology* **13**, 2157-2168.

Johnston SE, Orell P, Pritchard VL, Kent MP, Lien S, Niemelä E, *et al.* (2014) Genome-wide SNP analysis reveals a genetic basis for sea-age variation in a wild population of Atlantic salmon (*Salmo salar*). *Molecular Ecology* **23**, 3452-3468.

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

Kalinowski ST, Wagner AP, Taper ML (2006) ML-RELATE: a computer program for maximum likelihood estimation of relatedness and relationship. *Molecular Ecology Notes* **6**, 576-579.

Kass RE, Carlin BP, Gelman A, Neal RM (1998) Markov chain Monte Carlo in practice: A roundtable discussion. *Journal of the American Statistical Association* **52**, 93-100.

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

Knapp LA, Cadavid LF, Watkins DI (1998) The MHC-E locus is the most well conserved of all known primate class I histocompatibility genes. *Journal of Immunology* **160**, 189-196.

Kodama M, Brieuc MSO, Devlin RH, Hard JJ, Naish KA (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**, 1717-1730.

Lander E, Kruglyak L (1995) Genetic dissection of complex traits - guidelines for interpreting and reporting results. *Nature Genetics* **11**, 241-247.

Landry C, Bernatchez L (2001) Comparative analysis of population structure across environments and geographical scales at major histocompatibility complex and microsatellite loci in Atlantic salmon (*Salmo salar*). *Molecular Ecology* **10**, 2525-2539.
Langefors A, Lohm J, Grahn M, Andersen O, von Schantz T (2001) Association between major histocompatibility complex class IIB alleles and resistance to *Aeromonas salmonicida* in Atlantic salmon. *Proceedings of the Royal Society B-Biological Sciences* **268**, 479-485.
Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**.
Larson W, Seeb JE, Limborg MT, McKinney GJ, Everett MV, Seeb LW (Submitted) Identification of multiple QTL hotspots in sockeye salmon (*Oncorhynchus nerka*) using genotyping by sequencing and a dense linkage map.
Larson WA, Seeb JE, Dann TH, Schindler DE, Seeb LW (2014a) Signals of heterogeneous selection at an MHC locus in geographically proximate ecotypes of sockeye salmon. *Molecular Ecology* **23**, 5448-5461.
Larson WA, Seeb JE, Pascal CE, Templin WD, Seeb LW (2014b) Single-nucleotide polymorphisms (SNPs) identified through genotyping-by-sequencing improve genetic stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) from western Alaska. *Canadian Journal of Fisheries and Aquatic Sciences* **71**, 698-708.
Larson WA, Seeb LW, Everett MV, Waples RK, Templin WD, Seeb JE (2014c) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications* **7**, 355-369.
Larson WA, Utter FM, Myers KW, Templin WD, Seeb JE, Guthrie CM, *et al.* (2013) Single-nucleotide polymorphisms reveal distribution and migration of Chinook salmon (*Oncorhynchus tshawytscha*) in the Bering Sea and North Pacific Ocean. *Canadian Journal of Fisheries and Aquatic Sciences* **70**, 128-141.
Lemay MA, Russello MA (2015) Genetic evidence for ecological divergence in kokanee salmon. *Molecular Ecology* **24**, 798-811.
Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452.
Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, Davidson WS, *et al.* (2011) 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**.
Limborg MT, Waples RK, Allendorf FA, Seeb JE (Submitted) Linkage mapping reveals strong chiasma interference in sockeye salmon: Implications for interpreting genomic data.
Lin J, Quinn TP, Hilborn R, Hauser L (2008a) Fine-scale differentiation between sockeye salmon ecotypes and the effect of phenotype on straying. *Heredity* **101**, 341-350.
Lin J, Ziegler E, Quinn TP, Hauser L (2008b) Contrasting patterns of morphological and neutral genetic divergence among geographically proximate populations of sockeye salmon *Oncorhynchus nerka* in Lake Aleknagik, Alaska. *Journal of Fish Biology* **73**, 1993-2004.
Lingnau T (1996) Norton Sound and Kotzebue Sound management area salmon catch and escapement report, 1995 [online]. Alaska Department of Fish and Game, Regional Information Report No. 3A96-23. Available from http://www.adfg.alaska.gov/FedAidpdfs/RIR.3A.1996.23.pdf [accessed 30 January 2013].

