Effectiveness of managed gene flow to reduce genetic and phenotypic change associated with captive breeding of Chinook salmon

Charles D. Waters

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Kerry-Ann Naish, Chair

Steven Schroder

Jeffrey Hard

Program Authorized to Offer Degree:

School of Aquatic and Fishery Sciences

University of Washington

Abstract

Effectiveness of managed gene flow to reduce genetic and phenotypic change associated with captive breeding of Chinook salmon

Charles D. Waters

Chair of the Supervisory Committee:
Professor Kerry-Ann Naish
School of Aquatic and Fishery Sciences

Captive breeding programs can rebuild depressed populations and aid in the recovery of threatened or endangered species. However, associated genetic and phenotypic changes may decrease the fitness of captive individuals when they are released into the wild and thus reduce restoration success. Genetic changes include loss of genetic diversity and divergence from the wild population, inbreeding, and adaptation to captivity, while changes in fitness traits, which affect both individual reproductive success and population productivity and resilience, may arise from genetic factors and differences between the captive and natural environments. Incorporating wild individuals as broodstock in captive breeding programs, which we refer to here as managed gene flow, is one strategy that may mitigate these potential risks. While this approach has been widely adopted in salmon hatchery management throughout the Pacific Northwest, it has not been empirically tested in a comparative framework over multiple generations.

This dissertation characterizes genetic and phenotypic changes associated with captive breeding using novel genome-wide approaches and explicitly tests the effectiveness of managed gene flow to minimize these changes using two hatchery populations of Chinook salmon, *Oncorhynchus tshawytscha*. The hatchery populations were derived from the same wild population but are now managed as separate lines, one integrated with (i.e. managed gene flow) and one segregated from (i.e. no gene flow) the source stock, and thus provide an ideal system for comparing the two alternative management strategies. I used genomic and phenotypic data spanning five generations to examine the two hatchery populations across a range of measures.

First, I used over 9000 loci to test whether managed gene flow between natural and captive environments, when compared to broodstock segregation, reduced genome-wide divergence from the wild founding population over four generations of captive rearing (Chapters 1-2). Genetic divergence from the source population was minimal in the integrated hatchery line, which implemented managed gene flow by using only naturally-born adults as captive broodstock, but significant in the segregated line, which bred only captive-origin individuals. Estimates of effective number of breeders revealed that the rapid divergence observed in the segregated line was largely attributable to genetic drift. However, we also identified temporally-consistent signatures of adaptive divergence within the segregated line, possibly indicative of domestication selection. The results empirically demonstrated that using managed gene flow for propagating a captive-reared population reduces genetic divergence over the short term compared to one that relies solely on captive-origin parents. The findings also provided insight into the rate at which divergence may occur in integrated and segregated hatchery programs.

Second, I computed genomic-based estimates of pairwise relatedness and individual inbreeding within the integrated and segregated hatchery lines across four generations and

determined if managed gene flow successfully reduced the risks of inbreeding over time (Chapter 3). I also quantified the effect of inbreeding coefficient on eight fitness-related traits that had been measured in returning adults. The segregated line had slight but significantly lower levels of relatedness than the integrated line in the first generation but significantly higher levels in the third and fourth generations. Levels of inbreeding were similar between the two hatchery lines in the first, third, and fourth generations, despite 3- to 27-fold differences in estimates of effective numbers of breeders. However, inbreeding in the segregated line was significantly higher in the second generation. Inbreeding coefficient did not affect fecundity, reproductive effort, return timing, and fork length. In contrast, inbreeding significantly affected spawn timing, weight, condition factor, and daily growth coefficient, although the effects varied by sex, hatchery line, and generation. While the results indicated that managed gene flow may reduce the genetic risks of inbreeding, they also suggested that short-term risks may not be severe in small, segregated hatchery populations. The effects of inbreeding on fitness, however, require further examination, particularly at earlier life stages.

Third, I identified loci associated with six fitness-related traits in adult Chinook salmon using an approach suitable for polygenic traits, Random Forest, and then explored the use of trait-associated loci within a management context, namely, whether they could serve as tools for monitoring the effects of alternative management approaches on genetic change underlying phenotypic traits (Chapter 4). I identified 226 unique loci associated with the six traits. Mapping of these trait-associated loci, gene annotations, and integration of results across multiple studies revealed candidate regions potentially involved in fitness. Genotypes at trait-associated loci were then compared between the integrated and segregated hatchery lines. While no broad scale change was detected between the lines across four generations, there were numerous regions

where trait-associated loci overlapped with signatures of adaptive divergence identified in Chapters 1 and 2. Many regions of overlap, primarily with loci linked to return and spawn timing, were either unique to, or more divergent in, the segregated line, suggesting that these traits may be responding to domestication selection. This chapter is one of the first studies to utilize genomic approaches to demonstrate the effectiveness of a conservation strategy, managed gene flow, on trait-associated – and potentially adaptive – loci.

Last, I combined lessons learned from my analyses in Chapter 4 with information from colleagues and other studies to provide a simple, introductory guide to facilitate the use of Random Forest to identify genotype-phenotype associations in non-model organisms (Chapter 5). The guide first provides an overview of the Random Forest algorithm. Next, steps are described to prepare data for Random Forest, including initial data exploration and the identification of important covariates and possible confounding factors. Advice is then provided on the initiation and optimization of the Random Forest algorithm, along with a summary of methods for interpreting the results and identifying trait-associated, or predictor, loci. Annotated R tutorials are also included to assist users in implementing each step of the algorithm. This guide will hopefully facilitate the use of Random Forest in future ecological and evolutionary studies and contribute to the reporting of accurate and reproducible results.

This dissertation research encompassed a broad perspective to characterize multiple genetic risks of captive breeding using novel genomic approaches, to link these risks to important fitness traits that were measured in adults, and to demonstrate that managed gene flow successfully mitigated potential adverse effects across four generations. By comparing two alternative management strategies, the study provides insight into the range of outcomes that may occur in captive breeding programs, which is highly relevant to risk assessment in realistic

scenarios. Further, the results provide molecular tools to better monitor genetic change in hatchery and wild populations of salmon and further inform "best practices" in hatchery management to support declining wild populations.

The findings also lay the foundation for future research efforts. For example, Chapters 1 and 2 revealed that using 100% natural-origin broodstock reduced genetic divergence across four generations. However, removing individuals from the wild for broodstock may also have genetic and demographic costs to the wild population, particularly since most populations supplemented by hatcheries are already in decline. A future study could use experimental populations to explicitly examine the effects of different levels of managed gene flow on divergence. Such information would enable conservation hatcheries to minimize genetic risks while also reducing potential costs of broodstock removal. Chapter 3 examined the genetic and phenotypic risks of inbreeding in returning adults. Yet, inbred fish may have significantly higher levels of mortality in the marine environment than non-inbred fish. Thus, the examination of inbreeding in adults may be inherently biased. A future study could sample both juveniles and returning adults from each hatchery line to determine if levels of inbreeding differ by life stage and if selection in the marine environment mitigates the risks of inbreeding in hatcheries. The identification of loci associated with six key traits in Chapter 4 is a first step towards characterizing the functional genetic basis of fitness in Chinook salmon. The regions where trait-associated and outlier loci overlapped will also provide useful starting points for future sequencing efforts that aim to identify the specific genes responding to domestication selection. In addition to specific research questions, this body of work demonstrates the utility of genomic-based approaches in conservation monitoring and therefore provides an overall framework for other studies that aim to integrate genomics with the management of captive and wild populations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation is the culmination of seven years of sample collections, laboratory work, bioinformatics analyses, and interpretations. It not only represents my efforts, but those of many other colleagues and collaborators. This research would not have been possible without them, nor would I be the person and scientist that I am today without their guidance and support. I would like to take this opportunity to thank these individuals.

I would first like to thank my advisor, Kerry Naish. Kerry has always been supportive of me, my research, and my life outside of graduate school. She has taught me to think critically about every analysis, to interpret my results in a biological context, and countless other things. Without fail, Kerry always made time for me whenever I had a question, no matter how busy she was with other responsibilities. She also supported my love of Alaska by letting me return to the Little Port Walter Marine Research Station every summer to assist with their research efforts. She is truly an outstanding scientist, teacher, and mentor, and I cannot thank her enough.

I would also like to thank the other members (past and present) of the Molecular Ecology Research Lab for their friendship, support, and assistance. Lorenz Hauser provided valuable input with many genetic analyses and questions that arose during my project. Isadora Jimenez-Hidalgo taught me everything that I know about laboratory work, assisted me with sample processing, and provided guidance throughout my entire PhD. Marine Brieuc and Dan Drinan also gave their valuable time to assist me with laboratory work and data analysis, particularly when I was just beginning graduate school. The other members of our lab family – Jocelyn Lin, Shannon O'Brien, Daniel Peterson, Miyako Kodama, Eleni Petrou, Natalie Lowell, Molly Jackson, Mary Fisher, and Samuel May – supported me, both scientifically and personally, throughout this process.

In addition to Kerry, the other members of my committee – Jeff Hard, Lorenz Hauser, Maureen Purcell, Ryan Kelly, Steve Schroder, and Tim Thornton – were extremely supportive of my research and timeline and provided insightful comments on my projects. Jeff Hard, in particular, assisted me with various analyses on a regular basis and was kind enough to serve as my sponsor for the NMFS-Sea Grant Fellowship in Population Dynamics. Tim Thornton, in addition to his valuable research advice, saved the day by providing me with funding for my last quarter at the university. Thank you to Jeff, Lorenz, Maureen, Ryan, Steve, and Tim. It has been a privilege to work with each of you.

My research relied on an extraordinary data set from the Cle Elum Supplementation and Research Facility (CESRF). I would therefore like to thank all Yakama Nation and Washington Department of Fish and Wildlife personnel involved in the initiation and propagation of the hatchery lines at CESRF. I would specifically like to thank Dave Fast, Bill Bosch, and Curt Knudsen for their valuable input on my research and for answering my countless questions.

I would like to thank my funding sources for supporting the various projects of my dissertation, including the UW School of Aquatic and Fishery Sciences, the Hall Conservation Genetics Research Award, Washington Sea Grant, and the NMFS-Sea Grant Fellowship in Population and Ecosystem Dynamics.

Lastly, I would like to thank my family and girlfriend. Mom and Dad, thank you for raising three intelligent, considerate boys and for showing us the value of hard work. I credit you for why I became a scientist. Jimmy and Arelly, thank you for encouraging me from afar. Richard and Kim, thank you for your support and for the many free meals! And to my girlfriend Emma, thank you for listening to my highs and lows every day and for your continued support. I would not have been able to finish this degree without you.

GENERAL INTRODUCTION

Captive breeding programs have long been used to augment economically valuable species, supplement declining populations, and conserve threatened or endangered species (Frankham 2008; Fraser 2008; Lorenzen *et al.* 2010; Hayward 2011; Witzenberger & Hochkirch 2011). These programs can be effective management tools because they increase survival of individuals when in captivity. However, their efficacy to enhance natural populations remains controversial among scientists, managers, and the public (Snyder *et al.* 1996; Bowkett 2009), in part because captive breeding may cause genetic and phenotypic changes that may decrease the fitness of individuals after they are released into the wild. Such changes have been documented in a variety of taxa, from invertebrates to apex predators (Frankham 2008; Jule *et al.* 2008), and may be caused by multiple factors. First, rearing conditions may result in the relaxation of natural selective pressures and impose artificial selection favoring individuals suited for captivity instead of the natural environment, also known as domestication selection (Frankham 2008). The impacts of domestication selection on fitness have been studied across an array of species (e.g. Lewis & Thomas 2001; Shuster *et al.* 2005; Tian *et al.* 2009; Rubin *et al.* 2010; Weber *et al.* 2010). Second, genetic drift and inbreeding may reduce genetic diversity within captive populations and negatively affect fitness through inbreeding depression (Lacy 1993; Madsen *et al.* 1996; Joron & Brakefield 2003; Fraser 2008; Naish *et al.* 2008; Hammerly *et al.* 2013; Willoughby *et al.* 2015). As a result of these factors, captive and wild populations may become genetically divergent over time, and individuals raised in captivity for enhancement purposes may be less effective in contributing to the population's long-term viability than their wild-born counterparts (Berejikian & Ford 2004; Fraser 2008; Jule *et al.* 2008; Laikre *et al.* 2010).

One form of captive breeding – supportive breeding – might improve the success of recovery efforts because it avoids captive rearing for the full life cycle (Ryman & Laikre 1991). In such programs, a fraction of a population is taken into captivity for reproduction, and their progeny are released back into the natural environment to join wild-born conspecifics. Supportive breeding using hatcheries has become an important component of many recovery plans for anadromous Pacific salmon on the West Coast of North America, where wild populations have steadily declined due to anthropogenic disturbances (National Research Council 1996; Gustafson *et al.* 2007). However, despite being exposed to the captive environment for only part of their life cycle, hatchery-reared salmon are still at risk of genetic and phenotypic change. For example, domestication selection may occur because the fish experience high rearing densities, no predation, no sexual selection, unnatural flow regimes, an unlimited supply of food, and many other artificial conditions. Domestication selection has been the focus of most hatchery research recently, and its effects on fitness and fitness-related traits in adults and juveniles include decreased reproductive success (Araki *et al.* 2007; Christie *et al.* 2012; Milot *et al.* 2013; Christie *et al.* 2014), reduced response or increased vulnerability to predation (Fritts *et al.* 2007), lower survival (McGinnity *et al.* 2003), differences in growth rate and morphology (McGinnity *et al.* 2003; Busack *et al.* 2007), and reduced competitive abilities (Metcalfe *et al.* 2003).

Genetic drift and inbreeding may also reduce the fitness of hatchery-reared salmon, as these effects can be amplified in hatchery populations due to limited rearing capacity and numbers of breeders, variance in reproductive success, and unequal sex ratios (Wang *et al.* 2002; Naish *et al.* 2008). Indeed, numerous studies of experimental and captive populations observed adverse fitness effects that were due to inbreeding (Wang *et al.* 2002; Naish *et al.* 2013).

Notably, marine survival of rainbow and steelhead trout released into the wild was 78% lower in inbred families than in non-inbred controls (Thrower & Hard 2009), providing evidence that inbreeding can have serious demographic consequences for salmonid populations.

While the risks of hatchery rearing are widely acknowledged, several key questions remain unanswered. For example, the multigenerational effects of genetic and phenotypic change in hatchery fish to supplemented wild populations are unclear. The captive environment typically facilitates the reproduction of most individuals regardless of their fitness in the natural environment (Neff *et al.* 2011), but subsequent selection on F_1 offspring released in the wild might remove maladaptive individuals (Baskett & Waples 2013) and reduce their effects on the fitness of the supplemented population. Alternatively, ongoing broodstock collection and reduced fitness of F_1 hatchery offspring might have long-term negative demographic and genetic effects on the wild component (Ryman & Laikre 1991; Ford 2002; Baskett & Waples 2013). Furthermore, measuring the relative fitness of the offspring of hatchery fish that spawned in the wild (the F_2 generation, born in the wild) can be challenging, because matings that produce no offspring are typically unobserved (Araki *et al.* 2009), and subsequent analyses can have reduced power to detect an effect (Christie *et al.* 2014). Information on the multigenerational effects of hatchery fish on both hatchery and natural populations would help policy-makers weigh the relative risks and benefits of operating enhancement programs for extended periods of time.

The effects of hatchery management on potentially adaptive genetic variation are also unknown. Comparisons of fitness between hatchery- and wild-origin individuals, for example, do not reveal how captive rearing affects genetic variation underlying fitness traits or which traits are most susceptible to domestication selection. Many key traits in salmon – length, weight, and dates of return to freshwater and maturation – are correlated with individual fitness (e.g. Thorpe

et al. 1984; Schroder *et al.* 2010; Kodama *et al.* 2012). In addition, these traits have significant additive genetic variation on which selection can act (Hard 2004; Carlson & Seamons 2008) and can differ between hatchery and wild populations (e.g. Ford *et al.* 2006; Knudsen *et al.* 2006; Hoffnagle *et al.* 2008). Therefore, studying the genetic basis of specific traits may provide a better understanding of how domestication selection acts and, in turn, reveal possible mechanisms underlying the reduced fitness of hatchery-origin fish after they are released into the wild.

Despite such risks, captive breeding may be one of the only feasible tools for the conservation and recovery of declining populations and species, particularly as rates of extinction increase (Pimm *et al.* 2006; Williams & Hoffman 2009; Burkhead 2012). Therefore, there is a pressing need to address these and other unanswered questions. Furthermore, approaches that minimize genetic and phenotypic risks of captive breeding over sufficient generations for recovery to occur must be developed and tested (Williams & Hoffman 2009). One alternative management approach to mitigate potential risks is the intentional promotion of gene flow from wild to captive populations, or managed gene flow. Theoretical studies suggest that this strategy may mitigate the risks of domestication selection, inbreeding, and genetic drift associated with captive rearing (Lynch & O'Hely 2001; Duchesne & Bernatchez 2002; Ford 2002; Baskett & Waples 2013). In fact, managed gene flow in the form of integrated hatchery management has been adopted in many hatchery programs in the Pacific Northwest due to recent reform efforts (Mobrand *et al.* 2005; Paquet *et al.* 2011). Yet, the effectiveness of managed gene flow to reduce genetic and concomitant phenotypic changes that could occur from captive rearing has never been empirically evaluated.

A powerful way to evaluate the effectiveness of managed gene flow would be to study a system that is comparative in nature: namely, one that maintains a captive line that has been deliberately kept separate from wild individuals, and a second line where wild individuals are used exclusively as captive broodstock. Variation in each generation of the captive lines could be compared to the original founding population using genome-wide surveys, which offer a means to broadly estimate rates of genetic change. Levels of inbreeding and relatedness could also be measured within the two captive lines over time to compare potential risks of inbreeding depression. Similarly, genetic variation at loci linked to fitness traits could be compared to determine how the different management approaches affect potentially adaptive genetic variation, and which traits may be most susceptible to domestication selection within the hatchery. Such comparisons would provide guidelines for "best practices" in hatchery management, refine understanding on the optimal longevity of supportive breeding programs, and deliver insight on the range of possible outcomes of managed gene flow.

A hatchery program at the Cle Elum Supplementation and Research Facility (CESRF) was initiated in 1997 in response to declining anadromous spring Chinook salmon returning to the Yakima River, a tributary of the Columbia River USA. Wild adults were collected for founding broodstock from the upper Yakima River population from 1997–2002. Beginning with brood year 2002, the hatchery population was divided into a segregated (SEG) line, which was not allowed to interbreed with the source population, and an integrated (INT) line, which was allowed to spawn in the river. All first generation hatchery fish from the integrated line were allowed to spawn naturally. The proportion of hatchery fish from the integrated line spawning in the natural environment has varied between 0.2 and 0.76 (mean=0.56) from 2001–2013 (Fast *et al.* 2015). There is indirect evidence that these hatchery fish successfully contributed to the

natural population: hatchery and natural origin fish had similar distributions on the spawning grounds (Dittman *et al.* 2010), and redd (nest) abundance and spatial distribution has increased (Fast *et al.* 2015). Fish from the integrated and segregated hatchery lines are raised in the same facility but are differentially marked for external identification. The two lines have been reared for four generations, and DNA samples have been collected from every adult fish used as broodstock since the inception of the program (fish are intercepted at Roza Dam). Thus, this system provides an extraordinary opportunity to improve our understanding of genetic and phenotypic changes due to captive breeding and to experimentally evaluate efforts to reduce these risks.

This dissertation aims to better characterize genetic and phenotypic changes associated with captive breeding using genome-wide approaches, and to explicitly test the effectiveness of managed gene flow to minimize these risks using two hatchery populations of Chinook salmon, *Oncorhynchus tshawytscha*. The hatchery populations were derived from the same source but now maintain contrasting levels of gene flow with the natural population. I use genomic and phenotypic data across five generations to compare the hatchery populations across a range of measures. Specifically, I 1) test whether managed gene flow reduces overall genetic divergence (Chapters 1-2) and levels of inbreeding and inbreeding depression (Chapter 3) due to captive breeding, 2) characterize the genetic architecture of adult fitness traits and determine if they may be responding to genetic change in captivity (Chapter 4), and 4) combine lessons learned from my own analyses with information from colleagues and other studies to provide a simple, introductory guide to facilitate the use of Random Forest to identify genotype-phenotype associations in non-model organisms (Chapter 5). The findings of my doctoral research will

inform management of hatchery populations and aid the restoration of declining populations across an array of species.

REFERENCES

Araki H, Ardren WR, Olsen E, Cooper B, Blouin MS (2007) Reproductive success of captive-bred steelhead trout in the wild: Evaluation of three hatchery programs in the hood river. *Conservation Biology*, **21**, 181-190.

Araki H, Cooper B, Blouin MS (2009) Carry-over effect of captive breeding reduces reproductive fitness of wild-born descendants in the wild. *Biology Letters*, **5**, 621-624.

Baskett ML, Waples RS (2013) Evaluating alternative strategies for minimizing unintended fitness consequences of cultured individuals on wild populations. *Conservation Biology*, **27**, 83-94.

Berejikian BA, Ford MJ (2004) Review of relative fitness of hatchery and natural salmon. (ed. Commerce USD), p. 28. NOAA Tech Memo. NMFS-NWFSC.

Bowkett AE (2009) Recent captive-breeding proposals and the return of the ark concept to global species conservation. *Conservation Biology*, **23**, 773-776.

Burkhead NM (2012) Extinction rates in north american freshwater fishes, 1900-2010. *Bioscience*, **62**, 798-808.

Busack C, Knudsen CM, Hart G, Huffman P (2007) Morphological differences between adult wild and first-generation hatchery upper yakima river spring chinook salmon. *Transactions of the American Fisheries Society*, **136**, 1076-1087.

Carlson SM, Seamons TR (2008) A review of quantitative genetic components of fitness in salmonids: Implications for adaptation to future change. *Evolutionary Applications*, **1**, 222-238.

Christie MR, Ford MJ, Blouin MS (2014) On the reproductive success of early-generation hatchery fish in the wild. *Evolutionary Applications*, **7**, 883-896.

Christie MR, Marine ML, French RA, Blouin MS (2012) Genetic adaptation to captivity can occur in a single generation. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 238-242.

Dittman AH, May D, Larsen DA, Moser ML, Johnston M, Fast D (2010) Homing and spawning site selection by supplemented hatchery- and natural-origin yakima river spring chinook salmon. *Transactions of the American Fisheries Society*, **139**, 1014-1028.

Duchesne P, Bernatchez L (2002) An analytical investigation of the dynamics of inbreeding in multi-generation supportive breeding. *Conservation Genetics*, **3**, 47-60.

Fast DE, Bosch WJ, Johnston MV *et al.* (2015) A synthesis of findings from an integrated hatchery program after three generations of spawning in the natural environment. *North American Journal of Aquaculture*, **77**, 377-395.

Ford MJ (2002) Selection in captivity during supportive breeding may reduce fitness in the wild. *Conservation Biology*, **16**, 815-825.

Ford MJ, Fuss H, Boelts B, LaHood E, Hard J, Miller J (2006) Changes in run timing and natural smolt production in a naturally spawning coho salmon (O*ncorhynchus kisutch*)

population after 60 years of intensive hatchery supplementation. *Canadian Journal of Fisheries and Aquatic Sciences*, **63**, 2343-2355.
Frankham R (2008) Genetic adaptation to captivity in species conservation programs. *Molecular Ecology*, **17**, 325-333.
Fraser DJ (2008) How well can captive breeding programs conserve biodiversity? A review of salmonids. *Evolutionary Applications*, **1**, 535-586.
Fritts AL, Scott JL, Pearsons TN (2007) The effects of domestication on the relative vulnerability of hatchery and wild origin spring chinook salmon (*Oncorhynchus tshawytscha*) to predation. *Canadian Journal of Fisheries and Aquatic Sciences*, **64**, 813-818.
Gustafson RG, Waples RS, Myers JM *et al.* (2007) Pacific salmon extinctions: Quantifying lost and remaining diversity. *Conservation Biology*, **21**, 1009-1020.
Hammerly SC, Morrow ME, Johnson JA (2013) A comparison of pedigree- and DNA-based measures for identifying inbreeding depression in the critically endangered attwater's prairie-chicken. *Molecular Ecology*, **22**, 5313-5328.
Hard JJ (2004) Evolution of chinook salmon life history under size-selective harvest. In: *Evolution illuminated: Salmon and their relatives* (eds. Hendry A, Stearns S), pp. 315-337. Oxford University Press.
Hayward MW (2011) Using the iucn red list to determine effective conservation strategies. *Biodiversity and Conservation*, **20**, 2563-2573.
Hoffnagle TL, Carmichael RW, Frenyea KA, Keniry PJ (2008) Run timing, spawn timing, and spawning distribution of hatchery- and natural-origin spring chinook salmon in the imnaha river, oregon. *North American Journal of Fisheries Management*, **28**, 148-164.
Joron M, Brakefield PM (2003) Captivity masks inbreeding effects on male mating success in butterflies. *Nature*, **424**, 191-194.
Jule KR, Leaver LA, Lea SEG (2008) The effects of captive experience on reintroduction survival in carnivores: A review and analysis. *Biological Conservation*, **141**, 355-363.
Knudsen CM, Schroder SL, Busack CA *et al.* (2006) Comparison of life history traits between first-generation hatchery and wild upper yakima river spring chinook salmon. *Transactions of the American Fisheries Society*, **135**, 1130-1144.
Kodama M, Hard JJ, Naish KA (2012) Temporal variation in selection on body length and date of return in a wild population of coho salmon, *Oncorhynchus kisutch*. *Bmc Evolutionary Biology*, **12**, 12.
Lacy RC (1993) Impacts of inbreeding in natural and captive populations of vertebrates-implications for conservation. *Perspectives in Biology and Medicine*, **36**, 480-496.
Laikre L, Schwartz MK, Waples RS, Ryman N, Ge MWG (2010) Compromising genetic diversity in the wild: Unmonitored large-scale release of plants and animals. *Trends in Ecology & Evolution*, **25**, 520-529.
Lewis OT, Thomas CD (2001) Adaptations to captivity in the butterfly pieris brassicae (l.) and the implications for ex situ conservation. *Journal of Insect Conservation*, **5**, 55-63.
Lorenzen K, Leber KM, Blankenship HL (2010) Responsible approach to marine stock enhancement: An update. *Reviews in Fisheries Science*, **18**, 189-210.
Lynch M, O'Hely M (2001) Captive breeding and the genetic fitness of natural populations. *Conservation Genetics*, **2**, 363-378.
Madsen T, Stille B, Shine R (1996) Inbreeding depression in an isolated population of adders vipera berus. *Biological Conservation*, **75**, 113-118.