Lisi PJ, Schindler DE, Bentley KT, Pess GR (2013) Association between geomorphic attributes of watersheds, water temperature, and salmon spawn timing in Alaskan streams. *Geomorphology* **185**, 78-86.
Lukacs MF, Harstad H, Bakke HG, Beetz-Sargent M, McKinnel L, Lubieniecki KP, *et al.* (2010) Comprehensive analysis of MHC class I genes from the U-, S-, and Z-lineages in Atlantic salmon. *BMC Genomics* **11**.
Lynch M, Walsh B (1998) Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, MA.
Major RL, Ito J, Ito S, Godfrey H (1978) Distribution and abundance of Chinook salmon (*Oncorhynchus tshawytscha*) in offshore waters of the North Pacific. *International North Pacific Fish Commission Bulletin* **38**, 1-54.
Marcogliese DJ (2001) Implications of climate change for parasitism of animals in the aquatic environment. *Canadian Journal of Zoology-Revue Canadienne De Zoologie* **79**, 1331-1352.
Margolis L (1963) Parsites as indicators of the geographical origin of sockeye salmon, *Oncorhynchus nerka* (Walbaum), occurring in the North Pacific Ocean and adjacent seas. *International North Pacific Fish Commission Bulletin* **11**, 101-156.
Marlowe C, Bucsack C (1995) The effect of decreasing sample size on the precision of GSI stock composition estimates for chinook salmon (*Oncorhynchus tshawytscha*) using data from the Washington Coastal and Strait of Juan de Fuca troll fisheries in 1989-1990. Northwest Fishery Resource Bulletin, Project Report Series no. 2. Available from http://nwifc.dreamhosters.com/w/wp-content/uploads/2008/07/nwfrb_prs002.pdf [accessed 5 October 2012].
Martins EG, Hinch SG, Patterson DA, Hague MJ, Cooke SJ, Miller KM, *et al.* (2012) High river temperature reduces survival of sockeye salmon (*Oncorhynchus nerka*) approaching spawning grounds and exacerbates female mortality. *Canadian Journal of Fisheries and Aquatic Sciences* **69**, 330-342.
Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC (2014) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources* **15**, 28-41.
Matsumura M, Fremont DH, Peterson PA, Wilson IA (1992) Emerging principles for the recognition of peptide antigens by MHC class-I molecules. *Science* **257**, 927-934.
McClelland EK, Ming TJ, Tabata A, Kaukinen KH, Beacham TD, Withler RE, *et al.* (2013) Patterns of selection and allele diversity of class I and class II major histocompatibility loci across the species range of sockeye salmon (*Oncorhynchus nerka*). *Molecular Ecology* **22**, 4783-4800.
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.
McKeown NJ, Hynes RA, Duguid RA, Ferguson A, Prodoehl PA (2010) Phylogeographic structure of brown trout *Salmo trutta* in Britain and Ireland: glacial refugia, postglacial colonization and origins of sympatric populations. *Journal of Fish Biology* **76**, 319-347.
McKinney GJ, Seeb LW, Larson WA, Gomez-Uchida D, Limborg MT, Brieuc MSO, *et al.* (Submitted) An integrated linkage map reveals candidate genes underlying adaptive

variation in Chinook salmon (*Oncorhynchus tshawytscha*). Available for review on request from agajoh@oregonstate.edu.
McPhee MV, Tappenbeck TH, Whited DC, Stanford JA (2009) Genetic diversity and population structure in the Kuskokwim River drainage support the recurrent evolution hypothesis for sockeye salmon life histories. *Transactions of the American Fisheries Society* **138**, 1481-1489.
McPherson S, Bernard D, Clark JH, Pahlke K, Jones E, Hovanisian JD, *et al.* (2003) Stock status and escapement goals for Chinook salmon stocks in Southeast Alaska. Alaska Department of Fish and Game, Special Publication 03-01. Available from http://www.sf.adfg.state.ak.us/FedAidpdfs/Sp03-01.pdf [accessed 6 May 2012].
Mercer B, Wilson JK (2011) 2010 Chinook salmon sonar enumeration on the Big Salmon River [online]. Prepared for the Yukon River Panel Restoration and Enhancement Fund, CRE-41-10. http://yukonriverpanel.com/salmon/wp-content/uploads/2011/04/cre-41-10-big-salmon-sonar-final-report.pdf [accessed January 30, 2012].
Miller HC, Allendorf F, Daugherty CH (2010) Genetic diversity and differentiation at MHC genes in island populations of tuatara (*Sphenodon spp.*). *Molecular Ecology* **19**, 3894-3908.
Miller HC, Lambert DM (2004) Genetic drift outweighs balancing selection in shaping post-bottleneck major histocompatibility complex variation in New Zealand robins (*Petroicidae*). *Molecular Ecology* **13**, 3709-3721.
Miller KM, Kaukinen KH, Beacham TD, Withler RE (2001) Geographic heterogeneity in natural selection on an MHC locus in sockeye salmon. *Genetica* **111**, 237-257.
Miller KM, Teffer A, Tucker S, Li S, Schulze AD, Trudel M, *et al.* (2014) Infectious disease, shifting climates, and opportunistic predators: cumulative factors potentially impacting wild salmon declines. *Evolutionary Applications* **7**, 812-855.
Miller KM, Withler RE (1996) Sequence analysis of a polymorphic MHC class II gene in Pacific salmon. *Immunogenetics* **43**, 337-351.
Miller MR, Brunelli JP, Wheeler PA, Liu S, Rexroad CE, III, Palti Y, *et al.* (2012) A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology* **21**, 237-249.
Milner GB, Teel DJ, Utter FM, Winans GA (1985) A genetic method of stock identification in mixed populations of Pacific Salmon, *Oncorhynchus spp*. *Marine Fisheries Review* **47**, 1-8.
Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology* **12**.
Molyneaux DB, Brannian LK (2006) Review of escapement and abundance information for Kuskokwim area salmon stocks. Alaska Department of Fish and Game, Fishery Manuscript No. 06-08. Available from http://www.sf.adfg.state.ak.us/FedAidpdfs/fms06-08.pdf [accessed 6 May 2012].
Molyneaux DB, Dubois L (1999) Salmon age, sex and length catalog for the Kuskokwim area, 1998 progress report [online]. Alaska Department of Fish and Game, Regional Information Report No. 3A99-15. Available from http://www.sf.adfg.state.ak.us/FedAidpdfs/RIR.3A.1999.15.pdf [accessed 30 January 2013].

Moriya S, Sato S, Azumaya T, Suzuki O, Urawa S, Urano A, *et al.* (2007) Genetic stock identification of chum salmon in the Bering Sea and North Pacific Ocean using mitochondrial DNA microarray. *Marine Biotechnology* **9**, 179-191.

Mukhopadhyay S, Jackson PK (2011) The tubby family proteins. *Genome Biology* **12**.

Murphy JM, Templin WD, Farley EVJ, Seeb JE (2009) Stock-structured distribution of western Alaska and Yukon juvenile Chinook salmon (*Oncorhynchus tshawytscha*) from United States BASIS surveys, 2002–2007. *North Pacific Anadromous Fish Commission Bulletin* **5**, 51-59.

Myers JM, Kope RG, Bryant GJ, Teel D, Lierheimer LJ, Wainwright TC, *et al.* (1998) Status review of Chinook salmon from Washington, Idaho, Oregon, and California. U.S. Dept. Commer., NOAA Tech. Memo. NMFS-NWFSC-35, 443 p. Available from http://www.nwfsc.noaa.gov/publications/techmemos/tm35/index.htm#toc [accessed 6 May 2012].

Myers KW, Aydin KY, Walker RV, Fowler S, Dahlberg ML (1996) Known ocean ranges of stocks of Pacific salmon and steelhead as shown by tagging experiments, 1956-1995. *North Pacific Anadromous Fish Commission Document 192. Available from www.npafc.org [accessed 6 May 2012]*.

Myers KW, Celewycz AG, Farley EV (2004) High seas salmonid coded-wire tag recovery data, 2004. N. Pac. Anadr. Fish Comm. Doc. 804, SAFS-UW-04, School of Aquatic and Fishery Sciences, University of Washington, Seattle. 22p. Available from https://digital.lib.washington.edu/researchworks/handle/1773/4538 [accessed 6 May 2012].

Myers KW, Harris CK, Ishida Y, Margolis L, Ogura M (1993) Review of the Japanese landbased driftnet salmon fishery in the western North Pacific Ocean and the continent of origin of salmonids in this area. Int. N. Pac. Fish Comm. Bull. **52**: 86p.

Myers KW, Harris CK, Knudsen CM, Walker RV, Davis ND, Rogers DE (1987) Stock Origins of Chinook Salmon in the Area of the Japanese Mothership Salmon Fishery. *North American Journal of Fisheries Management* **7**, 459-474.

Myers KW, Klovach NV, Gritsenko OF, Urawa S, Royer TC (2007) Stock-specific distributions of Asian and North American salmon in the open ocean, interannual changes, and oceanographic conditions. *North Pacific Anadromous Fish Commission Bulletin* **4**, 159-177.

Myers KW, Rogers DE (1988) Stock origins of Chinook salmon in incidental catches by groundfish fisheries in the eastern Bering Sea. *N. Am. J. Fish. Manage.* **8**, 162-171.

Myers KW, Walker RV, Davis ND, Armstrong JA, Fournier WJ, Mantua NJ, *et al.* (2010) Climate-ocean effects on Chinook salmon. *Arctic Yukon Kuskokwim Sustainable Salmon Initiative, Project Final Product. SAFS-UW-1003, School of Aquatic and Fishery Sciences, University of Washington, Seattle. 249 p. Available from https://digital.lib.washington.edu/dspace/handle/1773/16308 [accessed 6 May 2012]*.