McGinnity P, Prodohl P, Ferguson K *et al.* (2003) Fitness reduction and potential extinction of wild populations of atlantic salmon, *salmo salar*, as a result of interactions with escaped farm salmon. *Proceedings of the Royal Society B-Biological Sciences*, **270**, 2443-2450.

Metcalfe NB, Valdimarsson SK, Morgan IJ (2003) The relative roles of domestication, rearing environment, prior residence and body size in deciding territorial contests between hatchery and wild juvenile salmon. *Journal of Applied Ecology*, **40**, 535-544.

Milot E, Perrier C, Papillon L, Dodson JJ, Bernatchez L (2013) Reduced fitness of atlantic salmon released in the wild after one generation of captive breeding. *Evolutionary Applications*, **6**, 472-485.

Mobrand LE, Barr J, Blankenship L *et al.* (2005) Hatchery reform in washington state: Principles and emerging issues. *Fisheries*, **30**, 11-23.

Naish KA, Seamons TR, Dauer MB, Hauser L, Quinn TP (2013) Relationship between effective population size, inbreeding and adult fitness-related traits in a steelhead (*oncorhynchus mykiss*) population released in the wild. *Molecular Ecology*, **22**, 1295-1309.

Naish KA, Taylor JE, Levin PS *et al.* (2008) An evaluation of the effects of conservation and fishery enhancement hatcheries on wild populations of salmon. In: *Advances in marine biology*, pp. 61-194. Elsevier Academic Press Inc, San Diego.

National Research Council (1996) *Upstream: Salmon and society in the pacific northwest* National Academy Press, Washington, D.C.

Neff BD, Garner SR, Pitcher TE (2011) Conservation and enhancement of wild fish populations: Preserving genetic quality versus genetic diversity. *Canadian Journal of Fisheries and Aquatic Sciences*, **68**, 1139-1154.

Paquet PJ, Flagg T, Appleby A *et al.* (2011) Hatcheries, conservation, and sustainable fisheries-achieving multiple goals: Results of the hatchery scientific review group's columbia river basin review. *Fisheries*, **36**, 547-561.

Pimm S, Raven P, Peterson A, Sekercioglu CH, Ehrlich PR (2006) Human impacts on the rates of recent, present, and future bird extinctions. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 10941-10946.

Rubin CJ, Zody MC, Eriksson J *et al.* (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, **464**, 587-U145.

Ryman N, Laikre L (1991) Effects of supportive breeding on the genetically effective population size. *Conservation Biology*, **5**, 325-329.

Schroder SL, Knudsen CM, Pearsons TN *et al.* (2010) Behavior and breeding success of wild and first-generation hatchery male spring chinook salmon spawning in an artificial stream. *Transactions of the American Fisheries Society*, **139**, 989-1003.

Shuster SM, Miller MP, Lang BK, Zorich N, Huynh L, Keim P (2005) The effects of controlled propagation on an endangered species: Genetic differentiation and divergence in body size among native and captive populations of the socorro isopod (crustacea : Flabellifera). *Conservation Genetics*, **6**, 355-368.

Snyder NFR, Derrickson SR, Beissinger SR *et al.* (1996) Limitations of captive breeding in endangered species recovery. *Conservation Biology*, **10**, 338-348.

Thorpe JE, Miles MS, Keay DS (1984) Developmental rate, fecundity and egg size in atlantic salmon, *salmo salar*. *Aquaculture*, **43**, 289-305.

Thrower FP, Hard JJ (2009) Effects of a single event of close inbreeding on growth and survival in steelhead. *Conservation Genetics*, **10**, 1299-1307.

Tian F, Stevens NM, Buckler ES (2009) Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9979-9986.

Wang SZ, Hard J, Utter F (2002) Salmonid inbreeding: A review. *Reviews in Fish Biology and Fisheries*, **11**, 301-319.

Weber KP, De S, Kozarewa I, Turner DJ, Babu MM, de Bono M (2010) Whole genome sequencing highlights genetic changes associated with laboratory domestication of c. Elegans. *PLoS ONE*, **5**, 10.

Williams SE, Hoffman EA (2009) Minimizing genetic adaptation in captive breeding programs: A review. *Biological Conservation*, **142**, 2388-2400.

Willoughby JR, Fernandez NB, Lamb MC, Ivy JA, Lacy RC, DeWoody JA (2015) The impacts of inbreeding, drift and selection on genetic diversity in captive breeding populations. *Molecular Ecology*, **24**, 98-110.

Witzenberger KA, Hochkirch A (2011) Ex situ conservation genetics: A review of molecular studies on the genetic consequences of captive breeding programmes for endangered animal species. *Biodiversity and Conservation*, **20**, 1843-1861.

Chapter 1. Effectiveness of managed gene flow in reducing genetic divergence associated with captive breeding[1]

1.1 ABSTRACT

Captive breeding has the potential to rebuild depressed populations. However, associated genetic changes may decrease restoration success and negatively affect the adaptive potential of the entire population. Thus, approaches that minimize genetic risks should be tested in a comparative framework over multiple generations. Genetic diversity in two captive-reared lines of a species of conservation interest, Chinook salmon (*Oncorhynchus tshawytscha*), was surveyed across three generations using genome-wide approaches. Genetic divergence from the source population was minimal in an integrated line, which implemented managed gene flow by using only naturally-born adults as captive broodstock, but significant in a segregated line, which bred only captive-origin individuals. Estimates of effective number of breeders revealed that the rapid divergence observed in the latter was largely attributable to genetic drift. Three independent tests for signatures of adaptive divergence also identified temporal change within the segregated line, possibly indicating domestication selection. The results empirically demonstrate that using managed gene flow for propagating a captive-reared population reduces genetic divergence over the short term compared to one that relies solely on captive-origin parents. These findings complement existing studies of captive breeding, which typically focus on a single management strategy and examine the fitness of one or two generations.

[1] This chapter has been published as Effectiveness of managed gene flow in reducing genetic divergence associated with captive breeding, Waters CD, Hard JJ, Brieuc MSO, Fast DE, Warheit KI, Waples RS, Knudsen CM, Bosch WJ, and Naish KA (2015), *Evolutionary Applications* **8**, 956-971.

1.2 INTRODUCTION

The genetic risks associated with captive breeding are widely recognized. Rearing conditions may result in the relaxation of natural selective pressures and impose artificial selection favoring individuals suited for captivity instead of the natural environment, also known as domestication selection (Frankham 2008). Additionally, genetic drift and inbreeding may reduce genetic diversity within captive populations (Fraser 2008; Naish *et al.* 2008). Thus, captive and wild populations may become genetically divergent over time, and individuals raised in captivity for supplementation purposes may be less effective in contributing to the population's long-term viability than their wild-born counterparts (Fraser 2008; Jule *et al.* 2008; Laikre *et al.* 2010). Yet, captive breeding may be the only feasible tool for species recovery and conservation, particularly as rates of extinction increase (Pimm *et al.* 2006; Burkhead 2012). Therefore, there is a pressing need to develop and test approaches that minimize these risks over sufficient generations for population recovery to occur (Williams & Hoffman 2009). Theoretical treatments (Lynch and O'Hely 2001; Duchesne & Bernatchez 2002; Ford 2002) suggest that intentional promotion of gene flow from wild to captive populations may mitigate genetic divergence caused by genetic drift, inbreeding, and selection during captive breeding (Mobrand *et al.* 2005; Frankham 2008; Paquet *et al.* 2011). This strategy, which we refer to here as managed gene flow, should be investigated in empirical settings to determine whether the approach reduces genetic changes that could adversely affect fitness and limit restoration success.

Supportive breeding is a form of captive breeding intended to enhance the sizes of populations in their natural environment (Ryman & Laikre 1991). In such programs, a fraction of a population is taken into captivity for reproduction, and their progeny are released back into the

natural environment to join wild-born conspecifics. This method avoids captive rearing for the full life cycle and might aid *in-situ* population recovery if captive offspring have high reintroduction success. Supportive breeding using hatcheries has become an important component of many recovery plans for anadromous Pacific salmon on the West Coast of North America, where wild populations have steadily declined due to anthropogenic disturbances (National Research Council 1996). Historically, many hatcheries used returning hatchery fish for broodstock, but potential divergence of these fish from the wild population might result in decreased fitness following interbreeding in the natural environment (Busack & Currens 1995; Campton 1995). Recent management reforms aimed at mitigating negative effects of captive rearing have led to the widespread use of locally derived broodstock and implementation of managed gene flow between the hatchery and naturally-derived components (Mobrand *et al.* 2005; Paquet *et al.* 2011).

There are several concerted research efforts aimed at understanding the impacts of supportive breeding on wild populations and the effectiveness of recent hatchery reforms. Pedigree-based studies have provided evidence of reduced reproductive success of hatchery origin fish spawning in the natural environment relative to that of wild fish in the same cohort (Araki *et al.* 2007; Milot *et al.* 2013; Christie *et al.* 2014). However, the significance of reduced fitness of first generation hatchery fish to the future viability of supported populations is unclear. The captive environment typically facilitates the reproduction of most individuals regardless of their fitness in the natural environment (Neff *et al.* 2011), but subsequent selection on F_1 offspring released in the wild might remove maladaptive individuals (Baskett & Waples 2013) and reduce their effects on population fitness. Alternatively, ongoing broodstock collection and reduced fitness of F_1 hatchery offspring might have long-term negative demographic and genetic

effects on the wild component (Ryman & Laikre 1991; Ford 2002; Baskett & Waples 2013). It is also challenging to measure the relative fitness of the offspring of hatchery fish that spawned in the wild (the F_2 generation, born in the wild) because matings that produce no offspring are typically unobserved (Araki *et al.* 2009), and subsequent analyses can have reduced power to detect an effect (Christie *et al.* 2014). Finally, many pedigree studies are necessarily opportunistic because they often rely on established systems, and specific outcomes might depend on local management objectives. It is therefore important to develop complementary experimental methods that examine the consequences of captive rearing on genetic diversity in hatchery-produced fishes, and investigate the potential for managed gene flow to reduce their genetic divergence from wild fish.

A powerful way to evaluate the use of managed gene flow in supportive breeding would be to study a system that is comparative in nature: namely, one that maintains a captive line that has been deliberately kept separate from wild individuals, and a second line where wild individuals are used exclusively as captive broodstock. Variation in each generation of the captive lines could be compared to the original founding population using genome-wide surveys, which offer a means to broadly estimate rates of genetic change. Such a comparison would deliver insight on the range of possible outcomes of managed gene flow and refine understanding on the optimal longevity of supportive breeding programs.

A hatchery program at the Cle Elum Supplementation and Research Facility (CESRF) was initiated in 1997 in response to declining anadromous spring Chinook salmon returning to the Yakima River, a tributary of the Columbia River USA (Figure 1.1). Wild adults were collected for founding broodstock from the upper Yakima River population from 1997–2002. Beginning with brood year 2002, the hatchery population was divided into a segregated (SEG)

line, which was not allowed to interbreed with the source population, and an integrated (INT) line, which was allowed to spawn in the river (Figure 1.2). All first generation hatchery fish from the integrated line were allowed to spawn naturally. The proportion of hatchery fish from the integrated line spawning in the natural environment has varied between 0.2 and 0.76 (mean=0.56) from 2001–2013 (Fast *et al.* 2015). There is indirect evidence that these hatchery fish successfully contributed to the natural population: hatchery and natural origin fish had similar distributions on the spawning grounds (Dittman *et al.* 2010), and redd (nest) abundance and spatial distribution has increased (Fast *et al.* 2015). Fish from the integrated and segregated hatchery lines are raised in the same facility but are differentially marked for external identification. The two lines have been reared for three generations, and DNA samples have been collected from every adult fish used as broodstock since the inception of the program (fish are intercepted at Roza Dam, Figure 1.1). Thus, this system provides an extraordinary opportunity to experimentally evaluate efforts to reduce genetic change in captive management.

The aim of our study was to determine whether managed gene flow between natural and captive environments was effective at reducing genetic divergence relative to a wild founding population over three generations in a Chinook salmon supportive breeding program. We used a population genomic approach to survey genetic variation at 9410 polymorphic restriction site-associated DNA (RAD) markers, including 4405 markers that were anchored to a dense linkage map representing all 34 chromosomes. These markers were first used to test whether genetic divergence occurred in the integrated and segregated hatchery lines over each generation since founding, and to determine whether managed gene flow was effective at reducing differentiation. We also investigated potential mechanisms underlying observed changes. Changes due to genetic drift were examined using estimates of effective number of breeders. Molecular signatures of

adaptive divergence, possibly indicative of domestication selection, were detected using multiple outlier analyses. By combining these results, we empirically validated an approach aimed at minimizing potential deleterious genetic effects of supportive breeding, and identified possible causes that might lead to reduced fitness of captive-reared individuals.

1.3 METHODS

Study system

The Yakima River is a tributary of the Columbia River located in south-central and eastern Washington State, USA (Figure 1.1). The basin is home to three genetically distinct populations of spring, stream-type Chinook salmon (*Oncorhynchus tshawytscha*) in the upper Yakima River, the Naches River, and the American River (Figure 1.1; Busack & Marshall 1991). Adult Chinook salmon return to the basin in spring, spawn in fall, and their offspring spend an entire year in freshwater before migrating to the ocean.

Annual returns of anadromous wild spring Chinook salmon to the Yakima River averaged approximately 1,600 individuals in the 1980s and 1990s (Figure 1.3). These population sizes represent a decline of 98% from estimates of historical returns (ref. in Lichatowich & Mobrand 1995). The hatchery program at CESRF was explicitly designed to test whether supportive breeding could increase harvest and production in the upper Yakima River Chinook population while minimizing ecological and genetic risks associated with captive rearing (RASP 1992).

Returning wild adults from the upper Yakima population were collected for broodstock from 1997–2002 as they passed the Roza Dam Adult Monitoring Facility (Roza Dam, Figure 1.1). Adults were spawned at CESRF, and eggs and fry were reared at the facility for

approximately 16 months. Juveniles were then transferred to three acclimation sites, which were designed to limit returns to the hatchery facility, expand the spatial influence of supplementation efforts, and permit related research on homing (Figure 1.1). Following an acclimation period of two months, fish were allowed to volitionally begin their migration to the ocean. Approximately 80-90% of upper Yakima River Chinook spend 2 years in the ocean and return to freshwater at age 4 to reproduce (Knudsen *et al.* 2006).

First generation hatchery fish began returning as adults in large numbers in 2001 and were allowed to spawn naturally. In 2002, CESRF spawned both wild and returning hatchery origin adults to create the segregated (SEG) and integrated (INT) hatchery lines (Figure 1.2). The segregated line relied solely on returning hatchery origin adults for broodstock, and no fish were allowed to spawn naturally. In contrast, broodstock for the integrated line comprised only natural origin fish, and all returning adults were free to spawn in the river. Natural origin fish are those from the upper Yakima River population; however, beginning in 2005, returning adults were no longer 100% "wild" due to possible influence from naturally spawning hatchery fish (Figure 1.2).

Sample collection

Tissue samples for DNA (operculum or axillary process) were collected from all fish during spawning at CESRF by facility staff or the Washington Department of Fish and Wildlife and stored in 100% ethanol in Olympia, WA. We sub-sampled tissues (N=681) from the 1998 wild founders (2^{nd} founding year; P_1 Founders) and hatchery brood years 2002 (F_1 Wild and F_1 Hatchery), 2006 (F_2 INT and F_2 SEG), and 2010 (F_3 INT and F_3 SEG; Figure 1.2). Since most individuals mature at age 4, these brood years represent four generations of adults. We

considered each generation of each hatchery line as separate "populations" for the purpose of this study.

DNA Sequencing and Genotyping

DNA was extracted using DNeasy Blood & Tissue kits (Qiagen, Valencia, CA) following the animal tissue protocol. Restriction site-associated (RAD) libraries (Baird *et al.* 2008) were prepared, with 24-36 individuals per lane, using the restriction enzyme *Sbf1* and sequenced using the Illumina HiSeq 2000 platform.

RAD sequences were processed with *Stacks* (v. 1.09, Catchen *et al.* 2013). Reads were demultiplexed and trimmed to 74 base pairs in length because sequencing errors increased after that length. Samples that showed signs of contamination were removed from the analysis. All reads were then aligned to a reference database for Chinook salmon comprising 48528 non-duplicated RAD loci, including 7146 mapped loci (Brieuc *et al.* 2014), using the "best" option in *Bowtie* (v. 0.12.8, Langmead *et al.* 2009) and allowing up to 3 nucleotide mismatches. If a read aligned to multiple loci in the baseline, then those loci were omitted from downstream analyses. Following the *Bowtie* alignment, monomorphic and polymorphic loci were identified in *Stacks* using the bounded-error SNP calling model with default error rates and a minimum stack depth of 10 reads. Loci were then filtered for those with two alleles. In an effort to correct for bias in genotype calls due to differences in read depth between two alleles at a locus, all bi-allelic loci were re-genotyped for each individual with a custom Python script. This script designates a genotype as heterozygous if both alleles had a minimum depth of two and a combined depth greater than 10 reads (Brieuc *et al.* 2014). Next, loci were filtered and retained if they had a minor allele frequency ≥ 0.05 in at least one population. These criteria were used to reduce error

associated with genotyping duplicated loci that have been retained in the salmonid genome following a whole duplication event, and to provide stringency for the detection of signatures of selection. Individual samples were removed if they had ≥50% missing genotypes across all filtered loci.

Heterozygosity and Population Differentiation

Observed (H_o) and expected (H_e) heterozygosity for each population, and tests for Hardy-Weinberg equilibrium (HWE) at each locus, were computed in the R-package *adegenet* (v. 1.3-9; Jombart 2008). Tests for HWE were conducted using the Monte Carlo procedure and $1x10^5$ permutations. The expected false discovery rate (q-value) for each locus was computed to correct for multiple testing using the R-package *qvalue* (v. 1.28.0; Storey 2002). A locus was deemed significantly out of HWE when the q-value was less than 0.05. Loci out of HWE, however, were retained for analyses because they may be of interest to present and future studies.

Two approaches were used to measure genetic change between each of the lines across generations. First, genetic differentiation, F_{ST}, was calculated between populations and at each locus using Weir and Cockerham's unbiased estimator in *Genepop* (v. 4.1, Weir & Cockerham 1984; Raymond & Rousset 1995). Significance of pairwise population F_{ST} comparisons was determined from tests of genotypic differentiation performed in *Genepop* using default parameters.

Second, a discriminant analysis of principal components (DAPC) was used to visualize temporal changes in genetic relationships between the lines and to identify loci that contribute most to population separation. Principal components analysis (PCA) is not optimal for the analysis of population structure because it maximizes total genetic variation explained but does

not account for underlying genetic structure (Jombart *et al.* 2010). On the other hand, discriminant analysis (DA) is affected by correlations between variables (loci) and requires the number of variables to be less than the number of objects (individuals) (Borcard *et al.* 2011). These constraints are problematic for analysis of genomic data because some loci are physically linked, and the number of loci is typically much greater than the number of individuals. DAPC overcomes these drawbacks by combining PCA and DA. First, PCA is used to identify synthetic, uncorrelated variables (principal components or PCs) that maximize the amount of genetic variation explained in the data. PCs are then retained as variables for DA; the actual number of PCs retained can be chosen to ensure that the number of variables is less than the number of objects. DA is subsequently performed on the retained PCs to explore divergence between groups.

DAPC was conducted in the R-package *adegenet* using all individuals and loci. Since the analysis requires a complete data set, missing values were replaced by the mean frequency of the corresponding allele, computed on the whole set of individuals. To avoid over-fitting the discriminant functions, the *optim.a.score* function was used to identify the optimal number of principal components to retain in the first step of the analysis based on the difference between observed and random discrimination.

Loci that contribute most to group separation, which we refer to as discriminatory loci, were identified using the *snpzip* function. The contributions of loci to the DAPC, or loadings, were first used to compute a distance matrix between loci. A hierarchical clustering analysis using the median clustering method was then performed on the distance matrix to separate loci into those that contribute to group divergence and those that do not. Discriminatory loci were

compared to explicit tests for molecular signatures of selection to provide support for loci consistent with adaptive divergence.

Effective number of breeders

Effective number of breeders, N_b, was compared between hatchery lines as a proxy metric for estimating the effects of genetic drift in each population. *N_EEstimator* (v. 2.01, Do *et al.* 2014) was used to obtain estimates of N_b by the linkage disequilibrium (LD) and temporal methods. To reduce potential bias due to selection, loci that were identified as outliers by *F_{TEMP}* and *Bayescan*, as well as DAPC discriminatory loci, were omitted. For each year sampled, we used only four year old adults, which represented a single cohort of individuals and thus provided the most appropriate measure of N_b.

For the LD estimates, we assumed random mating and used an allele frequency restriction of 0.05. Bias due to physical linkage was removed by excluding calculations between pairs of loci on the same chromosome (Larson *et al.* 2014). Estimates of N_b and 95% parametric confidence intervals were subsequently calculated for all populations (Larson *et al.* 2014). However, estimates of N_b obtained by the LD method may be biased by overlapping generations and fluctuating population size (Waples *et al.* 2014). We accordingly adjusted our estimates of N_b, since the study population was subject to both sources of potential bias (Figure 1.3), based on principles in Waples *et al*. 2014. Briefly, annual census data was first grouped into generations comprising four years each, since 80-90% of this population matures at age four (Knudsen *et al.* 2006). For each generation, we calculated the harmonic mean of total number of spawners per year (N_{census}), as low return years within a generation have a disproportionate effect on N_b, and multiplied the mean by four to obtain total number of spawners per generation (N_{gen}). We then

calculated the weighted harmonic mean of total spawners for the previous three generations (Wt. N_{gen}), since past demographic history influences LD estimates of N_b (Waples *et al.* 2014). Previous generations, starting from the most recent, received weights of 1/2, 1/4, and 1/8 because half of existing LD decays each generation (Waples *et al.* 2014). We then divided Wt. N_{gen} by N_{census} for the year of interest. This ratio approximates the ratio of N_e/N_b, assuming N_e/N_{gen} is proportional to N_b/N_{census}. Since N_b estimates are a function of the harmonic mean of N_e and true N_b (Waples *et al.* 2014), this ratio was then used to calculate bias and adjust our N_b estimates. Full calculations are shown in Table S1.6.

For the temporal method, N_b and 95% confidence intervals were calculated using the P_1 founders and three hatchery generations based on all non-outlier loci. We followed the Jorde and Ryman method, Plan II sampling, and used an allele frequency restriction of 0.05.

Detection of outlier loci consistent with adaptive divergence

Given that F_{ST} outlier tests are known to exhibit a high rate of false positive results (Lotterhos & Whitlock 2014), three independent tests for genomic regions indicative of diversifying selection were conducted to explore the potential role of selection in divergence of the two lines from the founder population. In our final interpretation, emphasis was placed on outlier regions that were identified by multiple tests, and were significantly divergent across multiple generations.