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, *et al.* (2012) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B-Biological Sciences* **367**, 343-353.

Nagasawa T, Azumaya T (2009) Distribution and CPUE trends in Pacific salmon, especially sockeye salmon in the Bering Sea and adjacent waters from 1972 to the mid 2000s. *North Pacific Anadromous Fish Commission Bulletin* **5**, 1-13.

Nandor FN, Longwill JR, Webb DL (2009) Overview of the coded wire tag program in the greater Pacific region of North America. Available from www.rmpc.org/files/Nandor_CWT_Overview.pdf [accessed 6 May 2012].
Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology* **22**, 2841-2847.
Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. *Journal of Molecular Evolution* **19**, 153-170.
Nelson PA, Hasbrouck JJ, Witteveen MJ, Bouwnes KA, Vining I (2006) Review of salmon escapement goals in the Alaska Peninsula and Aleutian Islands management areas- report to the Alaska Board of Fisheries, 2004. Alaska Department of Fish and Game, Fishery Manuscript No. 06-03. Available from http://www.sf.adfg.state.ak.us/fedaidpdfs/fms06-03.pdf [accessed 6 May 2012].
Neville HM, Isaak DJ, Dunham JB, Thurow RF, Rieman BE (2006) Fine-scale natal homing and localized movement as shaped by sex and spawning habitat in Chinook salmon: insights from spatial autocorrelation analysis of individual genotypes. *Molecular Ecology* **15**, 4589-4602.
Nielsen EE, Cariani A, Mac Aoidh E, Maes GE, Milano I, Ogden R, *et al.* (2012) Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature Communications* **3:851**.
Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009) Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology* **18**, 3128-3150.
Nielsen EE, Kenchington E (2001) A new approach to prioritizing marine fish and shellfish populations for conservation. *Fish and Fisheries* **2**, 328-343.
Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12**, 443-451.
Niskanen AK, Kennedy LJ, Ruokonen M, Kojola I, Lohi H, Isomursu M, *et al.* (2013) Balancing selection and heterozygote advantage in MHC loci of the bottlenecked Finnish wolf population. *Molecular Ecology* **23**, 875-889.
Nosil P, Egan SP, Funk DJ (2008) Heterogeneous genomic differentiation between walking-stick ecotypes: "Isolation by adaptation" and multiple roles for divergent selection. *Evolution* **62**, 316-336.
Nosil P, Feder JL (2012) Genomic divergence during speciation: causes and consequences *Philosophical Transactions of the Royal Society B-Biological Sciences* **367**, 332-342.
Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology* **18**, 375-402.
O'Malley KG, Sakamoto T, Danzmann RG, Ferguson MM (2003) Quantitative trait loci for spawning date and body weight in rainbow trout: Testing for conserved effects across ancestrally duplicated chromosomes. *Journal of Heredity* **94**, 273-284.
Oliver MK, Lambin X, Cornulier T, Piertney SB (2009) Spatio-temporal variation in the strength and mode of selection acting on major histocompatibility complex diversity in water vole (*Arvicola terrestris*) metapopulations. *Molecular Ecology* **18**, 80-92.
Olsen JB, Beacham TD, Wetklo M, Seeb LW, Smith CT, Flannery BG, *et al.* (2010) The influence of hydrology and waterway distance on population structure of Chinook salmon *Oncorhynchus tshawytscha* in a large river. *Journal of Fish Biology* **76**, 1128-1148.

Olsen JB, Crane PA, Flannery BG, Dunmall K, Templin WD, Wenburg JK (2011) Comparative landscape genetic analysis of three Pacific salmon species from subarctic North America. *Conservation Genetics* **12**, 223-241.
Palstra FP, Ruzzante DE (2008) Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? *Molecular Ecology* **17**, 3428-3447.
Palstra FP, Ruzzante DE (2011) Demographic and genetic factors shaping contemporary metapopulation effective size and its empirical estimation in salmonid fish. *Heredity* **107**, 444-455.
Park L (2011) Effective population size of current human population. *Genetics Research* **93**, 105-114.
Pavey SA, Nielsen JL, Hamon TR (2010) Recent ecological divergence despite migration in sockeye salmon (*Oncorhynchus nerka*). *Evolution* **64**, 1773-1783.
Pavlidis P, Jensen JD, Stephan W, Stamatakis A (2012) A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution* **29**, 3237-3248.
Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* **28**, 2537-2539.
Pearse DE, Miller MR, Abadía-Cardoso A, Garza JC (2014) Rapid parallel evolution of standing variation in a single, complex, genomic region is associated with life history in steelhead/rainbow trout. *Proceedings of the Royal Society B: Biological Sciences* **281**.
Peel D, Waples RS, Macbeth GM, Do C, Ovenden JR (2013) Accounting for missing data in the estimation of contemporary genetic effective population size (Ne). *Molecular Ecology Resources* **13**, 243-253.
Pella J, Masuda M (2001) Bayesian methods for analysis of stock mixtures from genetic characters. *Fishery Bulletin* **99**, 151-167.
Pella JJ, Milner GB (1987) Use of genetic marks in stock composition analysis. Pages 247-276 *in* N. Ryman, and F. Utter, editors. Population Genetics and Fishery Management. Washington Sea Grant, Univ. of Washington Press, Seattle.
Perry GML, Ferguson MM, Sakamoto T, Danzmann RG (2005) Sex-linked quantitative trait loci for thermotolerance and length in the rainbow trout. *Journal of Heredity* **96**, 97-107.
Pess GR, Quinn TP, Schindler DE, Liermann MC (2013) Freshwater habitat associations between pink (*Oncorhynchus gorbuscha*), chum (*O. keta*) and Chinook salmon (*O. tshawytscha*) in a watershed dominated by sockeye salmon (*O. nerka*) abundance. *Ecology of Freshwater Fish* **23**, 360-372.
Peterson DA, Hilborn R, Hauser L (2014) Local adaptation limits lifetime reproductive success of dispersers in a wild salmon metapopulation. *Nature Communications* **5**.
Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex. *Heredity* **96**, 7-21.
Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A (2004) GENECLASS2: A software for genetic assignment and first-generation migrant detection. *Journal of Heredity* **95**, 536-539.
Potts WK, Wakeland EK (1990) The maintenance of MHC polymorphism. *Immunology Today* **11**, 39-39.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559-575.
Quinn TP (2005) *The behavior and ecology of Pacific salmon and trout. University of Washington Press, Seattle*.
Quinn TP, Hendry AP, Wetzel LA (1995) The influence of life history trade-offs and the size of incubation gravels on egg size variation in sockeye salmon (*Oncorhynchus nerka*). *Oikos* **74**, 425-438.
Quinn TP, Wetzel L, Bishop S, Overberg K, Rogers DE (2001) Influence of breeding habitat on bear predation and age at maturity and sexual dimorphism of sockeye salmon populations. *Canadian Journal of Zoology* **79**, 1782-1793.
Radice AD, Bugaj B, Fitch DHA, Emmons SW (1994) Widespread occurrence of the TC1 transposon family - TC1-like transposons from teleost fish. *Molecular & General Genetics* **244**, 606-612.
Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 9197-9201.
Rastas P, Paulin L, Hanski I, Lehtonen R, Auvinen P (2013) Lep-MAP: fast and accurate linkage map construction for large SNP datasets. *Bioinformatics* **29**, 3128-3134.
Raymond M, Rousset F (1995) An exact test for population differentiation. *Evolution* **49**, 1280-1283.
Reeve HK, Sherman PW (1993) Adaptation and the goals of evolutionary research. *Quarterly Review of Biology* **68**, 1-32.
Reid DP, Szanto A, Glebe B, Danzmann RG, Ferguson MM (2005) QTL for body weight and condition factor in Atlantic salmon (*Salmo salar*): comparative analysis with rainbow trout (*Oncorhynchus mykiss*) and Arctic charr (*Salvelinus alpinus*). *Heredity* **94**, 166-172.
Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications* **4**, 1827.
Renaut S, Maillet N, Normandeau E, Sauvage C, Derome N, Rogers SM, *et al.* (2012) Genome-wide patterns of divergence during speciation: the lake whitefish case study. *Philosophical Transactions of the Royal Society B-Biological Sciences* **367**, 354-363.
Reno PW (1998) Factors involved in the dissemination of disease in fish populations. *Journal of Aquatic Animal Health* **10**, 160-171.
Rice WR (1989) Analyzing tables of statistical tests. *Evolution* **43**, 223-225.
Robinson ML, Gomez-Raya L, Rauw WM, Peacock MM (2008) Fulton's body condition factor K correlates with survival time in a thermal challenge experiment in juvenile Lahontan cutthroat trout (*Oncorhynchus clarki henshawi*). *Journal of Thermal Biology* **33**, 363-368.
Roesti M, Salzburger W, Berner D (2012) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology* **12**.
Roff DA (2007) A centennial celebration for quantitative genetics. *Evolution* **61**, 1017-1032.
Rousset F (2008) GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.
Ruggerone GT, Goetz FA (2004) Survival of Puget Sound Chinook salmon (*Oncorhynchus tshawytscha*) in response to climate-induced competition with pink salmon