1. *F_{TEMP} Method.* F_{TEMP} (Therkildsen *et al.* 2013) is designed to detect selection in a single population sampled across multiple generations by simulating genetic drift over time. The method is a modification of the widely-used Fdist approach (Beaumont & Nichols 1996), and is

based on an island model with drift but no migration. We parameterized simulations of genetic drift using estimates of N_b derived from the study populations, and explicitly identified loci that exceeded neutral expectations (i.e. consistent with signatures of adaptive divergence) in each hatchery line. Drift was simulated at $1x10^6$ loci for four generations. The model was first run using N_b estimates obtained by the temporal method using all loci (N_b values of 459 and 61 for the integrated and segregated hatchery lines, respectively). Yet, N_b estimates can themselves be biased by loci under selection. Therefore, F_{TEMP} outlier loci identified in the first run, as well as *Bayescan* outlier and DAPC discriminatory loci, were removed, and second estimates of N_b were calculated using only non-outlier loci. Estimates of N_b using non-outlier loci were 466 and 69 for the integrated and segregated hatchery lines, respectively, and final F_{TEMP} simulations incorporated these N_b estimates. However, to test the sensitivity of our results to N_b, we also ran F_{TEMP} simulations using the following estimates: 1) N_b estimates obtained from individuals of all ages (N_b of 388 and 59 for INT and SEG lines, respectively) instead of those calculated using only 4 year olds; 2) estimates of N_e instead of N_b (approx. $4*N_b$, Waples 2002), and 3) N_b estimates for the opposite hatchery line (i.e. N_b of 466 for the SEG line and 69 for the INT line).

Sample sizes (n) per generation were also parameterized in the F_{TEMP} simulations, where sample sizes equaled the harmonic means of the number of individuals sequenced in each generation (61 and 57 for the integrated and segregated lines), but randomly varied between 0.8*n and n to account for variation in the number of individuals genotyped per locus per generation. One thousand final F_{TEMP} simulations were performed to account for variability between runs. To correct for multiple testing and control the false discovery rate (Storey 2002), the q-value of each locus was computed from its corresponding p-value using the R-package *qvalue.* A locus was considered significant if its q-value was less than 0.05 in at least 95% of the

final simulations. Significant loci showed greater temporal allelic variation than is expected under genetic drift alone.

2. *Bayescan*. We also employed a Bayesian approach to identify outlier loci using the program *Bayescan* (v. 2.1; Foll & Gaggiotti 2008). However, this approach assumes that the sampled populations evolved independently from a common ancestral population; thus, it is not ideally suited for temporal data. In addition, the power of *Bayescan* to detect outlier loci is reduced with few populations (<6), low sample sizes (<30), and low neutral F_{ST} values (<0.01) (Foll & Gaggiotti 2008). Despite these limitations, we used *Bayescan* because it has lower type I and type II errors than other available methods (De Mita *et al.* 2013). The program was run with all populations combined with default parameters. A locus was considered to be an outlier if its q-value was less than 0.05 in each of five independent runs.

3. Test for outlier regions of the genome: Sliding window analysis. Moving averages of pairwise F_{ST} values between each hatchery line and the P_1 founders at mapped markers were calculated using a kernel smoothing sliding window approach (Hohenlohe *et al.* 2010; Brieuc *et al.* 2015). Each chromosome was divided into 100 windows spanning 18cM each; these parameters were optimized to detect small regions of divergence but minimize background noise due to sampling variance. Contributions of individual loci to the F_{ST} average were weighted by their distance from the center of the window. A null distribution with 95% confidence intervals was calculated for each window using $1x10^6$ replicates of randomly sampling F_{ST} values from all mapped loci, with replacement, where sample size was based on the number of loci within each window. The sliding window analysis was performed separately on the F_1, F_2, and F_3 generations of each

hatchery line. Regions of significantly elevated divergence were identified as those where the moving average exceeded the 95% confidence interval of the null distribution.

Outlier Alignment and Gene Function

All loci located within the genomic regions that exhibited overlap among the three tests of adaptive divergence (summarized in Table 1) were aligned to the rainbow trout genome (Berthelot *et al.* 2014) using *Bowtie* (Langmead *et al.* 2009) to identify genes associated with divergence and potential targets of domestication selection. *BLAST2GO* (Conesa *et al.* 2005; Götz *et al.* 2008) was then used to conduct a *BLAST* (Altschul *et al.* 1990) search on the NCBI non-redundant (nr) public database for the rainbow trout gene coding sequences that were identified. We used an *e*-value threshold of $1x10^{-10}$ in *BLAST* searches. Gene ontology (GO) terms, GO Slim terms, and functions associated with each gene were subsequently identified.

1.4 RESULTS

DNA Sequencing and Genotyping

We identified 9410 bi-allelic RAD loci, including 4405 loci that aligned to the Chinook salmon linkage map and had a minor allele frequency >0.05 in at least one population (Table S1.1). A total of 413 individuals were genotyped at >50% of these loci and retained for analysis (Table S1.1). Sample sizes for the P_1 founders and each generation of integrated and segregated hatchery lines are summarized in Table S1.2.

Heterozygosity and Population Differentiation

Observed (H_o) and expected (H_e) heterozygosity was similar between the integrated and segregated hatchery lines within each generation, though a small decrease was observed over time (Table S1.2). All populations had less than 3% of loci that deviated from Hardy-Weinberg equilibrium. Loci that deviated from HWE were included in further analyses because they may be of interest to the present study.

Population genetic differentiation, F_{ST}, for each of the generations compared to the P_1 founders was low in all pairwise comparisons, ranging from 0 to 0.0108 (Table S1.3). The F_1 hatchery fish and segregated line steadily diverged from the P_1 founders over time; the F_2 and F_3 SEG populations were significantly differentiated from the founders (F_{ST} =0.0049 and 0.0108 respectively, $p<0.001$). In contrast, the integrated line did not exhibit the same temporal trend of increasing genetic divergence. However, F_{ST} compared to the P_1 founders was significant for the F_2 and F_3 INT populations (F_{ST} =0.0026 and 0.0022 respectively, $p<0.001$). F_{ST} between the F_3 SEG and F_3 INT populations was also significant (F_{ST} = 0.0093, $p<0.001$), indicating divergence in the two hatchery lines.

The first step of the discriminant analysis of principal components (DAPC) identified 58 PCs to retain, which explained 29.3% of the observed genetic variation among individuals. A plot of the kernel density estimates of individuals along the first discriminant function, which explained 63.7% of the retained variation, showed separation of the F_1 hatchery fish and segregated line from the P_1 founders (Figure 1.4). The segregated line became more divergent over time; in comparison, the F_1 wild group and integrated hatchery line clustered closely with the P_1 founders. The second discriminant function explained 13.6% of the retained variation and, to a minor extent, represented temporal variability within the two lines. We identified 25 loci that contributed most to separation along the first discriminant function based on hierarchical

clustering analysis (termed discriminatory loci; Table S1.8). A majority of the discriminatory loci (17 out of 25) did not overlap with outlier loci identified by explicit tests of selection and thus were inferred to be selectively neutral.

Estimates of effective number of breeders

Estimates of the effective number of breeders, N_b, obtained by the LD method for each sample year largely reflected the effective number of parents that produced the samples (Waples 2005), but also had contributions from previous generations (Waples *et al.* 2014). Bias-adjusted estimates of N_b from the natural population revealed a 3-4 fold increase following establishment of the integrated hatchery line (Figure 1.5; Table S1.5a), although this period also experienced higher adult numbers returning to the basin (Figure 1.3). Estimates of N_b in the segregated line declined steadily over time and, in the latter two generations, were over an order of magnitude lower than in the integrated line (Figure 1.5; Table S1.5a), suggesting that genetic drift may have driven the observed genetic divergence. Differences between the two lines were not unexpected, since the average broodstock sizes were 363 (sd=56) and 85 (sd=15) for the integrated and segregated lines, respectively. Interestingly, wild fish that were spawned in the hatchery in 1998 had a lower N_b estimate than the wild fish that reproduced in the river, despite having similar census sizes and the same population history (Table S1.5a).

The temporal method provided single estimates of N_b for the integrated and segregated lines across the sampled generations (Figure 1.5; Table S1.5b). Temporal estimates closely agreed with the harmonic means of the LD estimates and showed that the integrated line had a larger N_b.

Detection of outlier loci and chromosomal regions of high divergence

F_{TEMP} identified 228 temporal outlier loci in the segregated line and 80 in the integrated line (Table S1.8). Thirty-two loci were identified as outliers in both lines. F_{TEMP} results did not change considerably when run with the N_b estimates obtained from individuals of all ages. Running F_{TEMP} with estimates of N_e instead of N_b did not alter results for the integrated line, but it significantly increased the number of temporal outliers in the segregated line. In addition, there were 369 F_{TEMP} outliers in the segregated line but just 48 outliers in the integrated line when the N_b estimates for one hatchery line were used for the opposite line. These results indicated that the numbers of outlier loci detected in the segregated line were consistently higher than in the integrated line across a range of N_b values, and that using the single cohort N_b estimates instead of N_e was a conservative choice.

The second analysis, conducted using *Bayescan*, identified 75 outlier loci across all combined populations (Table S1.8). Since the *Bayescan* analysis required combining populations, outlier loci could not be attributed to a specific hatchery line. However, there was considerable overlap with F_{TEMP} results. Twenty-eight *Bayescan* outlier loci were also F_{TEMP} outliers in both hatchery lines, while an additional 17 and 22 *Bayescan* loci were F_{TEMP} outliers in only the integrated or segregated lines, respectively.

Lastly, sliding window analyses conducted on the 4405 mapped loci identified several regions of significantly elevated F_{ST} values in both hatchery lines (Table S1.9). However, regions of high divergence in the segregated line were more temporally stable than those in the integrated line; that is, more regions were consistently identified across the F_1, F_2, and F_3 generations. Five genomic regions had significantly elevated F_{ST} values in two of the three generations within the segregated line, and four regions were significantly high in every

generation. In contrast, the integrated line contained three regions that were significantly elevated in two generations, and no region was divergent across all generations. Furthermore, plots of per locus F_{ST} values for markers positioned along the chromosome revealed that divergence from the P_1 founders increased over time at some regions in the segregated line, and these regions also contained loci that were identified as outliers by other tests (Figure 1.6). This directional, temporal trend was not observed in the integrated line, although a region on Ots11 was significantly divergent in both hatchery lines in the F_1 and F_3 generations.

Overall, seven genomic regions exhibited overlap among all three tests of diversifying selection in the segregated line (Table 1). One region showed consensus across tests in the integrated line, but this was also shared by the segregated line. These results, in conjunction with the temporal increase in F_{ST} at some regions, suggest that selection might also have contributed to divergence in the segregated line.

Outlier Alignment and Gene Function

Forty-three mapped loci were located within the seven genomic regions that exhibited overlap among the three tests of adaptive divergence, and 25 of these loci aligned to the rainbow trout (RBT) genome. Twenty-two loci, including five outliers, were located within or near annotated RBT genes. While the RBT genes had an array of functions, commonalities among genes were observed. Three genes, including one associated with an outlier locus, were related to the protein ubiquitin, which is important for protein degradation and immune response. Two other genes were also related to the function and response of the immune system. Three genes, with two linked to outlier loci, were involved in the utilization, breakdown, and reception of lipids, while another two genes, including one near an outlier locus, were essential for the

production of ribose, a sugar that can be used for energy or the synthesis of nucleotides. Three genes were linked to neurotransmission, including one gene that coded for receptors of gamma-aminobutyric acid (GABA), a major inhibitory transmitter in the central nervous system, and a second gene that has been shown to induce clustering of GABA receptors. Lastly, two of the RBT genes, including one from an outlier locus, were linked to the development of photoreceptors and eye pigmentation. The shared functions of many genes provided insight on the potential targets of domestication selection. Alignment results, as well as gene functions with references, are summarized in Table S1.10.

1.5 DISCUSSION

The aim of this study was to evaluate the effectiveness of managed gene flow to reduce genetic divergence in supportive breeding programs. This goal was achieved by comparing genome-wide diversity in three generations of integrated and segregated Chinook salmon hatchery lines to their founding wild population. Genetic distance measures at 9410 loci showed that divergence in the segregated line was significant by the second generation. Much of this change can be ascribed to genetic drift, as suggested by the small numbers of effective breeders. However, we also found evidence for a temporal trend in divergence at specific genomic regions, consistent with domestication selection. In contrast, genetic divergence in the corresponding integrated line was marginal over three generations, suggesting that the use of natural origin broodstock was effective at reducing genetic change in the short term. By comparing contrasting regimes, the results illustrate the range of possible outcomes that may occur when using deliberate gene flow to mitigate genetic impacts in captive breeding programs. This information

is therefore highly relevant when weighing the relative risks of different captive rearing strategies.

Like many such studies on supportive breeding programs (Araki *et al.* 2007; Hess *et al.* 2012; Anderson *et al.* 2013; Christie *et al.* 2014), this study does not have a contemporary control population to which the hatchery lines can be compared. After the first hatchery generation, the upper Yakima population was no longer "wild" due to influence from the integrated line. Therefore, ongoing evolutionary processes in the natural environment cannot be readily discriminated from those in the hatchery. However, the comparisons to the wild founding population strongly suggest that divergence in the segregated line was primarily due to repeated exposure to the hatchery environment, since differentiation did not occur in the integrated line.

It is also important to note that the contribution of integrated hatchery fish to natural population productivity has not been fully quantified. Higher adult returns in the natural population were observed following initiation of the hatchery (Figure 1.3), but this increase might be attributed to supplementation, a favorable shift in environmental conditions (Mantua *et al.* 1997; Fast *et al.* 2015), or a combination of factors (Scheuerell *et al.* 2015). Previous studies suggest that hatchery-reared individuals tend to be less fit than naturally born fish (Araki *et al.* 2007; Milot *et al.* 2013; Christie *et al.* 2014), but successfully spawning hatchery fish can still contribute to natural production. In this system, there are three lines of evidence that suggest that hatchery fish aided population productivity. First, the proportion of naturally-spawning fish that were of hatchery origin reached 70-75% in some years (Figures 1.2, 1.3) and, given these proportions, it is highly likely that these fish contributed to subsequent generations. Second, nest (redd) abundance and spatial distribution significantly increased following supplementation efforts, an increase that exceeded numbers observed in a nearby unsupplemented river (Fast *et al.*

2015). Third, comparisons of female reproductive traits (Knudsen *et al.* 2008) and breeding success experiments within an artificial stream (Schroder *et al.* 2008; Schroder *et al.* 2010) revealed little difference between wild and first generation hatchery salmon, though offspring of wild females had significantly higher survival to the fry stage (5.6%, p=0.04; Schroder *et al.* 2008). A full understanding of the contribution of hatchery-reared fish to population productivity will be gained through complementary pedigree-based studies.

An ideal experiment would have compared lines with equal numbers of broodstock. For example, in theoretical treatments on the effects of different management strategies on genetic change in supportive breeding programs, Duchesne and Bernatchez (2002) found that increasing the census size of the captive population had the most influence on reducing the inbreeding coefficient compared to other tactics such as changing the proportion of wild individuals in the broodstock or modifying the numbers of captive fish released. Here, the numbers of adults spawned in the segregated line were typically a quarter of those in the integrated line, an outcome of the fact that the Cle Elum facility serves broader restoration goals beyond the experiment. This difference in broodstock size partly explains the divergence from the founder population observed in the segregated line, and it is likely that simply increasing the broodstock size would have reduced divergence. However, one of the major objectives of managed gene flow is to minimize genetic drift. This goal is partly achieved through taking advantage of both wild and hatchery production, effectively increasing the "rearing space" of supportive breeding programs. Therefore the effective size of integrated programs can be expected to exceed comparable segregated programs in cases where the hatchery or the wild component does not become a "sink" for the "source" population.

Evaluations of the relative broodstock census size and effective number of breeders across the integrated and segregated lines provide insight into the relative importance of gene flow in mitigating genetic divergence from the founder population. The ratio of broodstock census sizes (N_{brood} INT/N_{brood} SEG) for the two lines can be compared to the ratio of N_b estimates (N_b INT/N_b SEG) in each generation (Table S1.7). These ratios would be approximately equal if the N_b estimates solely reflected an increase in broodstock size. However, the ratios of N_b INT/N_b SEG were larger than N_{brood} INT/N_{brood} SEG, particularly in the F_2 and F_3 generations (Figure 1.7, Table S1.7). This simple comparison suggests that managed gene flow disproportionally increased N_b in the integrated line and reduced genetic divergence by incorporating the naturally spawning segment of the population.

There is an interesting aspect of this study that illustrates the temporal variation in the effects of hatchery rearing on N_b. The 2002 (F_1 period) estimates reflect the N_b of fish that spawned in 1998 (P_1 period). In this first period, approximately equal numbers of wild fish spawned in the river and in the hatchery, yet the N_b of fish in the river was over two times greater than the N_b of fish in the hatchery. The result suggests that in this generation, the hatchery environment caused higher variance in reproductive success and decreased N_b, which is consistent with other studies of effective size in hatcheries (reviewed in Naish *et al.* 2008; Christie *et al.* 2012; Naish *et al.* 2013). In contrast, the ratio N_b/N decreased in the F_2 and F_3 generations of the integrated line despite higher N_b estimates relative to the segregated line, suggesting that ecological influences such as higher competition on the spawning grounds (Fleming & Gross 1994) or limited carrying capacity might have influenced this ratio as the population increased in size. Overall, these results emphasize the dynamic impacts of both natural processes and hatchery rearing on effective size, the significance of multigenerational

monitoring, and the need to understand the factors influencing the success of supportive breeding programs as a whole.

Multiple lines of evidence and consistent temporal trends in the segregated line provided compelling support for some genomic regions involved in domestication. In contrast, outlier regions detected in the integrated line had minimal overlap among tests and lacked temporal trends. Tests for F_{ST} outliers are known to exhibit high false positive rates (Lotterhos & Whitlock 2014), but they continue to be used because they serve as useful starting points for further investigations (Storz 2005). Here, we reduced the likelihood of false positive results by using three independent tests for outliers (e.g. de Villemereuil *et al.* 2014), including one specifically designed to detect temporal divergence. Signatures of domestication selection are also difficult to detect, particularly in early generations (Lotterhos & Whitlock 2014; Mäkinen *et al.* 2015), and power is reduced with few populations in the analyses (Karlsson & Moen 2010). However, the results presented here are comparative and point toward greater and more rapid changes in the segregated line. More outliers were observed in the line with a smaller effective size, an interesting finding considering that selection is expected to be more efficient in larger populations (Wright 1931). Thus, the strength and direction of selection likely differ between the hatchery and natural environments. The relaxation of natural selective pressures that occurs in the hatchery, coupled with intense natural selection after hatchery-reared fish are released into the wild, may also promote domestication by favoring individuals that are adapted to the hatchery environment (Reisenbichler & Rubin 1999). Similar results have been observed in other small, fragmented populations (Koskinen *et al.* 2002). Finally, we note one outlier region that was divergent in the F_1 and F_3 generations of both hatchery lines. Since this region was an outlier in the F_1 wild group, the results might indicate a false positive or the action of natural selection.

However, the region was also an outlier in the F_1 hatchery and F_3 groups, possibly providing evidence of selection in the hatchery at early life stages that has affected both lines. We cannot rule out additional regions of undetected selection in both lines, as it is currently not possible to survey the whole genome in Chinook salmon. Despite these potential limitations, outlier regions identified in the segregated line are supported by multiple lines of evidence and provide valuable insight for the genetic basis of adaptation to captivity.

Alignment of loci within outlier regions to the rainbow trout genome and classification of gene functions identified possible targets of domestication selection. Multiple genes were related to the regulation and response of the immune system, which may be affected by the higher rearing densities in hatcheries. Many gene products, including those associated with outlier loci (ubiquitin, alpha beta hydrolase, and ribokinase), were connected to the processing of proteins, lipids, and sugars and may be affected by hatchery feeding regimes. For example, in rainbow trout, the ubiquitin-proteasome pathway, which facilitates protein degradation, was affected by starvation (Martin *et al.* 2002) and feeding status (Seiliez *et al.* 2008). Additionally, differential gene expression was observed between groups of rainbow trout that were fed different diets (Panserat *et al.* 2009). Two other gene products were important for the development of photoreceptors and eye pigmentation (salmon are visual feeders). Together, these functions suggest that domestication selection may be occurring in the hatchery environment due to differences in food availability and composition.

The two GABA-related genes, one that coded for GABA receptors and a second that induces clustering of GABA receptors, suggest another potential target of domestication selection. GABA has been shown to stimulate the secretion of two gonadotropins, GTH-1 and GTH-2, in rainbow trout (Mañanos *et al.* 1999), and these gonadotropins have been shown to

stimulate gonad growth in juvenile rainbow trout (Suzuki *et al.* 1988). Domestication selection acting on this pathway may explain the higher rates of early male maturation commonly observed in hatchery populations of Chinook salmon (Larsen *et al.* 2010; Harstad *et al.* 2014). Furthermore, early male maturation is positively correlated with cumulative growth (size at release, Harstad *et al.* 2014). Domestication selection likely targets polygenic traits, and here there is evidence that such selection may affect both the GABA-related genes and the genes that process proteins, lipids, and sugars.

The results reported here add to the growing number of studies that have, to date, typically focused on the relative performance of one or two generations of captive individuals released into the wild, and only under a single broodstock management regime. In salmon species, a few such studies documented minimal differences between hatchery and wild fish, including comparable reproductive success (Anderson *et al.* 2013), breeding success in an artificial channel (Schroder *et al.* 2008; Schroder *et al.* 2010), reproductive traits (Knudsen *et al.* 2008), and spawning distributions (Dittman *et al.* 2010). However, many studies have found significant differences in hatchery fish, including reduced response or increased vulnerability to predation (Fritts *et al.* 2007), lower survival (McGinnity *et al.* 2003), differences in growth rate and morphology (McGinnity *et al.* 2003; Busack *et al.* 2007), and reduced reproductive success (Araki *et al.* 2007, 2009; Christie *et al.* 2014). Yet, the genetic basis of these differences, and the rate at which they occur, remain unclear. Using an experimental system that included contrasting lines founded from a single population and sampled extensively over multiple generations, we began to address these uncertainties by estimating the rate of genetic change in hatchery fish with unprecedented resolution. While the magnitude of divergence observed in the segregated line was small, it is important to consider that this change occurred after only three generations

of captive rearing. If the trend of divergence continues, then F_{ST} could equal or exceed values typically observed between natural Chinook salmon populations (e.g. Brieuc *et al.* 2015) within another three to five generations. Thus, we consider the rate at which genetic differentiation occurred to be noteworthy and of significant interest to conservation programs. We also identified processes driving change in the segregated line, including the first possible indicators consistent with domestication selection observed in Pacific salmon hatchery populations.

Determining the impact of different captive rearing approaches on population productivity in species of conservation interest is central to measuring their success, but empirical studies on such species are usually constrained by the natural systems in which they operate. An ideal experiment would demonstrate that integrated fish have greater fitness than segregated fish in the natural environment. However, permitting spawning of the segregated fish in the same environment as the integrated fish would promote gene flow between the lines, and negate both the experiment and the concurrent restoration efforts in the Chinook salmon system we have studied. Therefore, it was necessary to use proxies for fitness by identifying signatures of selection and measuring effective population size, consistent with best practices in genetic monitoring (Schwartz *et al.* 2007). These measures can be further enhanced by integrating genome-wide association studies with phenotypic measures of fitness.

This study is the first to empirically compare multigenerational consequences of alternative management approaches in a non-model species of conservation interest using genome-wide surveys, and adds to the range of available tools for assessing risks associated with supportive breeding. Multigenerational observations are advantageous because processes occurring in the wild, including natural selection, could mitigate or exacerbate the effects of captive rearing over time. Our findings provide the first empirical demonstration that using

natural-born parents in captive breeding programs is more effective at reducing genetic divergence than using only captive-born individuals. Interpreting these findings in a management context can be challenging on one hand, because determining "acceptable" levels of divergence depends in part on a clear link between the molecular measures we report and population productivity. On the other hand, the study is comparative in nature, and considers a range of possible outcomes, an approach that is highly relevant to risk assessment in realistic scenarios (Waples 1999; Duchesne & Bernatchez 2002; Ford 2002; Waples & Drake 2004; Naish *et al.* 2008). The experimental system we evaluated therefore provides guidance on "best practices" in supportive breeding. Conservation plans should thus consider establishing captive breeding programs using wild broodstock before the source population becomes too small, so that the benefits of starting such a program outweigh the demographic and genetic costs of removing individuals from the natural environment.

1.6 ACKNOWLEDGEMENTS

We thank everyone who was involved in establishing CESRF and shaping its research direction, including Levi George, Melvin Sampson, Steve Schroder, Craig Busack, past and present members of the Independent Scientific Review Panel, and the Yakama Nation Tribal Council. We also recognize the following individuals for project development, broodstock collection and sampling, and laboratory assistance: Michael Ford, all Yakama Nation and Washington Department of Fish and Wildlife personnel at the Roza Dam Adult Monitoring Facility and CESRF, Isadora Jimenez-Hidalgo, Katrina van Raay, and Daniel Drinan. Finally, we thank two anonymous referees and the Editor in Chief, Louis Bernatchez, for their insightful comments. Funding for this study was provided by NOAA Fisheries/Federal Columbia River

Power System (FCRPS) Biological Opinion Remand Funds (to K.A.N. and J.J.H.), Washington Sea Grant (Award NA14OAR4170078 to K.A.N), and the Marine Technology Society Graduate Student, John G. Peterson Endowed, Richard T. Whiteleather, and John N. Cobb Scholarships.

1.7 TABLES

Table 1.1 – Genomic regions that exhibited overlap among all three tests of adaptive divergence. The table indicates the number of generations (Gens) that a portion or all of each region was identified as a region of high divergence by the sliding window (SW) analyses for the integrated (INT) and segregated (SEG) lines, and the number of outlier loci within each region that were identified by F_{TEMP} or *Bayescan*.

Region	Chromosome	Map Position (cM)	Gens SW INT	Gens SW SEG	F_{TEMP}INT	F_{TEMP}SEG	Bayescan
1	Ots04	37.52-43.15		2		2	2
2	Ots05	87.43-95.05	1	3		1	1
3	Ots06	94.64-100.65	1	1		1	1
4	Ots11	57.77-71.46	2	3	1	2	1
5	Ots12	26.14-35.29	1	2		2	2
6	Ots15	132.33-139.26	1	1		1	1
7	Ots20	93.65-99.49	1	2		2	1

1.8 FIGURES

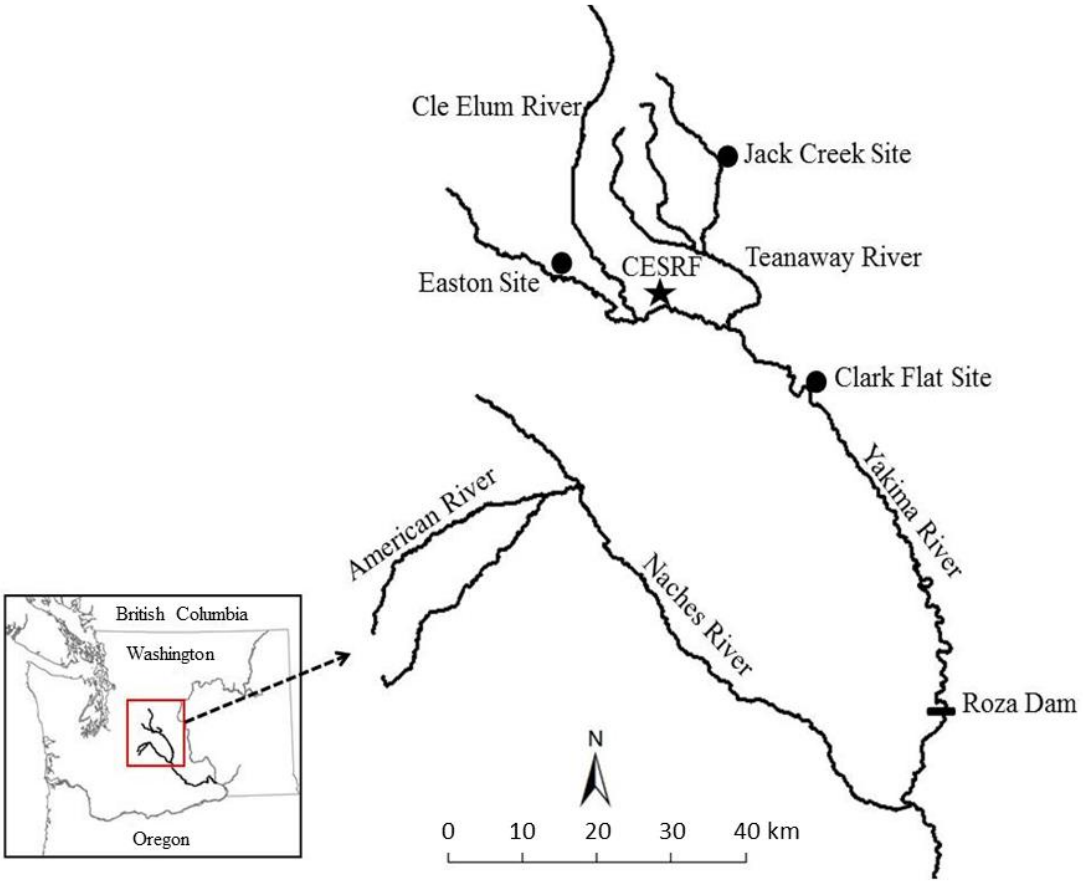

Figure 1.1 – Map of the Yakima River system. The three spring Chinook salmon populations spawn in the American River, the Naches River, and the upper Yakima River above Roza Dam. The upper Yakima population is the target of the Cle Elum Supplementation and Research Facility (CESRF). All adults returning to the upper Yakima River are sampled at Roza Dam and allowed to spawn naturally (natural origin and integrated line fish) or are removed from the system (all segregated line fish). Spawning and rearing for the hatchery lines occurs at CESRF. Prior to outmigration in spring, juveniles are transferred to the Easton, Jack Creek, and Clark Flat acclimation sites, where they are held for approximately two months before volitional release.

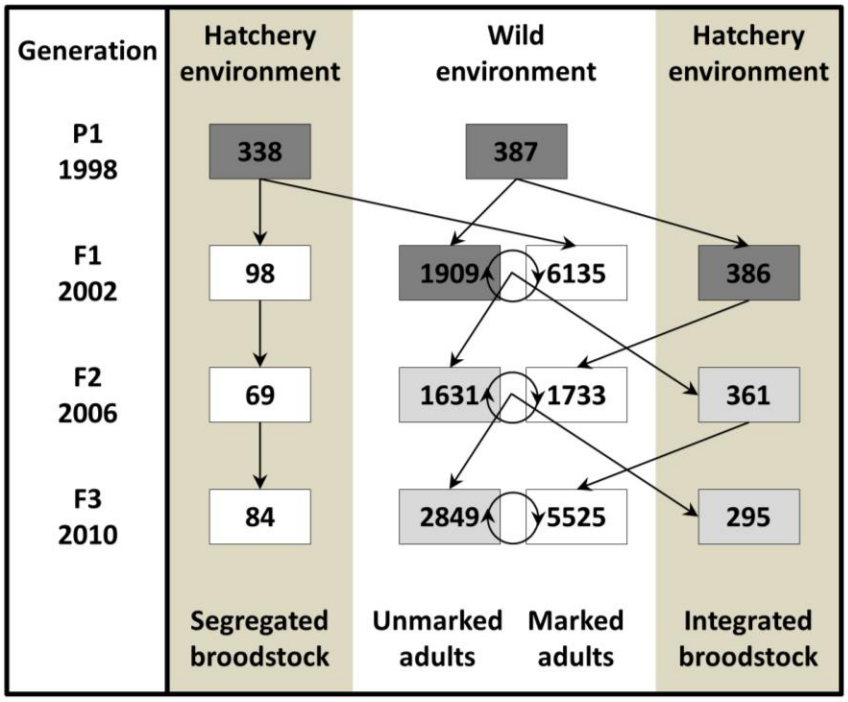

Figure 1.2 – Schematic illustrating the initiation (the founding P_1 generation) and subsequent broodstock management (F_1-F_3 generations) for the integrated and segregated hatchery lines of anadromous Chinook salmon that were surveyed. Numbers given in each box are the numbers of spawners (wild environment) and the number of broodstock (hatchery environment) for each brood year surveyed. Linear arrows denote the contribution of wild spawners or hatchery broodstock to the subsequent generation. Circular arrows represent unobserved mating between wild born (unmarked) and hatchery born (marked) spawners in the wild environment. Dark gray boxes represent wild adults, light gray boxes represent natural origin adults with hatchery, wild, or hybrid ancestry, and white boxes represent adults born in the hatchery. Only brood years sampled are illustrated, but the same design is implemented each year. Chinook salmon are semelparous but have overlapping generations - approximately 80-90% of adults at CSERF are four year olds.

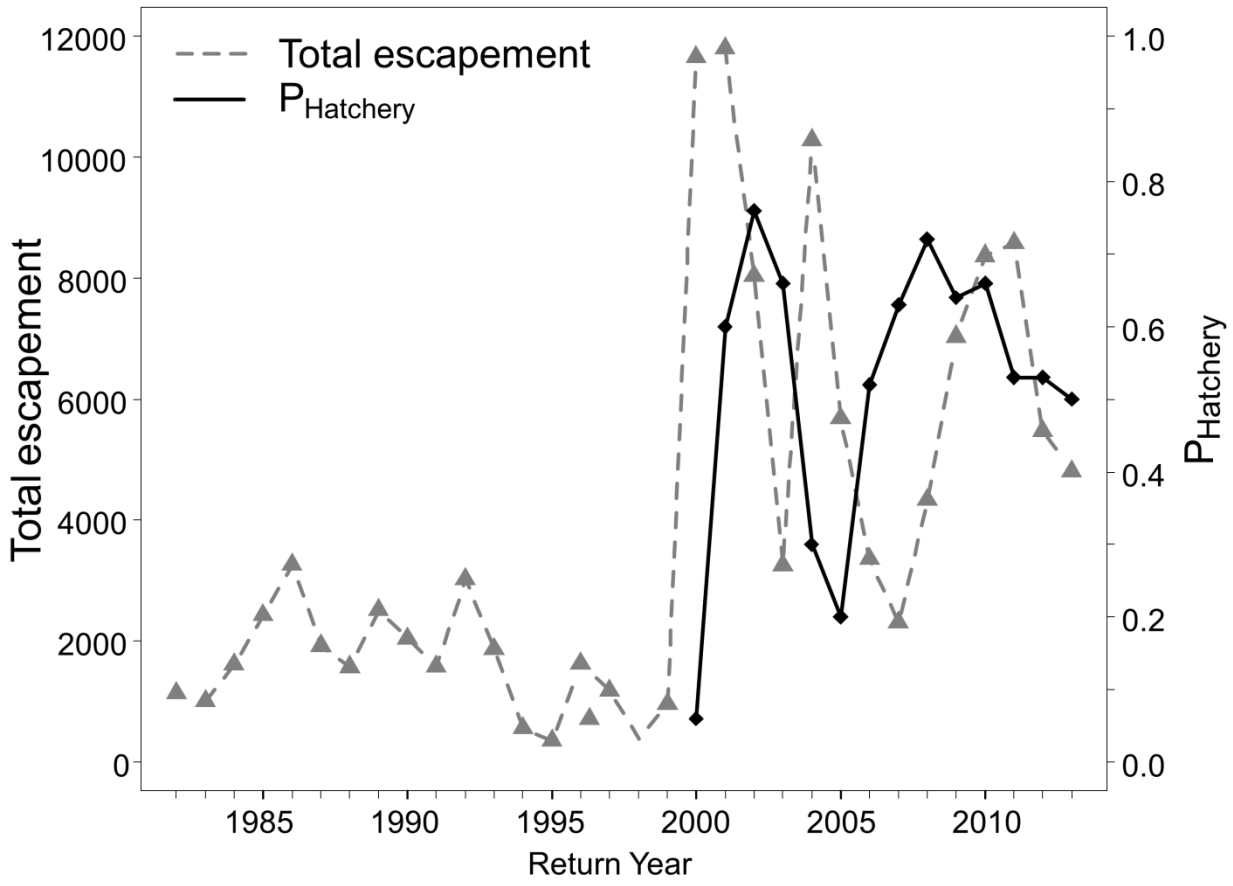

Figure 1.3 – Total annual escapement of adult Chinook in the upper Yakima river (gray, dashed line line) and the proportion of spawners that are of hatchery origin (black line).

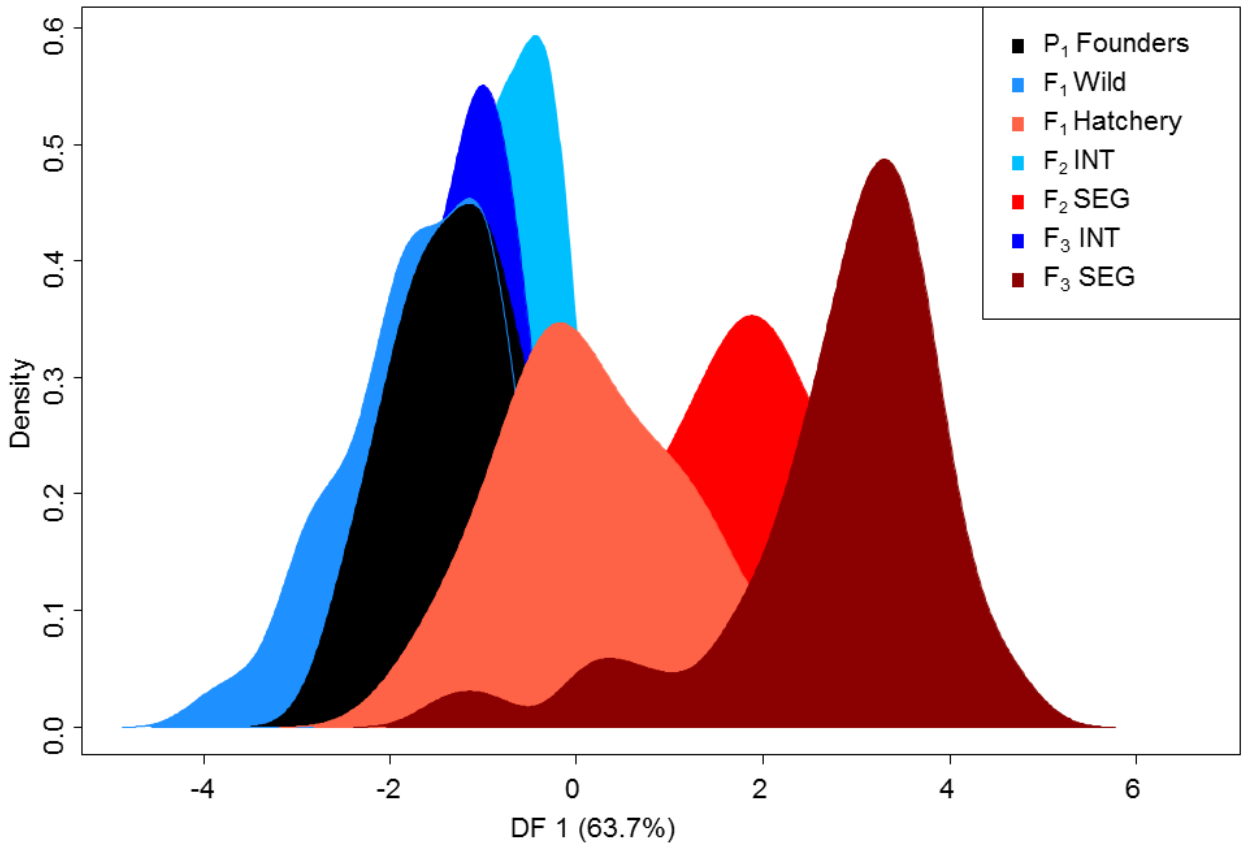

Figure 1.4 – Density plot of individuals from the wild founders (P_1 Founders, black) and three generations of the integrated (INT, blue colors) and segregated (SEG, red colors) hatchery lines along the first discriminant function from the discriminant analysis of principal components (DAPC).

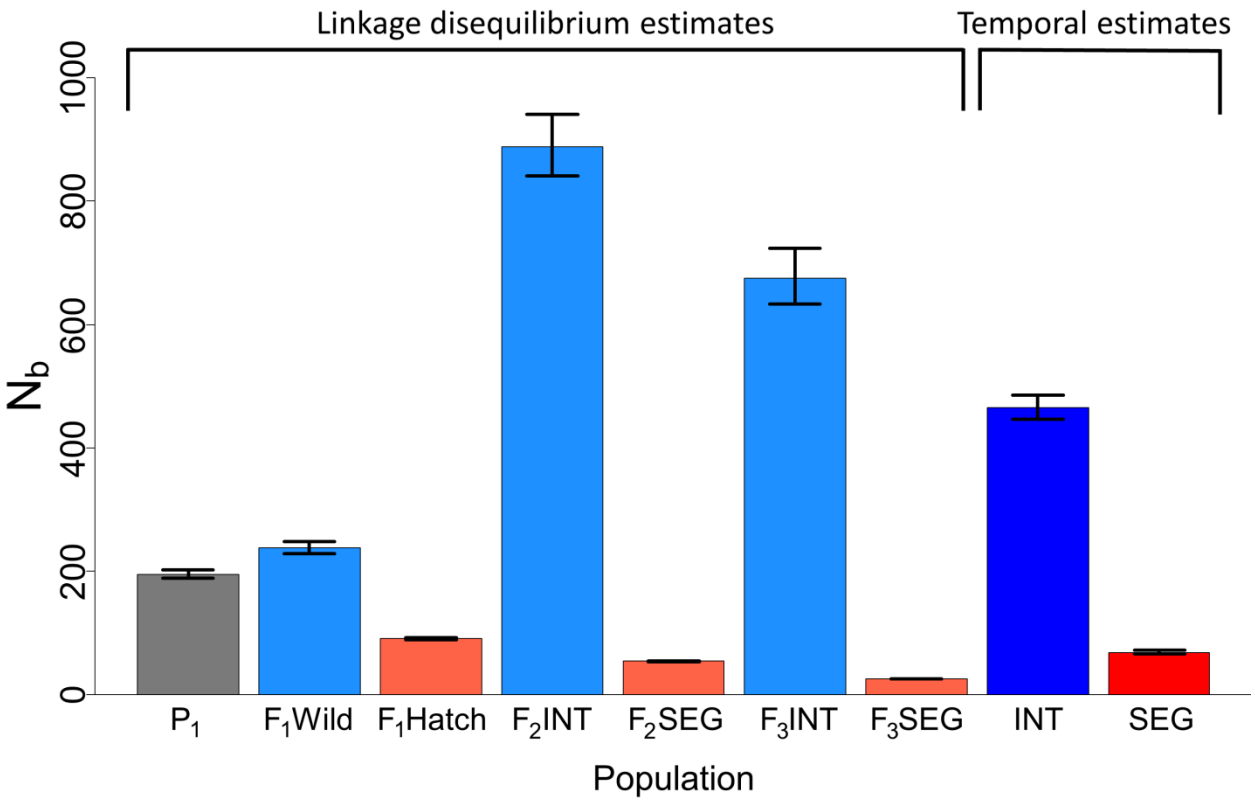

Figure 1.5 – Effective number of breeders, N_b, and 95% confidence intervals estimated by the linkage disequilibrium (LD) and temporal methods for the P_1 Founders (gray) and integrated (blue) and segregated (red) hatchery lines. The LD method allows estimation of N_b for every generation, while the temporal method yields a single estimate for the entire sampling period. LD estimates are adjusted for physical linkage and potential bias caused by overlapping generations and fluctuating population size.

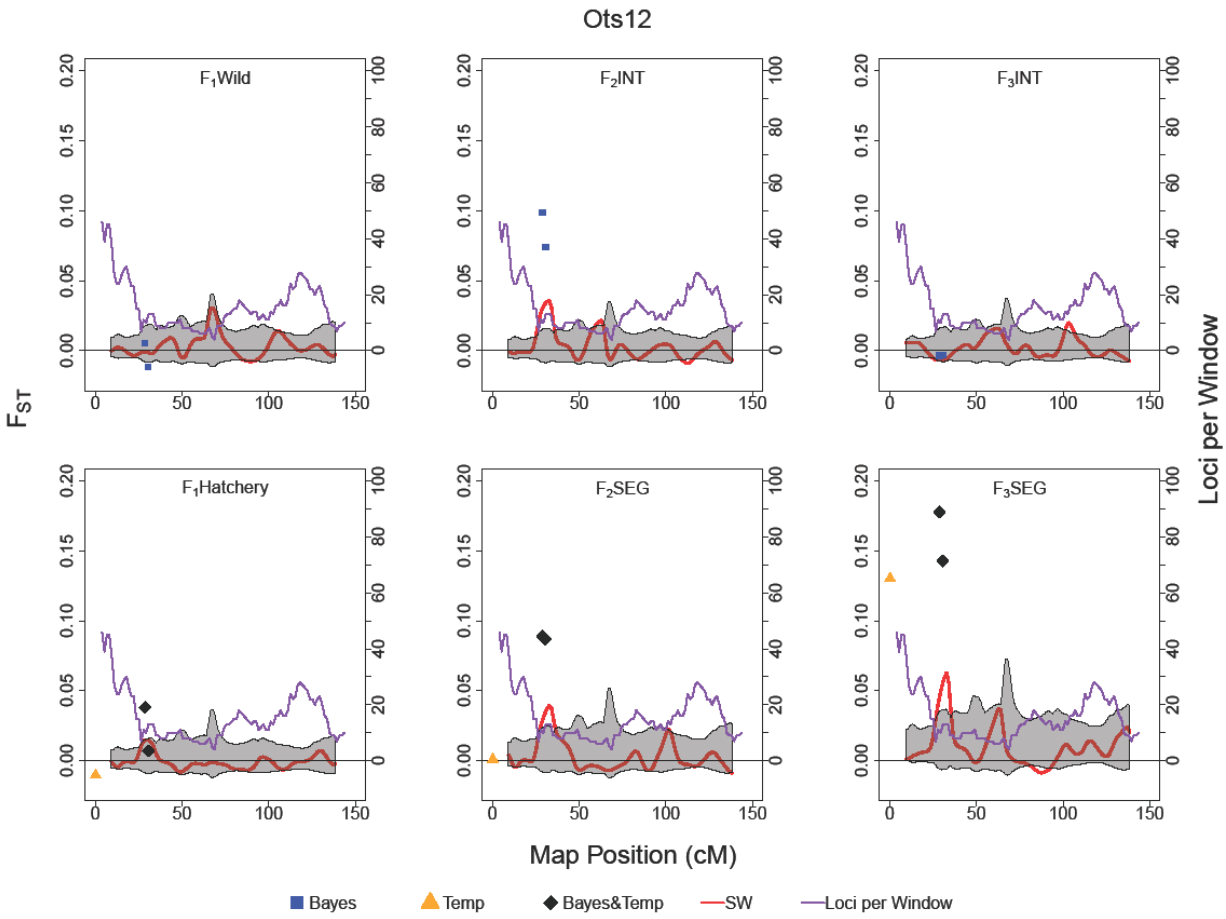

Figure 1.6 – Loci and regions of the genome showing signatures of adaptive divergence, based on pairwise F_{ST} compared to the P_1 founders, on chromosome Ots12 for the integrated (top panel) and segregated (bottom panel) hatchery lines through the F_1, F_2, and F_3 generations. Blue squares are loci that were identified as outliers with *Bayescan*, orange triangles are outliers identified by F_{TEMP}, and black diamonds are loci identified by both *Bayescan* and F_{TEMP}. The red line represents the kernel smoothed moving average of F_{ST} and the gray shaded area is the 95% confidence interval. The purple line shows the number of loci within each sliding window of the moving average.

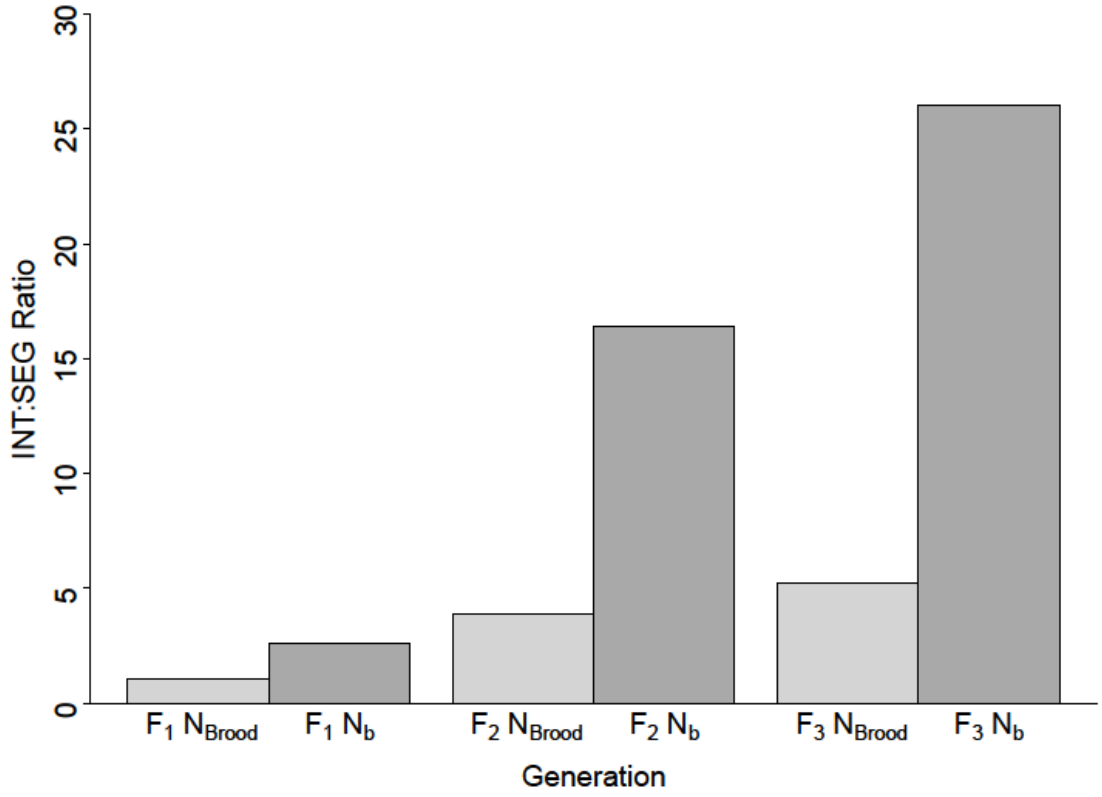

Figure 1.7 – The integrated:segregated ratios of broodstock sizes (N_{brood} INT/N_{brood} SEG) and effective number of breeders (N_b INT/N_b SEG) for the F_1, F_2, and F_3 hatchery generations. The N_b ratios were consistently higher than the N_{brood} ratios, indicating that differences in N_b between the hatchery lines were primarily due to managed gene flow and not differences in broodstock sizes.