(*Oncorhynchus gorbuscha*). *Canadian Journal of Fisheries and Aquatic Sciences* **61**, 1756-1770.
Russello MA, Kirk SL, Frazer KK, Askey PJ (2012) Detection of outlier loci and their utility for fisheries management. *Evolutionary Applications* **5**, 39-52.
Sagarin R, Carlsson J, Duval M, Freshwater W, Godfrey MH, Litaker W, *et al.* (2009) Bringing molecular tools into environmental resource management: Untangling the molecules to policy pathway. *Plos Biology* **7**, 426-430.
Sandone G (1995) Anvik River salmon escapement study, 1994 [online]. Alaska Department of Fish and Game, Regional Information Report No. 3A95-08. Available from http://www.sf.adfg.state.ak.us/FedAidPDFs/RIR.3A.1995.08.pdf [accessed 30 January 2013].
Santure AW, De Cauwer I, Robinson MR, Poissant J, Sheldon BC, Slate J (2013) Genomic dissection of variation in clutch size and egg mass in a wild great tit (*Parus major*) population. *Molecular Ecology* **22**, 3949-3962.
Schindler DE, Hilborn R, Chasco B, Boatright CP, Quinn TP, Rogers LA, *et al.* (2010) Population diversity and the portfolio effect in an exploited species. *Nature* **465**, 609-612.
Schlotterer C (2004) The evolution of molecular markers - just a matter of fashion? *Nature Reviews Genetics* **5**, 63-69.
Schroder G, Schroder J (1992) A single change of histidine to glutamine alters the substrate preference of a stilbene synthase. *Journal of Biological Chemistry* **267**, 20558-20560.
Seddon JM, Baverstock PR (1999) Variation on islands: major histocompatibility complex (Mhc) polymorphism in populations of the Australian bush rat. *Molecular Ecology* **8**, 2071-2079.
Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW (2011a) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources* **11**, 1-8.
Seeb JE, Pascal CE, Grau ED, Seeb LW, Templin WD, Harkins T, *et al.* (2011b) Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids. *Molecular Ecology Resources* **11**, 335-348.
Seeb JE, Pascal CE, Ramakrishnan R, Seeb LW (2009a) SNP genotyping by the 5 '-nuclease reaction: advances in high-throughput genotyping with nonmodel organisms. *In* Single nucleotide polymorphisms: methods and protocols, Second Edition. *Edited by* A.A. Komar. Methods in Molecular Biology 578: 277-292.
Seeb LW, Crane PA (1999a) Allozymes and mitochondrial DNA discriminate Asian and North American populations of chum salmon in mixed-stock fisheries along the south coast of the Alaska Peninsula. *Transactions of the American Fisheries Society* **128**, 88-103.
Seeb LW, Crane PA (1999b) High genetic heterogeneity in chum salmon in western Alaska, the contact zone between northern and southern lineages. *Transactions of the American Fisheries Society* **128**, 58-87.
Seeb LW, Crane PA, Kondzela CM, Wilmot RL, Urawa S, Varnavskaya NV, *et al.* (2004) Migration of Pacific Rim chum salmon on the high seas: insights from genetic data. *Environmental Biology of Fishes* **69**, 21-36.
Seeb LW, DeCovich NA, Barclay AW, Smith C, Templin WD (2009b) Timing and origin of Chinook salmon stocks in the Copper River and adjacent ocean fisheries using DNA markers [online]. Annual report for Study 04-507 USFWS Office Subsistence