1.9 SUPPLEMENTARY MATERIAL

S1 – Supplementary tables containing metadata and results of Chapter 1

S1.1 – Lists of the 9410 RAD loci and 413 individuals analyzed in this study.

S1.2 – Numbers of adults used as broodstock, the number of adults sampled in this study, the final sample size retained after filtering, the observed and expected heterozygosity for each population, and the number of loci that deviated from Hardy-Weinberg equilibrium.

S1.3 – Pairwise population differentiation (F_{ST}) compared to the P_1 Founders and between the two hatchery lines within each generation.

S1.4 – F_{ST} per locus for each population when compared to the P_1 Founders.

S1.5 – (a) Spawning groups and census sizes to which estimates of effective number of breeders apply, the proportions of adults that were of hatchery origin, and the years and populations from which estimates of effective number of breeders were obtained using the linkage disequilibrium (LD) method. LD estimates with 95% confidence intervals are shown. (b) Effective number of breeders for both hatchery lines from 1998-2010 estimated using the temporal method.

S1.6 – Bias calculations for estimates of effective number of breeders from the LD method.

S1.7 – Estimated contribution of naturally spawning fish to effective number of breeders in the integrated line.

S1.8 – Discriminatory loci identified for the first discriminant function of the DAPC and outlier loci identified by *Bayescan* and F_{TEMP}.

S1.9 – Regions identified by sliding window analyses as showing elevated divergence.

S1.10 – Gene ontology summary for loci within outlier regions of overlap.

1.10 REFERENCES

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

Anderson JH, Faulds PL, Atlas WI, Quinn TP (2013) Reproductive success of captively bred and naturally spawned chinook salmon colonizing newly accessible habitat. *Evolutionary Applications*, **6**, 165-179.

Araki H, Cooper B, Blouin MS (2007) Genetic effects of captive breeding cause a rapid, cumulative fitness decline in the wild. *Science*, **318**, 100-103.

Araki H, Cooper B, Blouin MS (2009) Carry-over effect of captive breeding reduces reproductive fitness of wild-born descendants in the wild. *Biology Letters*, **5**, 621-624.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid snp discovery and genetic mapping using sequenced rad markers. *Plos One*, **3**, 7.

Baskett ML, Waples RS (2013) Evaluating alternative strategies for minimizing unintended fitness consequences of cultured individuals on wild populations. *Conservation Biology*, **27**, 83-94.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B-Biological Sciences*, **263**, 1619-1626.

Berthelot C, Brunet F, Chalopin D *et al.* (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, **5**, 10.

Borcard D, Gillet F, Legendre P (2011) *Numerical ecology with r* Springer Verlag, New York.

Brieuc MSO, Ono K, Drinan DP, 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.

Burkhead NM (2012) Extinction rates in north american freshwater fishes, 1900-2010. *Bioscience*, **62**, 798-808.

Busack C, Knudsen CM, Hart G, Huffman P (2007) Morphological differences between adult wild and first-generation hatchery upper yakima river spring chinook salmon. *Transactions of the American Fisheries Society*, **136**, 1076-1087.

Busack C, Marshall A (1991) Genetic analysis of yfp chinook salmon stocks. In: *Yakima hatchery experimental design. Progress Report to Bonneville Power Administration* eds. Phelps S, Seiler D), Portland, Oregon.

Busack CA, Currens KP (1995) Genetic risks and hazards in hatchery operations: Fundamental concepts and issues. *American Fisheries Science Symposium*, **15**, 71-80.

Campton DE (1995) Genetic effects of hatchery fish on wild populations of pacific salmon and steelhead: What do we really know? *Transactions of the American Fisheries Society*, **15**, 337-353.

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

Christie MR, Ford MJ, Blouin MS (2014) On the reproductive success of early-generation hatchery fish in the wild. *Evolutionary Applications*, **7**, 883-896.

Christie MR, Marine ML, French RA, Waples RS, Blouin MS (2012) Effective size of a wild salmonid population is greatly reduced by hatchery supplementation. *Heredity*, **109**, 254-260.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2go: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674-3676.

De Mita S, Thuillet AC, Gay L *et al.* (2013) Detecting selection along environmental gradients: Analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383-1399.

de Villemereuil P, Frichot E, Bazin E, Francois O, Gaggiotti OE (2014) Genome scan methods against more complex models: When and how much should we trust them? *Molecular Ecology*, **23**, 2006-2019.

Dittman AH, May D, Larsen DA, Moser ML, Johnston M, Fast D (2010) Homing and spawning site selection by supplemented hatchery- and natural-origin yakima river spring chinook salmon. *Transactions of the American Fisheries Society*, **139**, 1014-1028.

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 (n-e) from genetic data. *Molecular Ecology Resources*, **14**, 209-214.

Duchesne P, Bernatchez L (2002) An analytical investigation of the dynamics of inbreeding in multi-generation supportive breeding. *Conservation Genetics*, **3**, 47-60.

Fast DE, Bosch WJ, Johnston MV *et al.* (2015) A synthesis of findings from an integrated hatchery program after three generations of spawning in the natural environment. *North American Journal of Aquaculture*, **77**, 377-395.

Fleming IA, Gross MR (1994) Breeding competition in a pacific salmon (coho: *Oncorhynchus kisutch*): Measures of natural and sexual selection. *Evolution*, **48**, 637-657.

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.

Ford MJ (2002) Selection in captivity during supportive breeding may reduce fitness in the wild. *Conservation Biology*, **16**, 815-825.

Frankham R (2008) Genetic adaptation to captivity in species conservation programs. *Molecular Ecology*, **17**, 325-333.

Fraser DJ (2008) How well can captive breeding programs conserve biodiversity? A review of salmonids. *Evolutionary Applications*, **1**, 535-586.

Fritts AL, Scott JL, Pearsons TN (2007) The effects of domestication on the relative vulnerability of hatchery and wild origin spring chinook salmon (*oncorhynchus tshawytscha*) to predation. *Canadian Journal of Fisheries and Aquatic Sciences*, **64**, 813-818.

Götz S, Garcia-Gomez JM, Terol J *et al.* (2008) High-throughput functional annotation and data mining with the blast2go suite. *Nucleic Acids Research*, **36**, 3420-3435.

Harstad DL, Larsen DA, Beckman BR (2014) Variation in minijack rate among hatchery populations of columbia river basin chinook salmon. *Transactions of the American Fisheries Society*, **143**, 768-778.

Hess MA, Rabe CD, Vogel JL, Stephenson JJ, Nelson DD, Narum SR (2012) Supportive breeding boosts natural population abundance with minimal negative impacts on fitness of a wild population of chinook salmon. *Molecular Ecology*, **21**, 5236-5250.

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**, 23.

Jombart T (2008) *Adegenet:* A r package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403-1405.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *Bmc Genetics*, **11**, 15.

Jule KR, Leaver LA, Lea SEG (2008) The effects of captive experience on reintroduction survival in carnivores: A review and analysis. *Biological Conservation*, **141**, 355-363.

Karlsson S, Moen T (2010) The power to detect artificial selection acting on single loci in recently domesticated species. *BMC Research Notes*, **3**, 232.

Knudsen CM, Schroder SL, Busack C, Johnston MV, Pearsons TN, Strom CR (2008) Comparison of female reproductive traits and progeny of first-generation hatchery and wild upper yakima river spring chinook salmon. *Transactions of the American Fisheries Society*, **137**, 1433-1445.

Knudsen CM, Schroder SL, Busack CA *et al.* (2006) Comparison of life history traits between first-generation hatchery and wild upper yakima river spring chinook salmon. *Transactions of the American Fisheries Society*, **135**, 1130-1144.

Koskinen MT, Haugen TO, Primmer CR (2002) Contemporary fisherian life-history evolution in small salmonid populations. *Nature*, **419**, 826-830.

Laikre L, Schwartz MK, Waples RS, Ryman N, Ge MWG (2010) Compromising genetic diversity in the wild: Unmonitored large-scale release of plants and animals. *Trends in Ecology & Evolution*, **25**, 520-529.

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**, 10.

Larsen DA, Beckman BR, Cooper KA (2010) Examining the conflict between smolting and precocious male maturation in spring (stream-type) chinook salmon. *Transactions of the American Fisheries Society*, **139**, 564-578.

Larson WA, Seeb LW, Everett MV, Waples RK, Templin WD, Seeb JE (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of chinook salmon (*oncorhynchus tshawytscha*). *Evolutionary Applications*, **7**, 355-369.

Lichatowich JA, Mobrand LE (1995) Analysis of chinook salmon in the columbia river from an ecosystem perspective. In: *Research Report to Bonneville Power Administration*, Portland, Oregon.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of f_{st} outlier tests. *Molecular Ecology*, **23**, 2178-2192.

Lynch MOH, Martin (2001) Captive breeding and the genetic fitness of natural populations. *Conservation Genetics*, **2**, 363-378.

Mäkinen H, Vasemägi A, McGinnity P, Cross TF, Primmer CR (2015) Population genomic analyses of early-phase atlantic salmon (*salmo salar*) domestication/captive breeding. *Evolutionary Applications*, **8**, 93-107.

Mañanos EL, Anglade I, Chyb J, Saligaut C, Breton B, Kah O (1999) Involvement of gamma-aminobutyric acid in the control of gth-1 and gth-2 secretion in male and female rainbow trout. *Neuroendocrinology*, **69**, 269-280.

Mantua NJ, Hare SR, Zhang Y, Wallace JM, Francis RC (1997) A pacific interdecadal climate oscillation with impacts on salmon production. *Bulletin of the American Meteorological Society*, **78**, 1069-1079.

Martin SAM, Blaney S, Bowman AS, Houlihan DF (2002) Ubiquitin-proteasome-dependent proteolysis in rainbow trout (*oncorhynchus mykiss*): Effect of food deprivation. *Pflugers Archiv-European Journal of Physiology*, **445**, 257-266.

McGinnity P, Prodohl P, Ferguson K *et al.* (2003) Fitness reduction and potential extinction of wild populations of atlantic salmon, *salmo salar*, as a result of interactions with escaped farm salmon. *Proceedings of the Royal Society B-Biological Sciences*, **270**, 2443-2450.

Milot E, Perrier C, Papillon L, Dodson JJ, Bernatchez L (2013) Reduced fitness of atlantic salmon released in the wild after one generation of captive breeding. *Evolutionary Applications*, **6**, 472-485.

Mobrand LE, Barr J, Blankenship L *et al.* (2005) Hatchery reform in washington state: Principles and emerging issues. *Fisheries*, **30**, 11-23.

Naish KA, Seamons TR, Dauer MB, Hauser L, Quinn TP (2013) Relationship between effective population size, inbreeding and adult fitness-related traits in a steelhead (*oncorhynchus mykiss*) population released in the wild. *Molecular Ecology*, **22**, 1295-1309.

Naish KA, Taylor JE, Levin PS *et al.* (2008) An evaluation of the effects of conservation and fishery enhancement hatcheries on wild populations of salmon. In: *Advances in marine biology*, pp. 61-194. Elsevier Academic Press Inc, San Diego.

National Research Council (1996) *Upstream: Salmon and society in the pacific northwest* National Academy Press, Washington, D.C.

Neff BD, Garner SR, Pitcher TE (2011) Conservation and enhancement of wild fish populations: Preserving genetic quality versus genetic diversity. *Canadian Journal of Fisheries and Aquatic Sciences*, **68**, 1139-1154.

Panserat S, Hortopan GA, Plagnes-Juan E *et al.* (2009) Differential gene expression after total replacement of dietary fish meal and fish oil by plant products in rainbow trout (*oncorhynchus mykiss*) liver. *Aquaculture*, **294**, 123-131.

Paquet PJ, Flagg T, Appleby A *et al.* (2011) Hatcheries, conservation, and sustainable fisheries-achieving multiple goals: Results of the hatchery scientific review group's columbia river basin review. *Fisheries*, **36**, 547-561.

Pimm S, Raven P, Peterson A, Sekercioglu CH, Ehrlich PR (2006) Human impacts on the rates of recent, present, and future bird extinctions. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 10941-10946.

RASP (1992) Supplementation in the columbia river basin summary report series. In: *Final Report to Bonneville Power Administration*, Portland, Oregon.

Raymond M, Rousset F (1995) Genepop (version 1.2) - population-genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248-249.

Reisenbichler RR, Rubin SP (1999) Genetic changes from artificial propagation of pacific salmon affect the productivity and viability of supplemented populations. *Ices Journal of Marine Science*, **56**, 459-466.

Ryman N, Laikre L (1991) Effects of supportive breeding on the genetically effective population size. *Conservation Biology*, **5**, 325-329.

Scheuerell MD, Buhle ER, Semmens BX, Ford MJ, Cooney T, Carmichael RW (2015) Analyzing large-scale conservation interventions with bayesian hierarchical models: A

case study of supplementing threatened pacific salmon. *Ecology and Evolution*, **5**, 2115-2125.
Schroder SL, Knudsen CM, Pearsons TN *et al.* (2010) Behavior and breeding success of wild and first-generation hatchery male spring chinook salmon spawning in an artificial stream. *Transactions of the American Fisheries Society*, **139**, 989-1003.
Schroder SL, Knudsen CM, Pearsons TN *et al.* (2008) Breeding success of wild and first-generation hatchery female spring chinook salmon spawning in an artificial stream. *Transactions of the American Fisheries Society*, **137**, 1475-1489.
Schwartz MK, Luikart G, Waples RS (2007) Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology & Evolution*, **22**, 25-33.
Seiliez I, Panserat S, Skiba-Cassy S *et al.* (2008) Feeding status regulates the polyubiquitination step of the ubiquitin-proteasome-dependent proteolysis in rainbow trout (*oncorhynchus mykiss*) muscle. *Journal of Nutrition*, **138**, 487-491.
Storey JD (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **64**, 479-498.
Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671-688.
Suzuki K, Kawauchi H, Nagahama Y (1988) Isolation and characterization of two distinct gonadotropins from chum salmon pituitary glands. *General and Comparative Endocrinology*, **71**, 292-301.
Therkildsen NO, Hemmer-Hansen J, Als TD *et al.* (2013) Microevolution in time and space: Snp analysis of historical DNA reveals dynamic signatures of selection in atlantic cod. *Molecular Ecology*, **22**, 2424-2440.
Waples RS (1999) Dispelling some myths about hatcheries. *Fisheries*, **24**, 12-21.
Waples RS (2002) Effective size of fluctuating salmon populations. *Genetics*, **161**, 783-791.
Waples RS (2005) Genetic estimates of contemporary effective population size: To what time periods do the estimates apply? *Molecular Ecology*, **14**, 3335-3352.
Waples RS, Antao T, Luikart G (2014) Effects of overlapping generations on linkage disequilibrium estimates of effective population size. *Genetics*, **197**, 769-U603.
Waples RS, Drake J (2004) Risk/benefit considerations for marine stock enhancement: A pacific salmon perspective. In: *Stock enhancement and sea ranching: Developments, pitfalls and opportunities* (eds. Leber KM, Kitada S, Blankenship HL, Svåsand T). Blackwell Publishing Ltd, Oxford, UK.
Weir BS, Cockerham CC (1984) Estimating f-statistics for the analysis of population structure. *Evolution*, **38**, 1358-1370.
Williams SE, Hoffman EA (2009) Minimizing genetic adaptation in captive breeding programs: A review. *Biological Conservation*, **142**, 2388-2400.
Wright S (1931) Evolution in mendelian populations. *Genetics*, **16**, 0097-0159.

Chapter 2. What can genomics tell us about the success of enhancement programs in anadromous Chinook salmon? A comparative analysis across four generations[2]

2.1 ABSTRACT

Population enhancement through the release of cultured organisms can be an important tool for marine restoration. However, there has been considerable debate about whether releases effectively contribute to conservation and harvest objectives, and whether cultured organisms impact the fitness of wild populations. Pacific salmonid hatcheries on the West Coast of North America represent one of the largest enhancement programs in the world. Molecular-based pedigree studies on one or two generations have contributed to our understanding of the fitness of hatchery-reared individuals relative to wild individuals, and tend to show that hatchery fish have lower reproductive success. However, interpreting the significance of these results can be challenging because the long-term genetic and ecological effects of releases on supplemented populations are unknown. Further, most salmon pedigree studies have been opportunistic, rather than hypothesis driven, and have not provided information on "best case" management scenarios. Here, we present a comparative, experimental approach based on genome-wide surveys of changes in diversity in two hatchery lines founded from the same population. We demonstrate that gene flow with wild individuals can reduce divergence from the wild source population over four generations. We also report evidence for consistent genetic changes in a closed hatchery

[2] This chapter complements Chapter 1 by adding a fourth generation to genome-wide comparisons between the integrated and segregated hatchery lines. It was published as Supplementary Material for Bernatchez *et al*. 2017, Harnessing the power of genomics to secure the future of seafood, *Trends in Ecology and Evolution* **32**, 665-680.

population that can be explained by both genetic drift and domestication selection. The results of this study suggest that genetic risks can be minimized over at least four generations with appropriate actions, and provide empirical support for a decision-making framework that is relevant to the management of hatchery populations.

2.2 INTRODUCTION

Enhancement, the release of cultured organisms to increase population abundance, is an important fishery management tool (Lorenzen *et al.* 2010). But genetic risks associated with artificial propagation are well known and may compromise the wild populations that enhancement is intended to support (Naish *et al.* 2008; Laikre *et al.* 2010). Supportive breeding programs are a form of enhancement used extensively in the management of Pacific salmon in North America. Such programs aim to increase population sizes by rearing a fraction of juveniles in captivity and then releasing them into the natural environment along with their wild-born conspecifics (Ryman & Laikre 1991). Concerted efforts have been directed at mitigating the effects of domestication selection, genetic drift and inbreeding (Mobrand *et al.* 2005) associated with these programs, because in many cases populations have not recovered and cannot support sustainable fisheries (Naish *et al.* 2008; Beamish *et al.* 2010; Scheuerell *et al.* 2015). Practical recommendations to mitigate genetic risks have focused on theoretical models that examine the influence of gene flow in reducing divergence between cultured and wild populations (Duschene & Bernatchez 2002; Ford 2002; Baskett & Waples 2013). Specifically, the intentional use of natural-origin broodstock in the creation of the hatchery population in each generation may reduce risks, especially when gene flow from the hatchery to the wild population is limited

(Mobrand *et al.* 2005). Such "managed gene flow" has seen widespread adoption in the Pacific Northwest of the USA (Paquet *et al.* 2011), but few practical examples on their efficacy exist.

An ideal way to test whether managed gene flow is effective at reducing genetic divergence between hatchery and wild populations is to empirically compare cultured populations with and without gene flow. Such a comparison would provide results on the range of possible outcomes of these management approaches, and would be especially informative if conducted longitudinally. The use of population genomic approaches provides a way to survey temporal changes in genetic divergence, to measure the rate of change with each generation since founding, and to identify factors driving divergence. We previously conducted such a study in two populations of Chinook salmon (*Oncorhynchus tshawytscha*) released from a hatchery on a tributary of the Columbia River (Waters *et al.* 2015). Both hatchery populations were founded from the same source population; however, one population remained integrated with the wild and used only wild-born broodstock in each generation, while the second hatchery population was maintained separately and received no gene flow from the wild. Our earlier results over three generations revealed little change in the integrated line compared to the founding population. Most of the genetic divergence in the segregated line could be attributed to genetic drift, but there was also evidence for directional selection at specific locations in the genome. However, it is unclear over how many generations managed gene flow may be effective at mitigating genetic risks, because processes occurring in the wild could mitigate or exacerbate the effects over time (Ford 2002; Baskett & Waples 2013). Here we aimed to test whether the use of natural-origin broodstock was effective at reducing divergence over several generations by extending our earlier study for an additional fourth generation.

2.3 METHODS

A spring-run Chinook salmon hatchery program was initiated in 1997 at the Cle Elum Supplementation and Research Facility (CESRF, Figure 2.1) to supplement the declining upper Yakima River Chinook salmon population while minimizing possible genetic and ecological risks associated with supportive breeding. Local, wild adults were collected for broodstock from 1997 to 2002 as they passed the Roza Dam Adult Monitoring Facility (Roza Dam, Figure 2.1). Adults were then transferred to CESRF and held until spawning. Eggs and juveniles were reared in the hatchery for approximately 18 months before they began their migration to the ocean. Adult hatchery fish first returned to the Yakima River in 2001 and were allowed to spawn naturally. In 2002, both wild and hatchery-origin adults were spawned at CESRF to create two contrasting hatchery lines. The integrated (INT) line is derived only from wild or natural-origin adults, and all fish from this line are allowed to spawn naturally. Here, natural-origin fish are those that were born in the river but may have some hatchery ancestry. The segregated (SEG) line, however, uses only hatchery-origin broodstock, and no fish from this line are allowed to spawn in the river.

Tissues for DNA were sampled from adults of both hatchery lines in 2014 during spawning at CESRF and stored in 100% ethanol. These adults represent the fourth (F_4) generation of each line. DNA was extracted using DNeasy Blood & Tissue kits (Qiagen, Valencia, CA, USA) following the animal tissue protocol. Restriction site-associated (RAD) libraries (Baird *et al.* 2008) were prepared using the restriction enzyme *SbfI* and sequenced on the Illumina HiSeq 2000 platform with 36 individuals per lane. All raw RAD sequences from the F_4 generation were combined with raw data from our previous comparative analysis (P_1 founders and F_1-F_3 generations, Waters *et al.* 2015). Filtering and genotyping were performed following

Waters *et al*. (2015), with two additional steps to improve data quality. First, loci were removed if more than 50% of individuals in any population were not genotyped. Then, loci were removed if they did not meet Hardy-Weinberg equilibrium conditions (q-value < 0.05) in more than one population, as determined by the Monte Carlo procedure with $1x10^5$ permutations in the R-package *adegenet* (v. 1.3-9, Jombart 2008). *Q*-values were computed using the R-package *qvalue* (v. 1.28.0, Storey 2002).

As the aim of the present study was to extend previous comparisons between the integrated and segregated hatchery lines by another generation, genetic change was evaluated using the same methods described in Waters *et al*. (2015). Population-level genetic change between each generation of the hatchery lines was evaluated using measures of F_{ST}, computed in *Genepop* (v. 4.1, Raymond & Rousset 1995), and a discriminant analysis of principal components (DAPC), conducted in the R-package *adegenet*. The relative effect of genetic drift within each hatchery line was determined using estimates of effective numbers of breeders, N_b. Temporal and linkage disequilibrium (LD) estimates of N_b were computed with N_E Estimator (v. 2.01, Do *et al.* 2014) using only four-year-old adults, which represented a single cohort of individuals. Steps taken to reduce potential bias in N_b estimates due to selection, overlapping generations, and fluctuating population size were identical to those of Waters *et al.* (2015). Lastly, loci and genomic regions exhibiting signals of diversifying selection in the hatchery lines were identified using three independent tests: F_{TEMP} (Therkildsen *et al.* 2013), *Bayescan* (Foll & Gaggiotti 2008), and a sliding-window approach (Brieuc *et al.* 2015; Waters *et al.* 2015). We focused on loci and regions that were identified by multiple tests and were divergent across multiple generations.

2.4 RESULTS

Tissues from 72 individuals (36 from each line) were sequenced from the F_4 generation. The raw data was combined with RAD sequences from the previous generations and filtered, yielding 9266 bi-allelic RAD loci with minor allele frequencies >0.05 in at least one population and less than 50% missing genotypes within each population. A total of 465 individuals from the five generations were genotyped at >50% of these loci and retained for analyses (Tables S2.1, S2.2; Waters *et al.* 2017). Tests of HWE identified 158 loci that significantly deviated from expectations in more than one population. Following removal of these loci, the final data set comprised 9108 loci (Table S2.1; Waters *et al.* 2017), including 4214 loci that aligned to the Chinook salmon linkage map (Brieuc *et al.* 2014). Population-level divergence of the two hatchery lines followed previously documented trends (Waters *et al.* 2015). Values of pairwise F_{ST} between the lines and the P_1 founders was approximately four times higher in the F_4 SEG population (F_{ST}=0.0125, p<0.001, Table S2.3; Waters *et al.* 2017) than in the F_4 INT population (F_{ST} =0.0033, p<0.001). Divergence between the two hatchery lines in the F_4 generation also continued to increase (F_{ST} =0.0126, p<0.001). Patterns of genetic change were further supported by a discriminant analysis of principal components, conducted on the first 63 PCs as recommended by the *optim.a.score* function in *adegenet*. The segregated line diverged from the P_1 founders and integrated line over time along the first discriminant function; this axis explained 59.2% of the retained variation (Figure 2.2). Genetic change between the later generations of the integrated line and the P_1 founders was evident along the second discriminant function, which explained 16.6% of the retained variation.

Bias-adjusted LD and temporal estimates of effective number of breeders, N_b, supported earlier results and suggested that the relative effect of genetic drift was much greater in the

segregated line than in the integrated line. The LD estimate of N_b in the F_4 INT population was nearly eight times higher than that obtained in the F_4 SEG population (Figure 2.3, Tables S2.4a, S2.5; Waters *et al.* 2017). The temporal N_b estimates, which applied to the entire sampling period of 1998-2014, were also markedly different between the two lines (Figure 2.3, Table S2.4b; Waters *et al.* 2017). While the average broodstock size of the integrated line (363 ± 15) exceeded that of the segregated line (85 ± 15), the difference was not sufficient to explain the higher N_b of the integrated line (Table S2.6; Waters *et al.* 2017).

In addition, the N_b estimates for the F_4 generation revealed a result that was not apparent from the census data alone. The ratio of effective number of breeders to census size (N_b/ N_{census}) declined from 0.21 (95% CI: 0.20-0.23) in the F_3 INT sample to 0.04 (95% CI: 0.03-0.04) in the F_4 INT sample, despite the fact that the census sizes increased from 3364 to 8374 adults. This result is important because the N_b/N_{census} ratio provides a metric for understanding factors which cause deviations from $N_b=N_{census}$ (e.g. variance in reproductive success) and affect genetic variation over time.

Three independent tests identified loci and genomic regions that exhibited signals of diversifying selection. The F_{TEMP} method identified 78 loci that exceeded neutral expectations in the integrated line and 198 in the segregated line (Table S2.7; Waters *et al.* 2017). Thirty-five loci were outliers in both hatchery lines. *Bayescan*, conducted using all populations combined, identified 120 loci putatively under diversifying selection (Table S2.7; Waters *et al.* 2017). There was considerable overlap between *Bayescan* and F_{TEMP}, as 48 and 72 *Bayescan* outliers were also identified by F_{TEMP} in the integrated and segregated lines, respectively. Genomic regions that exhibited significantly elevated levels of divergence compared to the P_1 founders were identified in both hatchery lines by sliding window analyses (Table S2.8; Waters *et al.* 2017). However,

divergence in the segregated line was more consistent across the F_1, F_2, F_3, and F_4 generations than in the integrated line. For example, seven regions were significantly elevated in at least three generations of the segregated line while none were observed in the integrated line (e.g. Figure 2.4). Five of these regions also contained outlier loci identified by F_{TEMP} and *Bayescan* (e.g. Figure 2.4), providing further support that selection – likely due to continued exposure to the hatchery environment – has also contributed to the higher levels of divergence observed in the segregated line. Previous work has identified genes in such regions of overlap that may be targeted by selection in captivity (Waters *et al.* 2015).

2.5 DISCUSSION

Here, we have demonstrated the utility of genomic-based methods to test alternative management approaches for population enhancement and to monitor fine scale genetic changes in populations over several generations. Many theoretical studies have indicated that ongoing gene flow between hatchery and wild fish may ultimately compromise the fitness of the natural population (Ford 2002; Baskett & Waples 2013). However, the degree to which the natural population is affected depends on many factors that likely fluctuate over time, such as selection intensity, the proportion of wild-origin individuals on the spawning grounds, reproductive rate in the hatchery and wild, and carrying capacity of the natural system. Thus, our multigenerational findings extend complementary studies that evaluate reproductive success in single populations over one or two generations (Christie *et al.* 2014). The results from the fourth hatchery generation largely supported observations from a previously published longitudinal study (Waters *et al.* 2015). Little genetic change occurred in the integrated hatchery line, which frequently exchanged migrants with the founding wild population. In contrast, consistent

temporal trends in divergence were documented in the segregated hatchery line, which is maintained as a closed population. Such consistency was observed on the population level and at specific genomic regions, despite the fact that environmental change likely occurred during the sixteen years over which this study was conducted. This result might be explained by domestication selection imposed by the relatively uniform hatchery environment on the segregated hatchery population. Genomic regions exhibiting potential signals of domestication selection can be further examined to identify candidate genes (e.g. Waters *et al.* 2015) and mechanisms underlying genetic adaptation to captivity, and to inform management practices to possibly reduce this risk.

Notably, extending the earlier study by another generation also revealed fluctuations in N_b/ N_{census} that would otherwise have been missed. We previously reported N_b/N_{census} ratios of 0.11 (95% CI: 0.10-0.12) and 0.21 (95% CI: 0.20-0.23) for the F_2 and F_3 INT samples, respectively, which reflected the first two generations of naturally-spawning adults that included hatchery fish from the integrated line. These estimates show a positive trend in N_b/N_{census}, which, if taken alone, could possibly be attributed to successful supplementation efforts. However, it is important to acknowledge temporal variability, and the decline of N_b/N_{census} to 0.04 in the F_4 INT sample may have two explanations. The first is that the results may indicate the influence of the Ryman-Laikre effect (Ryman & Laikre 1991), where supportive breeding reduces the effective size of a wild population. Alternatively, temporal fluctuations in N_b/N_{census} could simply reflect changes in the natural environment that influence demographic factors; the observed ratios of 0.04-0.21 are within the range of those documented in natural populations of many species (including salmonids; Frankham 1995; Naish *et al.* 2013). It is impossible to identify the true source(s) of the observed fluctuations, particularly since there is no wild control population for

comparison. Nevertheless, our results emphasize the importance of continued monitoring and the viability of integrating processes affecting the productivity of natural systems with enhancement efforts. Finally, while this study does not evaluate fitness directly and lacks an unsupplemented control population, rates of genetic divergence measured here provide a range of multigenerational outcomes for contrasting management regimes. These comparative findings, in turn, can assist managers and policy-makers when assessing the relative benefits and risks of conservation decisions, particularly in cases where population recovery may depend on supportive breeding.

2.6 ACKNOWLEDGEMENTS

We thank everyone who was involved in establishing CESRF, shaping its research direction, and sampling broodstock, including Levi George, Melvin Sampson, Steve Schroder, Craig Busack, past and present members of the Independent Scientific Review Panel, and the Yakama Nation Tribal Council. We are grateful to Maren Wellenreuther and Louis Bernatchez, and the OECD Co-operative Research Programme for the opportunity to present this research at the World Fisheries Congress in Korea. Funding for this study was provided by NOAA Fisheries/Federal Columbia River Power System (FCRPS) Biological Opinion Remand Funds (to K.A.N. and J.J.H.), Washington Sea Grant (Award NA14OAR4170078 to K.A.N), and the Hall Conservation Genetics Research Award from the University of Washington (to C.D.W.).

2.7 FIGURES

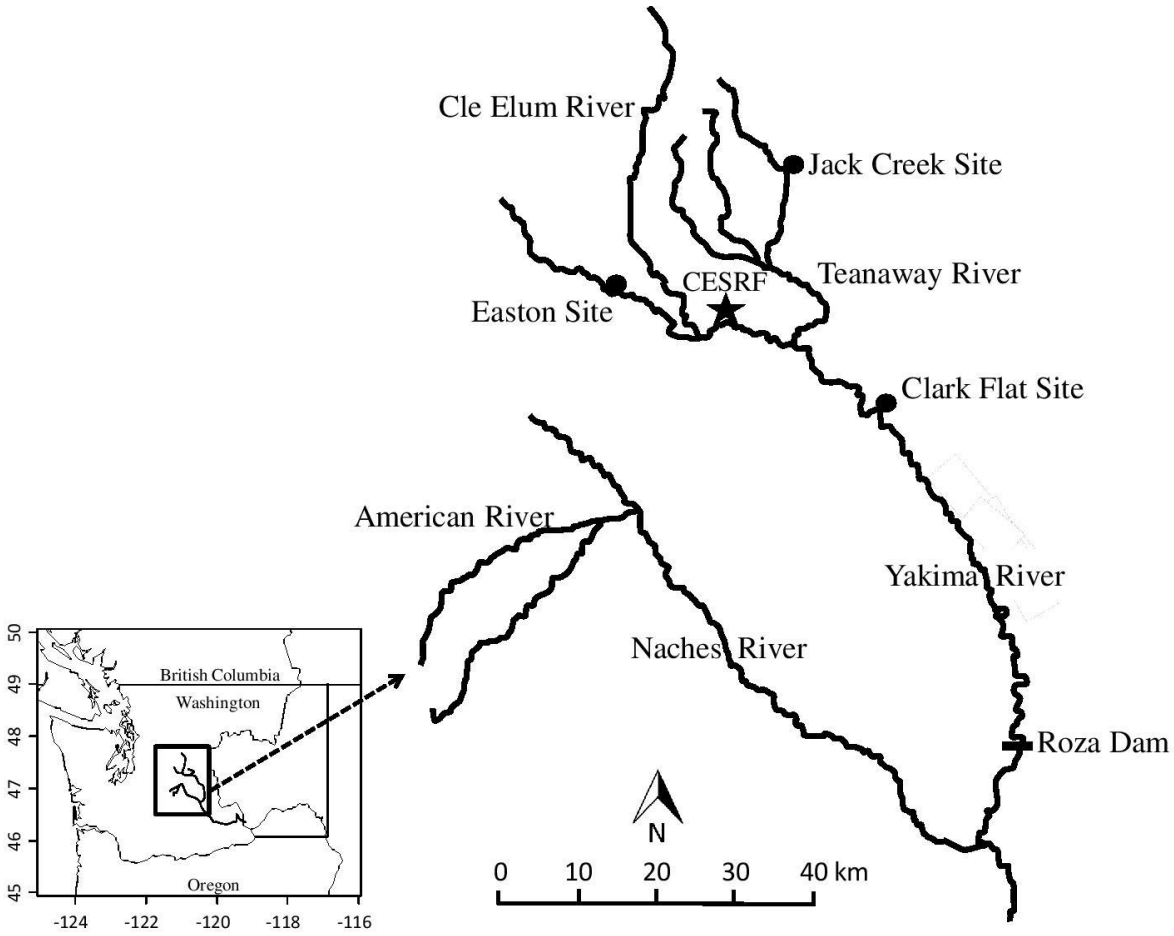

Figure 2.1 – From Waters *et al*. (2015). Map of the Yakima River system. The upper Yakima Chinook salmon population is the target of the Cle Elum Supplementation and Research Facility (CESRF). All returning adults are sampled at Roza Dam and allowed to spawn naturally (natural origin and integrated line fish) or are removed from the system (all segregated line fish). Spawning and rearing for the hatchery lines occurs at CESRF. Prior to outmigration in spring, juveniles are transferred to the Easton, Jack Creek, and Clark Flat acclimation sites, where they are held for approximately two months before volitional release.

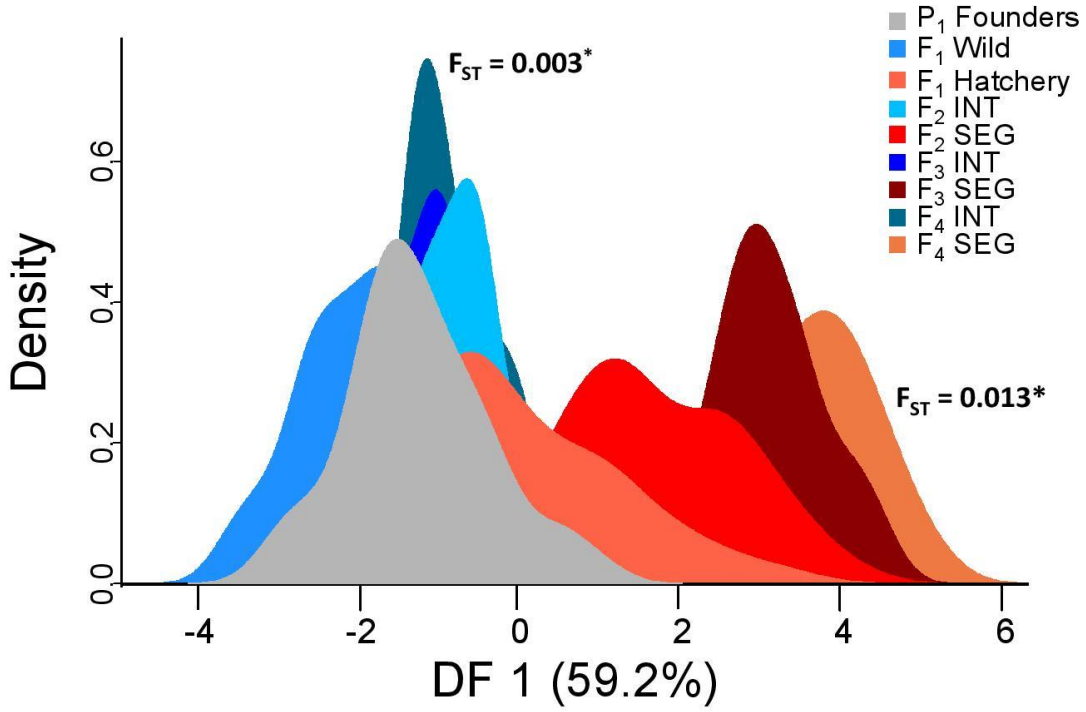

Figure 2.2 – Density plot of individuals along the first discriminant function from the discriminant analysis of principal components (DAPC) for the wild founders (P_1 Founders, black) and four generations of the integrated (INT, blue colors) and segregated (SEG, red colors) hatchery lines. Pairwise F_{ST} values for the F_4 generation compared to the P_1 founders are shown for each hatchery line.

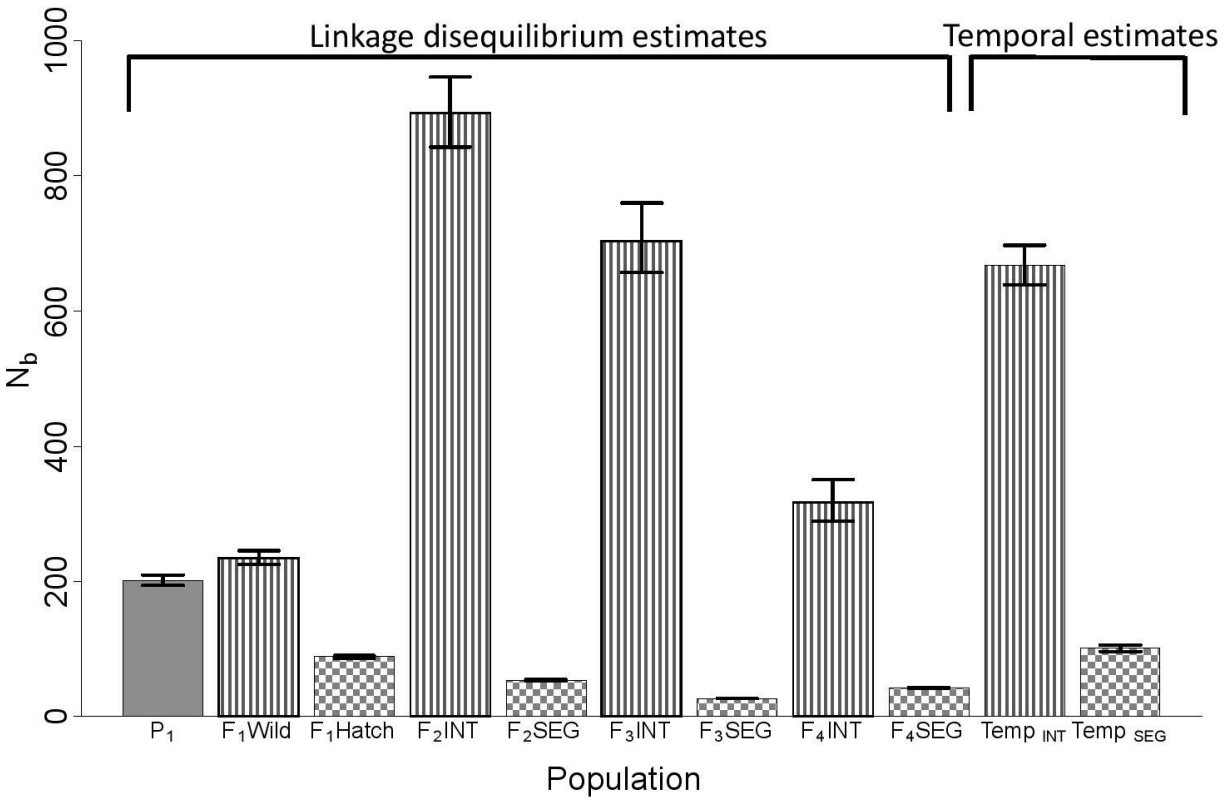

Figure 2.3 – Estimates of effective number of breeders, N_b, and 95% confidence intervals produced by the linkage disequilibrium (LD) and temporal methods. The LD method enables estimation of N_b for every generation, while a single estimate for the sampling period is produced by the temporal method. LD estimates are adjusted for physical linkage and other potential biases as described in Waters *et al*. (2015).

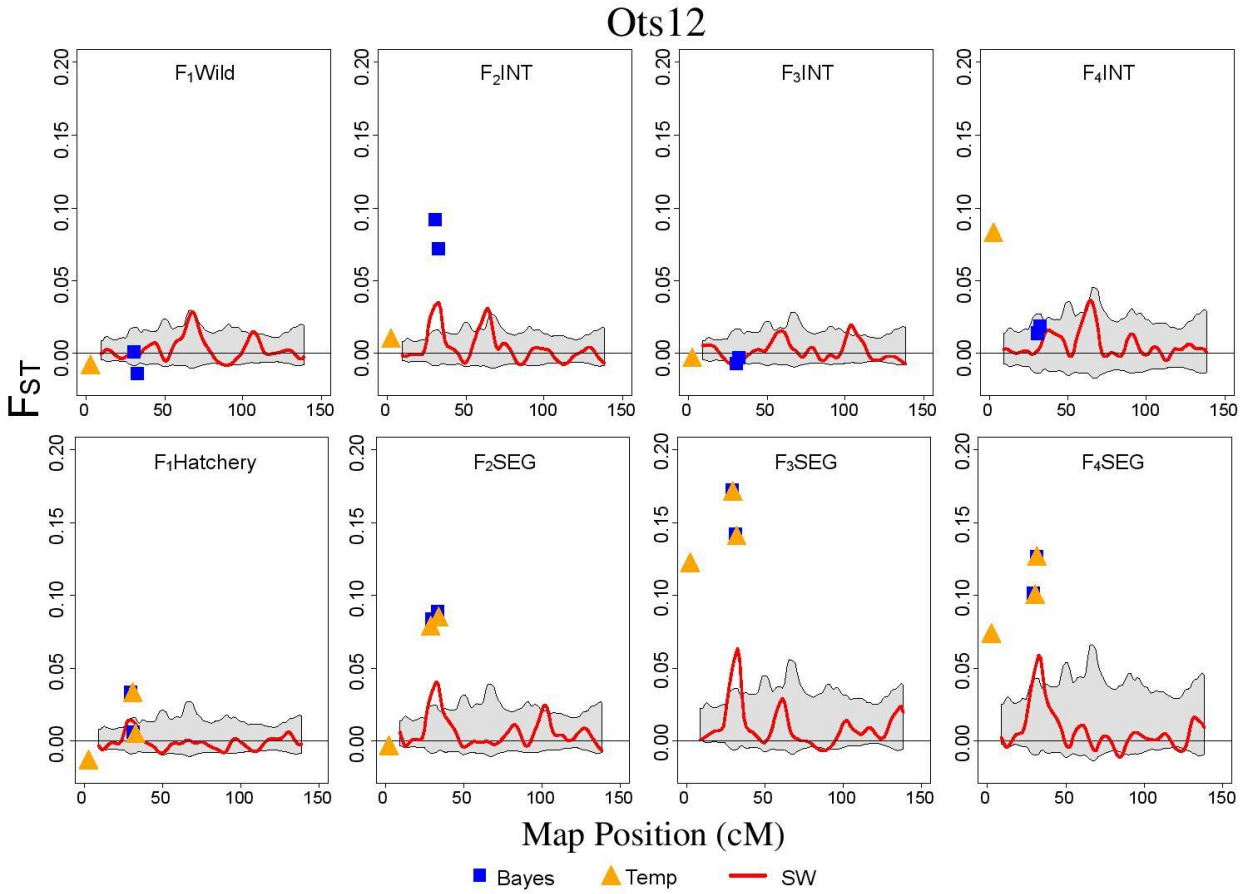

Figure 2.4 – Loci and regions of the genome showing signatures of adaptive divergence, based on pairwise F_{ST} compared to the P_1 founders, on chromosome Ots12 for the integrated (top panel) and segregated (bottom panel) hatchery lines through the F_1, F_2, F_3, and F_4 generations. Blue squares are loci that were identified as outliers with *Bayescan* and orange triangles are outliers identified by F_{TEMP}. The red line represents the kernel smoothed moving average of F_{ST} and the grey shaded area is the 95% confidence interval.

2.8 SUPPLEMENTARY MATERIAL

S2 – Supplementary tables containing metadata and results of Chapter 2

S2.1 – Lists of the 9108 RAD loci and 465 individuals analyzed in this study.

S2.2 – Numbers of adults used as broodstock, the number of adults sampled in this study, the final sample size retained after filtering, the observed and expected heterozygosity for each population, and the number of loci that deviated from Hardy-Weinberg equilibrium.

S2.3 – Pairwise population differentiation (F_{ST}) compared to the P_1 Founders and between the two hatchery lines within each generation.

S2.4 – (a) Spawning groups and census sizes to which estimates of effective number of breeders apply, the proportions of adults that were of hatchery origin, and the years and populations from which estimates of effective number of breeders were obtained using the linkage disequilibrium (LD) method. LD estimates with 95% confidence intervals are shown. (b) Effective number of breeders for both hatchery lines from 1998-2014 estimated using the temporal method.

S2.5 – Bias calculations for estimates of effective number of breeders from the LD method.

S2.6 – Estimated contribution of naturally spawning fish to effective number of breeders in the integrated line.

S2.7 – List of outlier loci identified by *Bayescan* and F_{TEMP}.

S2.8 – Regions identified by sliding window analyses as showing elevated divergence.

2.9 REFERENCES

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid snp discovery and genetic mapping using sequenced rad markers. *Plos One*, **3**, 7.

Baskett ML, Waples RS (2013) Evaluating alternative strategies for minimizing unintended fitness consequences of cultured individuals on wild populations. *Conservation Biology*, **27**, 83-94.

Beamish RJ, Sweeting RM, Lange KL, Noakes DJ, Preikshot D, Neville CM (2010) Early marine survival of coho salmon in the strait of georgia declines to very low levels. *Marine and Coastal Fisheries*, **2**, 424-439.

Brieuc MSO, Ono K, Drinan DP, 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.

Christie MR, Ford MJ, Blouin MS (2014) On the reproductive success of early-generation hatchery fish in the wild. *Evolutionary Applications*, **7**, 883-896.

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

Duschene P, Bernatchez L (2002) An analytical investigation of the dynamics of inbreeding in multi-generation supportive breeding. *Conservation Genetics*, **3**, 45-58.

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.

Ford MJ (2002) Selection in captivity during supportive breeding may reduce fitness in the wild. *Conservation Biology*, **16**, 815-825.

Frankham R (1995) Effective population size/adult population size ratios in wildlife: A review. *Genetical Research*, **66**, 95-107.

Jombart T (2008) *Adegenet:* A r package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403-1405.

Laikre L, Schwartz MK, Waples RS, Ryman N, GeM_Working_Group (2010) Compromising genetic diversity in the wild: Unmonitored large-scale release of plants and animals. *Trends in Ecology & Evolution*, **25**, 520-529.

Lorenzen K, Leber KM, Blankenship HL (2010) Responsible approach to marine stock enhancement: An update. *Reviews in Fisheries Science*, **18**, 189-210.

Mobrand LE, Barr J, Blankenship L *et al.* (2005) Hatchery reform in washington state: Principles and emerging issues. *Fisheries*, **30**, 11-23.

Naish KA, Seamons TR, Dauer MB, Hauser L, Quinn TP (2013) Relationship between effective population size, inbreeding and adult fitness-related traits in a steelhead (*oncorhynchus mykiss*) population released in the wild. *Molecular Ecology*, **22**, 1295-1309.

Naish KA, Taylor JE, Levin PS *et al.* (2008) An evaluation of the effects of conservation and fishery enhancement hatcheries on wild populations of salmon. *Advances in Marine Biology*, **53**, 61-194.

Paquet PJ, Flagg T, Appleby A *et al.* (2011) Hatcheries, conservation, and sustainable fisheries-achieving multiple goals: Results of the hatchery scientific review group's columbia river basin review *Fisheries*, **36**, 547-561.

Raymond M, Rousset F (1995) Genepop (version 1.2) - population-genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248-249.