Management Fisheries Resource Monitoring Program. Available from http://www.adfg.alaska.gov/FedAidpdfs/fds09-58.pdf [accessed 6 December 2013].
Seeb LW, Habicht C, Templin WD, Tarbox KE, Davis RZ, Brannian LK, *et al.* (2000) Genetic diversity of sockeye salmon of Cook Inlet, Alaska, and its application to management of populations affected by the *Exxon Valdez* oil spill. *Transactions of the American Fisheries Society* **129**, 1223-1249.
Seeb LW, Templin WD, Sato S, Abe S, Warheit K, Park JY, *et al.* (2011c) Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. *Molecular Ecology Resources* **11**, 195-217.
Shaklee JB, Beacham TD, Seeb L, White BA (1999) Managing fisheries using genetic data: case studies from four species of Pacific salmon. *Fisheries Research* **43**, 45-78.
Slate J (2005) Quantitative trait locus mapping in natural populations: progress, caveats and future directions. *Molecular Ecology* **14**, 363-379.
Slatkin M (1996) A correction to the exact test based on the Ewens sampling distribution. *Genetical Research* **68**, 259-260.
Smith CT, Antonovich A, Templin WD, Elfstrom CM, Narum SR, Seeb LW (2007) Impacts of marker class bias relative to locus-specific variability on population inferences in Chinook salmon: A comparison of single-nucleotide polymorphisms with short tandem repeats and allozymes. *Transactions of the American Fisheries Society* **136**, 1674-1687.
Smith CT, Elfstrom CM, Seeb LW, Seeb JE (2005a) Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Molecular Ecology* **14**, 4193-4203.
Smith CT, Seeb JE, Schwenke P, Seeb LW (2005b) Use of the 5 '-nuclease reaction for single nucleotide polymorphism genotyping in Chinook salmon. *Transactions of the American Fisheries Society* **134**, 207-217.
Smith CT, Templin WD, Seeb JE, Seeb LW (2005c) Single nucleotide polymorphisms provide rapid and accurate estimates of the proportions of US and Canadian Chinook salmon caught in Yukon River fisheries. *North American Journal of Fisheries Management* **25**, 944-953.
Smith MJ, Pascal CE, Grauvogel Z, Habicht C, Seeb JE, Seeb LW (2011) Multiplex preamplification PCR and microsatellite validation enables accurate single nucleotide polymorphism genotyping of historical fish scales. *Molecular Ecology Resources* **11**, 268-277.
Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, *et al.* (2014) Stick insect genomes reveal natural selection's role in parallel speciation. *Science* **344**, 738-742.
Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978-989.
Stewart IJ, Quinn TP, Bentzen P (2003) Evidence for fine-scale natal homing among island beach spawning sockeye salmon, *Oncorhynchus nerka*. *Environmental Biology of Fishes* **67**, 77-85.
Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* **100**, 158-170.
Storer CG, Pascal CE, Roberts SB, Templin WD, Seeb LW, Seeb JE (2012) Rank and order: evaluating the performance of SNPs for individual assignment in a non-model organism. *PLoS ONE* **7**, e49018.

Stram DL, Ianelli JN (2009) Eastern Bering Sea pollock trawl fisheries: variation in salmon bycatch over time and space. *In* Pacific salmon: ecology and management in western Alaska's populations. *Edited by* C.C. Krueger and C.E. Zimmerman. Am. Fish. Soc., Symp. No. 70, Bethesda, Maryland. pp. 827-850.

Sved JA, Cameron EC, Gilchrist AS (2013) Estimating effective population size from linkage disequilibrium between unlinked loci: theory and application to fruit fly outbreak populations. *PLoS ONE* **8**, e69078.

Takezaki N, Nei M, Tamura K (2010) POPTREE2: software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. *Molecular Biology and Evolution* **27**, 747-752.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2731-2739.

Taylor EB, Harvey S, Pollard S, Volpe J (1997) Postglacial genetic differentiation of reproductive ecotypes of kokanee *Oncorhynchus nerka* in Okanagan Lake, British Columbia. *Molecular Ecology* **6**, 503-517.

Templin WD, Seeb JE, Jasper JR, Barclay AW, Seeb LW (2011) Genetic differentiation of Alaska Chinook salmon: the missing link for migratory studies. *Molecular Ecology Resources* **11**, 226-246.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, *et al.* (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**, 520-526.

Thorgaard GH (1978) Sex-chromosomes in sockeye salmon - y - autosome fusion. *Canadian Journal of Genetics and Cytology* **20**, 349-354.

Thorgaard GH, Allendorf FW, Knudsen KL (1983) Gene-centromere mapping in rainbow trout-high interference over long map distances. *Genetics* **103**, 771-783.

Tonsor SJ (2012) Population genomics and the causes of local differentiation. *Molecular Ecology* **21**, 5393-5395.

Trudel M, Fisher J, Orsi JA, Morris JFT, Thiess ME, Sweeting RM, *et al.* (2009) Distribution and migration of juvenile Chinook salmon derived from coded wire tag recoveries along the continental shelf of western North America. *Transactions of the American Fisheries Society* **138**, 1369-1391.

Tucker S, Trudel M, Welch DW, Candy JR, Morris JFT, Thiess ME, *et al.* (2011) Life history and seasonal stock-specific ocean migration of juvenile Chinook salmon. *Transactions of the American Fisheries Society* **140**, 1101-1119.

Tucker S, Trudel M, Welch DW, Candy JR, Morris JFT, Thiess ME, *et al.* (2009) Seasonal Stock-Specific Migrations of Juvenile Sockeye Salmon along the West Coast of North America: Implications for Growth. *Transactions of the American Fisheries Society* **138**, 1458-1480.

Urawa S, Nagasawa K, Margolis L, Moles A (1998) Stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) in the North Pacific Ocean and Bering Sea by parasite tags. *N. Pac. Anadr. Fish Comm. Bull.* **1**, 199-204.

Utter F (2004) Population genetics, conservation and evolution in salmonids and other widely cultured fishes: some perspectives over six decades. *Reviews in Fish Biology and Fisheries* **14**, 125-144.

Utter F, Hodgins H, Allendorf F (1974) Biochemical genetic studies of fishes: potentialities and limitations. Pages 213-38 in D.C. Malins & J.R. Sargents, eds. Biochemical and biophysical perspectives in marine biology Vol. 1. Academic Press, San Francisco, CA.