Ryman N, Laikre L (1991) Effects of supportive breeding on the genetically effective population size. *Conservation Biology*, **5**, 325-329.

Scheuerell MD, Buhle ER, Semmens BX, Ford MJ, Cooney T, Carmichael RW (2015) Analyzing large-scale conservation interventions with bayesian hierarchical models: A case study of supplementing threatened pacific salmon. *Ecology and Evolution*, **5**, 2115-2125.

Storey JD (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **64**, 479-498.

Therkildsen NO, Hemmer-Hansen J, Als TD *et al.* (2013) Microevolution in time and space: Snp analysis of historical DNA reveals dynamic signatures of selection in atlantic cod. *Molecular Ecology*, **22**, 2424-2440.

Waters CD, Hard JJ, Brieuc MSO *et al.* (2015) Effectiveness of managed gene flow in reducing genetic divergence associated with captive breeding. *Evolutionary Applications*, **8**, 956-971.

Waters CD, Hard JJ, Brieuc MSO *et al.* (2017) What can genomics tell us about the success of enhancement programs in anadromous chinook salmon? A comparative analysis across four generations. Supplementary material for bernatchez et al. (2017) harnessing the power of genomics to secure the future of seafood. *Trends in Ecology & Evolution*, 665-680.

Chapter 3. Genetic and phenotypic effects of inbreeding across two different hatchery management regimes in Chinook salmon

3.1 ABSTRACT

The consequences of inbreeding in captive breeding programs have received insufficient attention compared to domestication selection, partly because detection of inbred individuals traditionally relied on pedigrees. However, genomic approaches now enable inbreeding coefficients to be accurately estimated from molecular markers. Here, we quantified levels of pairwise relatedness and inbreeding, as well as the effects of inbreeding on eight fitness-related traits, in two hatchery populations of adult Chinook salmon across four generations using 6805 restriction-site associated (RAD) loci. The hatchery populations were derived from the same source but are now managed as two lines that are integrated with and segregated from the founding wild stock. Relatedness and inbreeding were then compared between the two management strategies to evaluate the effectiveness of integrated management, or managed gene flow, to reduce risks of inbreeding in captive populations. Levels of inbreeding were similar between the two hatchery lines in the first, third, and fourth generations, despite 3- to 27-fold differences in estimates of effective numbers of breeders. However, inbreeding in the segregated line was significantly higher in the second generation. The segregated line also had slight but significantly lower levels of relatedness than the integrated line in the first generation but significantly higher levels in the third and fourth generations. Inbreeding coefficient did not affect fecundity, reproductive effort, return timing, and fork length of surviving adults. In contrast, inbreeding significantly affected spawn timing, weight, condition factor, and daily

growth coefficient in these fish, although the effects varied by sex, hatchery line, and generation. While our results indicate that integrated management may reduce the genetic risks of inbreeding, they also suggest that short-term risks may not be severe in small, segregated hatchery populations. The effects of inbreeding on fitness varied and thus require further exploration, particularly at earlier life stages.

3.2 INTRODUCTION

The use of captive breeding as a tool for population and species recovery remains controversial (Snyder *et al.* 1996; Bowkett 2009), in part due to accompanying genetic and ecological risks. One genetic risk, inbreeding, occurs when related individuals interbreed and is more likely to occur in small populations (Frankham *et al.* 2002; Keller & Waller 2002). Numerous factors can reduce census and genetic population sizes within captive breeding programs and thus increase the risk of inbreeding. Such factors include limited rearing capacity and numbers of breeders, differences in family sizes released from the program, variance in reproductive success, and unequal sex ratios (Allendorf 1993; Frankham *et al.* 2002; Naish *et al.* 2008). Inbreeding may lead to inbreeding depression, which is a reduction in fitness caused by an increased expression of deleterious recessive alleles or the loss of heterozygosity at loci with a heterozygote advantage (Charlesworth & Willis 2009). Inbreeding depression has been documented across an array of plants and animals (Keller & Waller 2002; Woodworth *et al.* 2002; Brekke *et al.* 2010; Elias *et al.* 2013; Hammerly *et al.* 2013; Hoffman *et al.* 2014; Hedrick & Garcia-Dorado 2016; Huisman *et al.* 2016) and, in addition to affecting individual fitness, can reduce population productivity and increase the risk of extinction (Saccheri *et al.* 1998; Frankham 2005; O'Grady *et al.* 2006).

Despite their widespread recognition and importance, the genetic, phenotypic, and demographic effects of inbreeding in natural populations remain unclear (Kardos *et al.* 2016). One of the key reasons for this limited understanding is that detecting inbred individuals has traditionally relied on full pedigrees over at least three generations; such pedigrees are typically intractable in wild and long-lived populations. Further, inbreeding coefficients derived from pedigrees may lack precision if the pedigree comprises few generations or is incomplete (Reid *et al.* 2014; Taylor *et al.* 2015), and individuals with the same pedigree can have different inbreeding coefficients due to physical linkage and recombination (Kardos *et al.* 2015). Marker-based estimates of inbreeding, until recently, were also imprecise because they were obtained using small numbers of loci (Hoffman *et al.* 2014; Taylor 2015). However, current genomic approaches provide thousands of molecular markers across the genome and permit the accurate and precise estimation of individual inbreeding coefficients within natural populations (Kardos *et al.* 2015). Incorporating these genomic estimates into studies of inbreeding depression have indicated that the negative fitness effects of inbreeding may be more severe than previously thought (Hoffman *et al.* 2014; Hedrick & Garcia-Dorado 2016; Huisman *et al.* 2016). Therefore, there is a pressing need to quantify the effects of inbreeding using genomic approaches in other populations, especially those of conservation concern.

Pacific salmon on the West Coast of North America may be particularly vulnerable to the genetic and phenotypic effects of inbreeding. Indeed, many wild populations have declined over the last century due to anthropogenic disturbances (National Research Council 1996; Gustafson *et al.* 2007) and may have an increased risk of inbreeding due to small population sizes. Additionally, conservation-focused hatchery populations, which are intended to supplement depressed wild stocks, may inadvertently have higher levels of inbreeding due to management

practices (e.g. broodstock collection and spawning procedures) and limited rearing capacity (Naish *et al.* 2008). Numerous studies have documented adverse fitness effects in salmonids that were attributed to inbreeding (Calaprice 1969; Ryman 1970; Kincaid 1976; Hard *et al.* 2000; Wang *et al.* 2002; Thrower & Hard 2009; Naish *et al.* 2013). However, the potential consequences of inbreeding over multiple generations can now be more precisely estimated in these species with genomic approaches.

The Cle Elum Supplementation and Research Facility (CESRF) in Cle Elum, Washington, U.S.A. provides an excellent opportunity to evaluate the genetic and phenotypic effects of inbreeding in captive-reared Pacific salmon that are intended to support conservation efforts. CESRF was initiated in 1997 in response to declining annual returns of anadromous wild spring Chinook salmon (*Oncorhynchus tshawytscha*) to the upper Yakima River, a tributary of the Columbia River. The program was designed to test whether supportive breeding could increase harvest and production in the upper Yakima River Chinook population while minimizing ecological and genetic risks associated with captive rearing (RASP 1992). To achieve this goal, the hatchery population was divided into a "segregated" (SEG) line, which is not allowed to interbreed with the source population, and an "integrated" (INT) line, which has unrestricted gene flow with wild individuals (Fast *et al.* 2015). Notably, tissue samples for DNA and extensive phenotypic data, including fork length, weight, and return timing, have been collected from every adult fish spawned in the hatchery since its inception, forming one of the most comprehensive collections available for any Pacific salmon population, wild or hatchery.

The propagation of the integrated and segregated hatchery lines at CESRF also permits the testing of an alternative management approach, managed gene flow, aimed at reducing the risks of inbreeding in captive-reared populations. Managed gene flow in the form of integrated

hatchery management has been widely implemented throughout the Pacific Northwest (Mobrand *et al.* 2005; Paquet *et al.* 2011), because theoretical studies suggest that it can mitigate the risks of domestication selection, inbreeding, and genetic drift associated with captive rearing (Lynch & O'Hely 2001; Duchesne & Bernatchez 2002; Ford 2002; Mobrand *et al.* 2005; Paquet *et al.* 2011; Baskett & Waples 2013). We recently provided the first empirical evidence that managed gene flow successfully reduces genome-wide differentiation (Waters *et al.* 2015; Waters *et al.* 2017) and divergence at trait-associated loci (Waters *et al.* 2018) when compared to a segregated management approach over four generations. This system also permits direct examination of the extent to which managed gene flow may reduce inbreeding and its potential adverse effects on fitness compared to segregated management over multiple generations.

Here, we estimated pairwise relatedness, a measure that may impact inbreeding in subsequent generations, and individual inbreeding coefficients across four generations of the integrated and segregated Chinook salmon hatchery lines at CESRF using 6805 restriction site-associated (RAD) loci. Multigenerational observations are advantageous because processes occurring in the wild, including natural selection, could mitigate or exacerbate the effects of inbreeding over time. Further, assessing multiple generations permits the quantification of the accumulated effects of inbreeding since the initiation of the hatchery. We then determined if managed gene flow successfully reduced the incidence of inbreeding due to captive-rearing by comparing levels of inbreeding and relatedness in the two hatchery lines over time. Next, the relationships between inbreeding coefficient and eight fitness-related traits – date of return to freshwater spawning grounds (return timing), fork length, weight, and condition factor at return, fecundity, reproductive effort, daily growth coefficient, and spawn timing – were quantified using linear regression models. The results will provide a more comprehensive understanding of

genetic and phenotypic change that may occur in captive-reared populations and will further inform “best” practices to minimize risks in conservation-focused captive breeding programs.

3.3 METHODS

Population description

The initial hatchery population at CESRF was founded from 1997–2002 using returning wild adults from the anadromous upper Yakima River Spring Chinook salmon population. Wild adults were collected at the Roza Dam Adult Monitoring Facility (RAMF), located 90 river kilometers south of CESRF, as they returned from the ocean to their freshwater spawning grounds. Adults were collected randomly and proportionately (based on average migration timing patterns) over the duration of the run and held at CESRF until maturation (Knudsen *et al.* 2006), at which point they were spawned following a 3x3 factorial mating design (when possible) to increase the effective number of breeders compared to a single pair design (Busack & Knudsen 2007). Offspring were reared at CESRF for approximately 16 months before being transferred to three acclimation sites; these sites were designed to expand the spatial influence of supplementation efforts while also enabling related research. The fish were acclimated at these sites for two months, after which they were allowed to volitionally begin their seaward migration.

In 2002, the initial hatchery population was divided into the integrated (INT) and segregated (SEG) hatchery lines by spawning wild and first generation hatchery-origin adults, respectively. Broodstock for the integrated line comprises only fish born in the wild, and all returning adults from this line are allowed to spawn in the river. The segregated line, in contrast, uses only returning hatchery-origin fish as broodstock, and no adults from this line are allowed to

reproduce naturally; fish from the two lines are differentially marked for external identification, so that all SEG adults are removed from the system at RAMF. With these practices, fish from the integrated line are exposed to hatchery conditions for just one generation while the segregated line is exposed every generation. Broodstock collection, spawning, and rearing procedures for subsequent generations of both lines are conducted in the same manner as the founding generation. Notably, the numbers of adults spawned in the segregated line are typically a quarter of those spawned in the integrated line (Waters *et al.* 2015), an outcome of the fact that the Cle Elum facility serves broader restoration goals. Additional details regarding the background of CESRF and the initiation of the integrated and segregated hatchery lines have been described elsewhere (Knudsen *et al.* 2006; Fast *et al.* 2015; Waters *et al.* 2015; Waters *et al.* 2018).

Sample collection

Tissue samples for DNA were collected from all fish during spawning at CESRF and stored in 100% ethanol. Adults from the following years were sub-sampled for this study: the 1998 wild founders (second founding year; P_1 Founders) and hatchery brood years 2002 (F_1 Wild and F_1 Hatchery), 2006 (F_2 INT and F_2 SEG), 2010 (F_3 INT and F_3 SEG), and 2014 (F_4 INT and F_4 SEG). A majority (>75%; Knudsen *et al.* 2006) of Chinook salmon from this system spend two years in the ocean and return at age four to reproduce, so these samples represent five generations.

Phenotypic data from returning adults were collected annually upon arrival at RAMF and, for fish used as broodstock, during spawning at CESRF. Eight fitness-related traits were examined for changes in fitness with inbreeding (Table 3.1). Return timing corresponds to day of arrival at RAMF. Spawn, or maturation, timing of all broodstock was estimated weekly at

CESRF by manually checking for gonadal ripeness. Gravimetric estimates of fecundity and reproductive effort (proportion of body mass allocated to gamete production) were calculated for females following Knudsen *et al*. (2008). Fecundity estimates were reduced by 5.5% to correct for bias from residual ovarian fluid remaining within the egg mass (Knudsen *et al.* 2008). Fork length and weight were measured at both RAMF and upon spawning at CESRF. However, only measurements from RAMF were analyzed here to minimize possible influences from time spent in the hatchery after collection at RAMF and from the development of secondary sex traits. Fulton's condition factor *K* (Ricker 1975), a measure of quality or well-being, was calculated using fork length and weight measurements at RAMF:

$$K = 100^* \frac{W}{L^3} \qquad (1)$$

where W was weight at RAMF in grams and L was length in centimeters. Daily growth coefficient (DGC), which has been shown to be more independent of initial weight than other measures, was calculated according to Cho (1992) and Dupont-Nivet et al. (2010):

$$DGC = 100 \times [((\text{final weight})^{1/3} - (\text{initial weight})^{1/3})/\text{days}] \qquad (2)$$

where initial weight was the weight at RAMF, final weight was weight at spawning, and days was the number of days between arrival at RAMF and spawning at CESRF. Here, DGC provided a measure related to weight lost during maturation and holding in the tanks at CESRF.

DNA sequencing and genotyping

DNA from all tissue samples was previously sequenced to investigate population divergence since founding (Waters *et al.* 2015; Waters *et al.* 2017) and the effects of domestication selection on genetic diversity underlying fitness-related traits (Waters *et al.* 2018).

Therefore, DNA extraction, RAD library preparation, and bioinformatic processing followed the methods reported in Waters *et al*. (2018) with two exceptions. First, loci were removed if they did not have a minor allele frequency ≥0.05 in the P_1 Founders, rather than ≥0.05 in at least one population, because unbiased estimates of inbreeding rely on polymorphic loci with allele frequencies estimated from a wild base population (Kardos *et al.* 2015). Secondly, all loci potentially under selection (i.e. outlier loci) identified from these samples in a previous study (Waters *et al.* 2017) were removed, because non-neutral loci may also bias inbreeding estimates.

Relatedness and inbreeding

In addition to inbreeding, pairwise relatedness was quantified as it may impact future levels of inbreeding within the two hatchery lines. Two sets of loci were used to estimate relatedness and inbreeding. The first set comprised all loci that passed filtering criteria. The second set of loci contained those that remained after pruning for linkage disequilibrium (LD) in PLINK (Purcell *et al.* 2007), because correlations between loci may influence estimates (Blouin 2003; Wang 2007; Santure *et al.* 2010; Kardos *et al.* 2015). PLINK requires base pair-specific information for each locus in order to prune marker sets with LD. Therefore, all study loci were aligned to the Chinook salmon genome (*Oncorhynchus tshawytscha;* Accession PRJNA432585 released Jan. 16, 2018 on NCBI by Fisheries and Oceans Canada) using *Bowtie2* (v. 2.2.9, Langmead & Salzberg 2012) with default parameters. Loci that aligned to the Chinook genome with a mapping quality ≥10 were assigned positions and imported into PLINK. LD pruning was then conducted using conservative parameters (window size = 50, step size = 5, variance inflation factor = 1).

Genomic measures of pairwise relatedness, R_{xy}, and individual inbreeding coefficient, *F*, within each generation of each hatchery line were then estimated in PLINK using the *make-rel* and *het* functions, respectively, using allele frequencies from the P_1 Founders, the wild base population which we assumed comprised largely unrelated individuals. The *het* measure of *F*, which compares the observed number of homozygous genotypes to the expected mean number under random mating, was chosen over an alternative measure, runs of homozygosity (ROH; Broman & Weber 1999; Curik *et al.* 2014), for two reasons. First, the marker density (~1 mapped SNP per 0.5 Mb) was likely too low to accurately detect runs of homozygosity; typical studies that use ROH as a measure regularly employ >50,000 loci (e.g. Ferenčaković *et al.* 2013; Bosse *et al.* 2015; Zhang *et al.* 2015). Second, simulations suggest that *F* based on homozygosity has equal or higher precision than other measures, including ROH, when 5,000 – 10,000 markers are used, as well as low bias when allele frequencies are derived from a wild base population (Kardos *et al.* 2015). Therefore, we considered this measure of *F* to be the most appropriate, given our data.

We hypothesized that hatchery rearing may inadvertently cause an increase in R_{xy} and *F* over time due to factors such as space limitations, broodstock collection practices, and spawning procedures. Further, levels of R_{xy} and *F* were expected to be higher in the segregated line than in the integrated line, since each generation of the segregated line has 3- to 27-fold fewer effective numbers of breeders (Waters *et al.* 2015; Waters *et al.* 2017). Differences in the distributions of R_{xy} and *F* between successive generations within each hatchery line, as well as differences between the two lines within each generation, were tested using Wilcoxon rank-sum tests conducted in R (R Core Team 2017), as the data were non-normally distributed.

Effects of inbreeding on traits influencing fitness

The effects of inbreeding coefficient on eight fitness-related traits (Table 3.1) were quantified using mixed linear models in R (R Core Team 2017). Only four-year-old fish were analyzed because they represented 90% of all adults sampled for this study, and sample sizes for three- and five-year-old fish were too small to accurately estimate age effects. Samples from the P_1 Founders were also excluded, as fish from this group were not present in any of the subsequent generations; the effect of the P_1 Founders could therefore not be separated from the generational effect. Thus, each generation analyzed comprised only fish from the integrated and segregated hatchery lines (i.e. a paired design).

Each phenotypic trait was modeled as a function of inbreeding coefficient, hatchery line, generation, and sex. We considered both linear and non-linear effects of inbreeding on fitness, the latter by including a quadratic term (i.e. F^2). We also included first order interactions of covariates with F to determine if the effects of inbreeding varied by hatchery line, sex, and generation. We considered including length and weight as covariates for specific traits, namely return timing, spawn timing, fecundity, and reproductive effort. No significant relationships were observed for return and spawn timing, while length and weight were highly correlated with fecundity and reproductive effort; we therefore did not include length and weight as explanatory variables in the models. All traits were modeled with a Gaussian distribution and an identity link function. The general model form was:

$$y_i = u + F_i + F^2_i + \text{Line}_{ij} + \text{Gen}_{ik} + \text{Sex}_{il} + F_i*\text{Line}_{ij} + F_i*\text{Gen}_{ik} + F_i*\text{Sex}_{il} + \text{Line}_{ij}*\text{Gen}_{ik} + \text{Line}_{ij}*\text{Sex}_{il} + \text{Gen}_{ik}*\text{Sex}_{il} + e_i \quad (3)$$

where y_i was the trait measurement for individual i, F_i was the fixed effect of inbreeding coefficient for individual i (F^2_i its square), Line_{ij} was the fixed effect of hatchery line j for

individual *i*, Gen_{ik} was the fixed effect of generation *k* for individual *i*, and Sex_{il} was the fixed effect of sex *l* for individual *i*, respectively. Line, sex, and generation were treated as categorical variables. Models for gravimetric fecundity and reproductive effort did not include sex, as these traits were only measured in females. There was no multicollinearity between predictors (variance inflation factors <2).

Backward model selection was performed on the full model using AIC scores to identify the best model. The selection process was implemented with the *stepAIC* function of the *MASS* package (Venables & Ripley 2002) in R (R Core Team 2017). The main effect of inbreeding coefficient was retained in the best model for each trait, even if non-significant, to quantify its effect on fitness. A significant, negative effect of *F* on a phenotypic trait was considered as evidence of inbreeding depression.

3.4 RESULTS

Sample collection

Tissues of 753 adult Chinook salmon were sub-sampled for DNA from the P_1 Founders and four generations of the integrated and segregated hatchery lines (Table S3.2.1). Phenotypic data were also recorded for these individuals upon their arrival at RAMF and again upon spawning at CESRF.

DNA sequencing and genotyping

Good quality DNA and RAD sequences were obtained for 696 of the 753 tissue samples, and bioinformatic processing using *Stacks* identified 11,809 biallelic loci. Subsequent filtering of loci and individuals following the steps of Waters *et al*. (2018), with the exception of the minor

allele frequency filter, reduced the data set to 452 individuals and 7,002 loci. Then, any of the 276 outlier loci that had been previously identified from these samples using two independent methods (Waters *et al.* 2017) were excluded. The final data set comprised 452 individuals genotyped at 6,805 loci (Tables S3.2.2 and S3.2.3).

Relatedness and inbreeding

All 6,805 loci that remained after filtering served as the first data set from which estimates of R_{xy} and F were obtained (i.e. full data set). All loci were then aligned to the Chinook salmon genome to identify base pair positions for LD pruning in PLINK (Table S3.2.4a). The 5,600 loci that aligned to the Chinook genome with mapping quality ≥10 were assigned positions, and LD pruning in PLINK identified 1,100 putatively unlinked loci for the second data set (i.e. LD-pruned data set; Table S3.2.4b). Genomic estimates of R_{xy} and F obtained from these two data sets were comparable (Pearson's r = 0.76 and 0.98 for R_{xy} and F, respectively; Figures 3.1 and 3.2, Table S3.2.5). However, we focused on estimates obtained using all 6,805 loci, as previous simulations indicated that additional markers conferred higher precision (Figure S3.1.1).

a. Relatedness

Pairwise relatedness in the P_1 Founders was centered near zero but significantly increased in the F_1 Wild and F_1 Hatchery groups following initiation of the hatchery (Figure 3.1, Table 3.2, Table S3.2.5a). Comparisons between subsequent generations revealed a temporally consistent, significant increase in R_{xy} for the segregated line, while R_{xy} in the integrated line only increased in the F_4 generation (Table 3.2). For example, the proportion of pairwise comparisons with R_{xy}

values ≥ 0.1 consistently increased in the F_2 (0.07), F_3 (0.14), and F_4 (0.19) generations of the segregated line but only increased in the last generation of the integrated line (proportions of 0.006, 0.005, and 0.025 for the F_2, F_3, and F_4 generations, respectively; Figure 3.1, Table S3.2.5a). Significant differences between the distributions of R_{xy} for the integrated and segregated hatchery lines were observed in the F_1, F_3, and F_4 generations. Specifically, R_{xy} of the F_1 Hatchery population was slightly lower than the F_1 Wild population, while R_{xy} for the segregated line was higher than the integrated line in the F_3 and F_4 generations (Figure 3.1, Table 3.2). Despite the differences, median values of R_{xy} remained close to zero in both hatchery lines across all generations.

b. Inbreeding

The distribution of inbreeding coefficients in the P_1 Founders was centered near zero, although individuals with relatively large *F* values (positive and negative) were also observed (Figure 3.2, Table S3.2.5b). Levels of *F* then significantly increased in the F_1 Wild and F_1 Hatchery groups following initiation of the hatchery (Figure 3.2, Table 3.2). Distributions of inbreeding values did not significantly change in the F_2 and F_3 generations of the segregated line, while a marginally significant increase in the distribution of *F* was observed in the F_3 generation of the integrated line (Table 3.2). Distributions of *F* in both hatchery lines then significantly decreased in the F_4 generation. Levels of *F* were not significantly different between the hatchery lines in the F_1, F_3, and F_4 generations (Figure 3.2, Table 3.2). The distribution of *F* values, however, was significantly higher in the segregated line for the F_2 generation.

Effects of inbreeding on traits influencing fitness

The best model to quantify the effect of individual inbreeding coefficient on fitness varied for each phenotypic trait (Table 3.3). Models for fecundity and reproductive effort comprised only main effects while those for the other six traits included both main effects and interactions. Return timing was the only trait for which a non-linear effect of *F* was included in the best model.

The effects of *F* on fitness varied by trait, hatchery line, generation, and sex. For example, inbreeding coefficient did not significantly affect the fecundity and reproductive effort of females, and the return timing and fork length of both sexes, in any generation of either hatchery line (Figure 3.3, Table 3.3, Figures S3.1.2 and S3.1.3). Spawn timing and weight of both males and females in the integrated line were significantly affected by *F* across all generations (Figure 3.4, Table 3.3, Figure S3.1.4). Specifically, inbred fish in the integrated line tended to mature later and weigh more than non-inbred fish. In contrast, spawn timing and weight of fish in the segregated line were not significantly affected by *F* in any generation. *F* had a significant positive effect on the condition factor of males and females in both hatchery lines in the F_1 generation, but no such effects were observed in the subsequent generations (Figure 3.5, Table 3.3). Daily growth coefficient of males from both hatchery lines was negatively affected by inbreeding coefficient across all generations (Table 3.3, Figure S3.1.5). That is, inbred males had a higher rate of weight loss during holding in the hatchery compared to non-inbred males. No such effect was observed in females.