Utter F, Ryman N (1993) Genetic markers and mixed stock fisheries. *Fisheries* **18**, 11-21.

Van der Walt JM, Nel LH, Hoelzel AR (2001) Characterization of major histocompatibility complex DRB diversity in the endemic South African antelope *Damaliscus pygargus:* a comparison in two subspecies with different demographic histories. *Molecular Ecology* **10**, 1679-1688.

van Duyvenvoorde HA, Lui JC, Kant SG, Oostdijk W, Gijsbers ACJ, Hoffer MJV, *et al.* (2014) Copy number variants in patients with short stature. *European Journal of Human Genetics* **22**, 602-609.

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

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology* **17**, 4334-4345.

Visscher PM, Thompson R, Haley CS (1996) Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**, 1013-1020.

Waples RK, Seeb LW, Seeb JE (2015) Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Molecular Ecology Resources*. Online early. DOI: 10.1111/1755-0998.12394.

Waples RS (1990a) Conservation genetics of Pacific salmon. 3. Estimating effective population-size. *Journal of Heredity* **81**, 277-289.

Waples RS (1990b) Temporal changes of allele frequency in Pacific salmon - implications for mixed-stock fishery analysis. *Canadian Journal of Fisheries and Aquatic Sciences* **47**, 968-976.

Waples RS (1998) Separating the wheat from the chaff: Patterns of genetic differentiation in high gene flow species. *Journal of Heredity* **89**, 438-450.

Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics* **7**, 167-184.

Waples RS, Dickhoff WW, Hauser L, Ryman N (2008) Six decades of fishery genetics: Taking stock. *Fisheries* **33**, 76-79.

Waples RS, Do C (2010) Linkage disequilibrium estimates of contemporary N_e using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications* **3**, 244-262.

Warheit K, Seeb LW, Templin WD, Seeb JE (2013) Moving GSI into the Next Decade: SNP Coordination for Pacific Salmon Treaty Fisheries. Chinook Technical Committee. Washington Department of Fish and Wildlife, Report FPT 13-09. http://wdfw.wa.gov/publications/01629/wdfw01629.pdf.

Watterson GA (1978) Homozygosity test of neutrality. *Genetics* **88**, 405-417.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. *Evolution* **38**, 1358-1370.

Welch DW, Melnychuk MC, Payne JC, Rechisky EL, Porter AD, Jackson GD, *et al.* (2011) In situ measurement of coastal ocean movements and survival of juvenile Pacific salmon. *Proceedings of the National Academy of Sciences* **108**, 8708-8713.

Williams D, Shelden C (2011) Kogrukluk River salmon studies, 2010 [online]. Alaska Department of Fish and Game, Fishery Data Series No. 11-49. http://www.adfg.alaska.gov/FedAidpdfs/FDS11-49.pdf [accessed January 30, 2012].
Wirgin, II, Waldman JR (1994) What DNA can do for you. *Fisheries* **19**, 16-27.
Wood CC, Bickham JW, Nelson RJ, Foote CJ, Patton JC (2008) Recurrent evolution of life history ecotypes in sockeye salmon: implications for conservation and future evolution. *Evolutionary Applications* **1**, 207-221.
Woody CA (2012) Assessing reliability of Pebble Limited Partnership's salmon escapement studies [online]. Fisheries research and consulting, Anchorage, Alaska. http://www.pebblescience.org/pdfs/Woody_EBD_EscapementFINAL27June2012.pdf [accessed January 30, 2012].
Worm B, Barbier EB, Beaumont N, Duffy JE, Folke C, Halpern BS, *et al.* (2006) Impacts of biodiversity loss on ocean ecosystem services. *Science* **314**, 787-790.
Xu SZ (2003) Theoretical basis of the Beavis effect. *Genetics* **165**, 2259-2268.
Yeaman S (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences* **110**, 1743-1751.
Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection balance. *Evolution* **65**, 1897-1911.
Yoon M, Jin D-H, Abe S (2009) Preliminary estimation of chum salmon stock composition in the Bering Sea and North Pacific Ocean using polymorphic microsatellite DNA markers. *Ichthyological Research* **56**, 37-42.
Zavolokin AV (2009) Forage base of Pacific salmon in the western Bering Sea and adjacent Pacific waters in 2002-2006. *North Pacific Anadromous Fish Commission Bulletin* **5**, 165-172.
Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* **18**, 821-829.
Zueva KJ, Lumme J, Veselov AE, Kent MP, Lien S, Primmer CR (2014) Footprints of directional selection in wild Atlantic salmon populations: evidence for parasite-driven evolution? *PLoS ONE* **9**, e91672.