3.5 DISCUSSION

Effective management of both captive and wild populations requires a comprehensive understanding of demographic, ecological, and genetic risks. Inbreeding is one genetic risk that

may affect individual fitness, population productivity, and probability of extinction (Keller & Waller 2002; Frankham 2005; O'Grady *et al.* 2006; Kardos *et al.* 2016). Yet, the magnitude of such effects remains unclear, largely due to limitations in obtaining precise estimates of individual inbreeding coefficients. Here, we assessed multigenerational levels of pairwise relatedness, which may impact inbreeding in subsequent generations, and individual inbreeding in integrated and segregated hatchery lines of Chinook salmon using a genomic approach that has been shown to provide accurate and precise estimates. The effects of inbreeding on eight fitness-related traits were also quantified. The two hatchery lines, which were derived from the same wild population but are now maintained separately, represent demographic extremes that provide insight into a range of possible outcomes relevant to risk assessment of captive breeding programs.

Relatedness and inbreeding over time

The distributions of relatedness and inbreeding coefficients in the P_1 Founders were centered near zero (Figures 3.1 and 3.2), as would be expected in a wild, randomly mating population of moderate size. This result supports the decision to use allele frequencies from the P_1 Founders to minimize potential bias in estimates of relatedness and inbreeding. However, a small number of individuals with relatively large (positive and negative) inbreeding coefficients were also observed (Figure 3.2), indicating a low rate of natural inbreeding and outbreeding.

The small but significant increases in levels of relatedness and inbreeding detected in the F_1 Wild and F_1 Hatchery groups may be due to the initiation of the hatchery program, as these values are largely influenced by spawning of the parental generation. Nearly 800 wild adult salmon returned to the system in 1998 (Waters *et al.* 2015), the second founding year from which

a majority of fish in the F_1 generation was derived. Approximately half of these adults were collected as hatchery broodstock (i.e. parents of F_1 Hatchery group) while the other half were allowed to spawn naturally in the river (i.e. parents of F_1 Wild group). The division of the population into two groups, and the resultant decrease in population sizes, may have inadvertently increased the rate of inbreeding during spawning. The decreases in population sizes may also explain the observed increases in relatedness in the subsequent generation, as the returning F_1 adults in each line represented offspring from a smaller number of parents, and the probability of sampling related individuals for this study may therefore have increased. However, as the study lacks a control population to which the hatchery lines can be compared, the effects of hatchery initiation on relatedness and inbreeding cannot be explicitly quantified.

A significant, temporally consistent increase in relatedness was observed in the F_2 to F_4 generations of the segregated line, while the integrated line only exhibited an increase in relatedness for the F_4 generation. In addition, levels of relatedness in the segregated line were significantly higher than those in the integrated line for the F_3 and F_4 generations (Figure 3.1, Table 3.2). Large differences in estimates of effective numbers of breeders, N_b, between the two hatchery lines (values for the INT line were 3- to 27-fold higher, Waters *et al.* 2015; Waters *et al.* 2017) likely contributed to the changes in relatedness within and between the two lines. Yet, the detection of significant differences may also be an artifact of the large number of pairwise comparisons within each sampling group (Figure 3.1, Table S3.2.5a) and the associated power to detect even the slightest differences, as suggested by the fact that differences were detected even though the median values of R_{xy} remained close to zero in both hatchery lines across all generations (Figure 3.1).

Distributions of inbreeding values did not display temporally consistent trends in either hatchery line for the F_2 to F_4 generations (Figure 3.2). Further, levels of *F* were different between the two lines only in the F_2 generation, where the segregated line had higher levels of inbreeding. Such results were surprising given the large differences in estimates of effective numbers of breeders between the hatchery lines. However, similarities between the lines may be due to the "best-practice" management principles employed at the Cle Elum Supplementation and Research Facility, which aim to reduce negative ecological and genetic effects of hatchery rearing (Fast *et al.* 2015) and include the use of 3x3 factorial matings to increase the effective number of breeders (Busack & Knudsen 2007). Together, the results of relatedness and inbreeding suggest that managed gene flow may reduce these genetic risks in the short term. However, our findings also suggest that the risks may not be severe in small, segregated hatchery populations, at least over a few generations.

Reduced levels of inbreeding in the F_4 generation

The distributions of individual *F* decreased in the F_4 generation for both hatchery lines, a result that was unexpected, particularly for the smaller segregated line. The inclusion of paralogous loci or contamination between samples, both of which would cause an increase in individual heterozygosity and a concomitant decrease in *F*, are unlikely explanations for the observed decrease. First, all sequence reads were aligned to a reference database of non-duplicated RAD loci for Chinook salmon (Brieuc *et al.* 2014), which minimized or eliminated the inclusion of paralogous loci. Second, individuals from the integrated and segregated hatchery lines were sampled, processed, and sequenced together. Contamination would therefore produce individuals with anomalous genotypes or those that are intermediate between the two hatchery

lines. However, a discriminant analysis of principal components (DAPC) showed that individuals from the F_4 generation clustered with previous generations from their respective hatchery lines (Figure S3.1.6; Waters *et al.* 2017) and did not display any inconsistent patterns. Further, the F_4 generation did not differ from previous generations with respect to other indicators of possible contamination, including individual observed heterozygosity (Figure S3.1.7), the distribution of F_{IS} per locus, and allelic read depths at heterozygous genotypes.

The decrease of F in the F_4 generation was therefore likely due to a sampling or an environmental effect, especially since it was observed in both hatchery lines. Due to external constraints, sample sizes in the F_4 generation were approximately half of those from the previous generations and may have reduced power to accurately estimate the distributions of individual F. Another explanation is that inbred fish in the F_4 generation experienced an unusually high degree of mortality in the marine environment, perhaps due to the anomalous warm-water mass (i.e. "the Blob") that first appeared in the northeastern Pacific in 2013 (the last year at sea for fish from the F_4 generation) and had negative ecological impacts throughout the region (Cavole *et al.* 2016). A final possible explanation for the observed decrease in F in the F_4 generation is a possible change in broodstock collection or mating procedures that occurred in the hatchery during the F_3 (the parental) generation. For example, a higher proportion of crosses between unrelated individuals would produce less inbred offspring. However, hatchery records do not show any deviations from protocols in that particular generation.

Effects of inbreeding on traits influencing fitness

Inbreeding did not affect fecundity, reproductive effort, return timing, and fork length of the adult Chinook salmon, while significant effects of inbreeding on phenotypes were observed

for spawn timing, weight, daily growth coefficient, and condition factor. The effects of inbreeding, however, varied by hatchery line, sex, and generation.

Inbred males and females in the integrated hatchery line tended to spawn later and weigh more than non-inbred individuals across all generations. Inbred males and females of both hatchery lines had higher condition factors (i.e. were "fatter") than non-inbred fish in the F_1 generation, while inbred males of both hatchery lines had larger daily growth coefficients (i.e. lost more weight between arrival at Roza Dam and spawning at CESRF). The biological impacts of these inbreeding effects, however, have yet to be quantified. Later spawn timing may confer higher fitness, as these individuals may not have to expend as much energy to defend spawning territories as earlier-maturing fish. However, later spawn timing may also negatively affect fitness, because later-maturing individuals may face higher levels of competition for mates. Similarly, higher weights and condition factors may confer higher fitness due to increased competitive abilities (Schroder *et al.* 2008; Schroder *et al.* 2010) and higher fecundity. Alternatively, higher weights and condition factors may impair the ability of fish to efficiently swim through the water and catch prey. The larger (more negative) daily growth coefficients of inbred fish indicate that they expend more energy before spawning; this may negatively affect gamete production and quality.

The varied effects of inbreeding across generations indicate that environmental conditions may exacerbate or mitigate the fitness consequences of inbreeding. The different effects of inbreeding between hatchery lines may result from an interaction between inbreeding and other genetic differences (i.e. genome-wide divergence and domestication selection) between the integrated and segregated hatchery lines that have been previously identified (Waters *et al.* 2015; Waters *et al.* 2017). Additionally, differences in phenotypic variability and variation in

levels of inbreeding may affect statistical power to detect correlations between inbreeding and fitness in the two hatchery lines. Further exploration is therefore needed to assess the consequences of inbreeding to the fitness of captive-reared fish.

Measuring relatedness and inbreeding

Genomic-based estimates of inbreeding have been shown to be accurate and precise in many circumstances (Kardos *et al.* 2015). The reduction in error variance associated with genomic-based estimates of inbreeding also improves the ability to detect effects of inbreeding on fitness (Hoffman *et al.* 2014). Here, we used 6,805 markers and allele frequencies from a wild base population to estimate relatedness and inbreeding and then quantified the effects of inbreeding on eight fitness-related traits. While possible bias in the genetic and fitness effects of inbreeding may have been minimized, their magnitude may have been influenced by the sampling design of this study. Specifically, adult Chinook salmon returning from the ocean to their freshwater spawning grounds were sampled. These adults are the small fraction of fish that survived the selective pressures imposed by the freshwater and marine environments and may represent the "most fit" individuals from the population. Estimates of relatedness and inbreeding, along with fitness effects, may therefore be downwardly biased if inbred fish experienced a disproportionately high level of mortality prior to adulthood. This possibility is particularly relevant when the results of Thrower and Hard (2009) are considered, which showed that inbred rainbow trout and steelhead experienced substantially lower (~80%) survival than non-inbred fish when released into the wild, marine environment (survivals were not significantly different during captivity in freshwater). Likewise, the sampling of adults may explain why levels of relatedness and inbreeding were comparable between the integrated and segregated hatchery

lines despite large differences in estimates of effective numbers of breeders (Waters *et al.* 2015; Waters *et al.* 2017). The influence of adult sampling on estimates could be explicitly tested in a future study by sampling both juveniles and returning adults from each hatchery line and comparing the distributions of relatedness and inbreeding between the life stages.

Management implications

Wild and hatchery-reared salmon provide millions of dollars to local economies and are important to the ecology and culture of the Pacific Northwest. However, numerous wild salmon populations in California, Oregon, Idaho, and Washington are declining (Gustafson *et al.* 2007), and many are listed under the Endangered Species Act. Supportive breeding programs in the form of hatcheries exist to supplement these declining populations, and much effort has been aimed at understanding the long term impacts of these programs. However, research has largely focused on the impacts of domestication selection on the fitness of hatchery fish. In contrast, genetic population sizes in hatcheries are typically smaller than census sizes and are likely unrecognized causes of reduced fitness. It is important to fully understand all potential risks in order to develop comprehensive benefit-risk assessments for supplementation programs and to maximize their effectiveness to rebuild natural populations.

This study is one of the first to assess the multigenerational risks of inbreeding in Pacific salmon using genomic approaches, which permit accurate and precise estimation of inbreeding and its effects on fitness. We also provide the first empirical evaluation of the effectiveness of integrated management, or managed gene flow, to minimize possible risks of inbreeding over time. While the results suggest that managed gene flow may reduce levels of relatedness and the rate of inbreeding, they also indicate that these risks may not be severe in small, segregated

hatchery populations. These findings provide a more comprehensive understanding of genetic and phenotypic change that may occur in captive-reared populations and will further inform "best" practices to minimize risks in conservation-focused captive breeding programs.

3.6 ACKNOWLEDGEMENTS

We thank the following individuals for project development, broodstock collection and sampling, and laboratory and analytical assistance: all Yakama Nation and Washington Department of Fish and Wildlife personnel at the Roza Dam Adult Monitoring Facility and the Cle Elum Supplementation and Research Facility, James Thorson, Michael Ford, Isadora Jimenez-Hidalgo, Lorenz Hauser, Daniel Drinan, Eleni Petrou, Natalie Lowell, Molly Jackson, Mary Fisher, Samuel May, Carita Pascal, and Garrett McKinney. We also thank everyone who was involved in establishing CESRF and shaping its research direction, including Levi George, Melvin Sampson, Steve Schroder, Craig Busack, past and present members of the Independent Scientific Review Panel, and the Yakama Nation Tribal Council. Funding for this study was provided by Washington Sea Grant (to K.A.N.), the National Marine Fisheries Service–Sea Grant Fellowship in Population and Ecosystem Dynamics (to C.D.W.), and the Hall Conservation Genetics Research Award from the University of Washington (to C.D.W.).

3.7 TABLES

Table 3.1 – Phenotypic traits measured in adult Chinook salmon at the Roza Dam Adult Monitoring Facility (RAMF) and, if used as broodstock, at the Cle Elum Supplementation and Research Facility (CESRF). Measurements from RAMF were used for fork length, weight, and condition factor in the analyses.

Trait Category	Trait (units)	Locations Measured	
Life history	return timing (day of year)	RAMF	
	spawn timing (day of year)		CESRF
	fecundity (females only; number of eggs)		CESRF
	reproductive effort (females only; no units)		CESRF
	daily growth coefficient (no units)	RAMF to CESRF	
Morphometric	fork length (cm)	RAMF	CESRF
	weight (kg)	RAMF	CESRF
	condition factor (no units)	RAMF	CESRF

Table 3.2 – Test statistics, *p* values, and the direction of significant differences from Wilcoxon rank-sum tests conducted between subsequent generations (gens) within the integrated (INT) and segregated (SEG) hatchery lines. Tests were also conducted between the hatchery lines within each generation. Significant differences are in bold.

Comparison		R_{xy}			F		
		W	***p* value**	**Difference**	**W**	***p* value**	**Difference**
between gens	P_1 vs. F_1 Wild	**142910**	**<2e-16**	**$F_1 > P_1$**	**1067**	**0.002**	**$F_1 > P_1$**
	F_1 Wild vs. F_2 INT	1329300	0.702	NA	1824	0.34	NA
	F_2 INT vs. F_3 INT	1871900	0.995	NA	**1564**	**0.049**	**$F_3 > F_2$**
	F_3 INT vs. F_4 INT	282120	**4.67e-4**	**$F_4 > F_3$**	**1513**	**1.86e-9**	**$F_3 > F_4$**
	P_1 vs. F_1 Hatch	**161720**	**<2e-16**	**$F_1 > P_1$**	**889.5**	**0.001**	**$F_1 > P_1$**
	F_1 Hatch vs. F_2 SEG	**825390**	**1.38e-5**	**$F_2 > F_1$**	1257	0.440	NA
	F_2 SEG vs. F_3 SEG	**1014800**	**2.72e-11**	**$F_3 > F_2$**	1746	0.289	NA
	F_3 SEG vs. F_4 SEG	213460	**3.23e-6**	**$F_4 > F_3$**	**1292**	**5.96e-8**	**$F_3 > F_4$**
between lines	F_1 Wild vs. F_1 Hatch	**1214000**	**4.20e-7**	**INT > SEG**	1392	0.489	NA
	F_2 INT vs. F_2 SEG	1089900	0.676	NA	**1111**	**0.017**	**SEG > INT**
	F_3 INT vs. F_3 SEG	**1678100**	**<2e-16**	**SEG > INT**	2126	0.667	NA
	F_4 INT vs. F_4 SEG	**29902**	**8.30e-9**	**SEG > INT**	333	0.519	NA

Table 3.3 – Summary of the best fit linear models for eight fitness-related traits based on AIC. The coefficient estimate, standard error, *t* value, and *p* value are provided for each term included in the best model, in addition to sample sizes for each trait. The reference level is females from the integrated hatchery line in the F_1 generation. Terms indicating an overall significant effect of *F* on trait value are in bold. Note that a significant interaction term indicates a significant difference from the reference level but may not represent a significant overall effect of *F* on trait value.

Trait	Term	Estimate	Std. Error	*t* value	*p* value
Fecundity	Intercept	4062.5	139.1	29.214	<2e-16
	F	-112.5	335.0	-0.336	0.737
n = 204	Hatchery Line	-186.6	119.5	-1.561	0.120
	F_2 gen.	-701.5	160.5	-4.371	2e-5
	F_3 gen.	149.3	154.6	0.966	0.335
	F_4 gen.	-254.6	217.7	-1.169	0.244
Reproductive effort	Intercept	0.198	0.004	52.746	<2e-16
	F	0.009	0.010	0.964	0.336
n = 201	F_2 gen.	-0.004	0.005	-0.790	0.431
	F_3 gen.	0.011	0.004	2.378	0.018
	F_4 gen.	0.013	0.006	2.085	0.038
Return timing	Intercept	165.267	2.777	59.519	<2e-16
	F	8.641	9.724	0.889	0.375
n = 336	F^2	-29.746	19.246	-1.546	0.123
	Sex	0.399	2.909	0.137	0.891
	Hatchery Line	5.998	2.693	2.227	0.027
	F_2 gen.	6.523	2.859	2.281	0.023
	F_3 gen.	-4.107	2.759	-1.488	0.138
	F_4 gen.	-4.271	3.767	-1.134	0.258
	Hatchery Line : Sex	-9.941	4.426	-2.246	0.025
Spawn timing	Intercept	264.729	1.213	218.223	<2e-16
	F	**7.011**	**3.247**	**2.159**	**0.032**

n = 334	Sex	2.548	1.431	1.781	0.076
	Hatchery Line	-3.284	1.534	-2.141	0.033
	F_2 gen.	-0.223	1.537	-0.145	0.885
	F_3 gen.	-0.097	1.521	-0.063	0.949
	F_4 gen.	-5.843	2.158	-2.708	0.007
	F : Hatchery Line	-7.435	4.133	-1.799	0.073
	F_2 gen : Hatchery Line	-4.008	1.897	-2.113	0.035
	F_3 gen : Hatchery Line	-4.908	1.856	-2.644	0.009
	F_4 gen : Hatchery Line	1.903	2.547	0.747	0.455
	F_2 gen : Sex	0.156	1.956	0.080	0.937
	F_3 gen : Sex	-4.311	1.911	-2.256	0.025
	F_4 gen : Sex	-2.103	2.492	-0.844	0.399
Fork length	Intercept	71.393	0.758	94.251	<2e-16
	F	3.976	2.245	1.771	0.078
n = 335	Sex	0.898	0.502	1.789	0.075
	Hatchery Line	-0.172	1.058	-0.162	0.871
	F_2 gen.	-2.464	0.901	-2.734	0.007
	F_3 gen.	1.026	0.861	1.191	0.234
	F_4 gen.	3.663	1.174	3.120	0.002
	F : Hatchery Line	-4.783	2.853	-1.676	0.095
	F_2 gen : Hatchery Line	-1.037	1.305	-0.794	0.428
	F_3 gen : Hatchery Line	0.202	1.264	0.160	0.873
	F_4 gen : Hatchery Line	-6.273	1.720	-3.647	3.09e-4
Weight	Intercept	4.221	0.140	30.173	<2e-16
	F	**0.861**	**0.415**	**2.077**	**0.039**
n = 336	Sex	0.183	0.093	1.978	0.049
	Hatchery Line	-0.026	0.195	-0.132	0.895
	F_2 gen.	-0.599	0.166	-3.596	3.73e-4
	F_3 gen.	0.171	0.159	1.072	0.284
	F_4 gen.	1.545	0.217	7.127	6.66e-12
	F : Hatchery Line	-0.901	0.527	-1.711	0.088
	F_2 gen : Hatchery Line	-0.278	0.241	-1.154	0.249

	F_3 gen : Hatchery Line	-0.028	0.233	-0.118	0.906
	F_4 gen : Hatchery Line	-1.482	0.318	-4.663	4.55e-6
Condition factor	Intercept	1.144	0.012	97.403	<2e-16
	F	**0.111**	**0.045**	**2.435**	**0.015**
n = 335	Hatchery Line	-0.018	0.008	-2.173	0.030
	F_2 gen.	-0.044	0.015	-2.942	0.003
	F_3 gen.	0.018	0.016	1.138	0.256
	F_4 gen.	0.193	0.016	11.890	<2e-16
	F : F_2 gen.	-0.116	0.058	-2.018	0.044
	F : F_3 gen.	-0.151	0.068	-2.225	0.027
	F : F_4 gen.	-0.213	0.117	-1.813	0.071
Daily growth coefficient	Intercept	-0.068	0.004	-16.489	<2e-16
	F	0.002	0.010	0.255	0.799
n = 328	Sex	-0.019	0.004	-5.122	5.27e-7
	Hatchery Line	-0.006	0.005	-1.043	0.298
	F_2 gen.	0.003	0.005	0.650	0.516
	F_3 gen.	0.017	0.005	3.409	0.001
	F_4 gen.	-0.069	0.007	-10.429	<2e-16
	***F* : Sex**	**-0.032**	**0.016**	**-2.037**	**0.042**
	F_2 gen : Hatchery Line	0.014	0.008	1.921	0.056
	F_3 gen : Hatchery Line	0.021	0.007	2.945	0.003
	F_4 gen : Hatchery Line	0.020	0.009	2.113	0.035

3.8 FIGURES

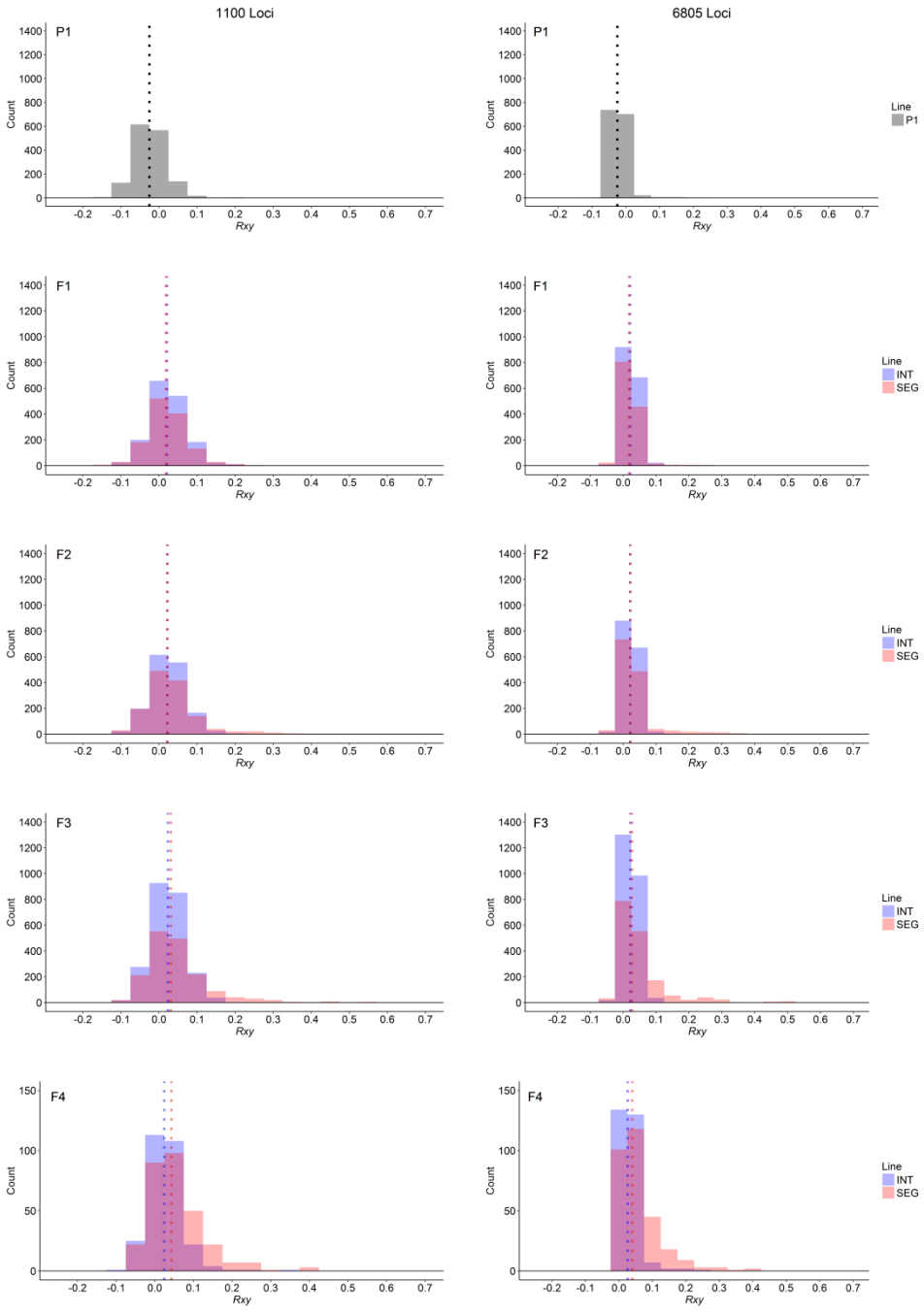

Figure 3.1 – Histograms of pairwise relatedness, R_{xy}, for all pairs of individuals within the P_1 founders (gray) and the F_1, F_2, F_3, and F_4 generations of the integrated (INT, in blue) and

segregated (SEG, in red) hatchery lines estimated from 1,100 unlinked loci (left column) and all 6,805 study loci (right column). The median values for each distribution are denoted by dotted vertical lines. Note that the y-axis differs for the F_4 generation due to fewer numbers of pairwise comparisons.

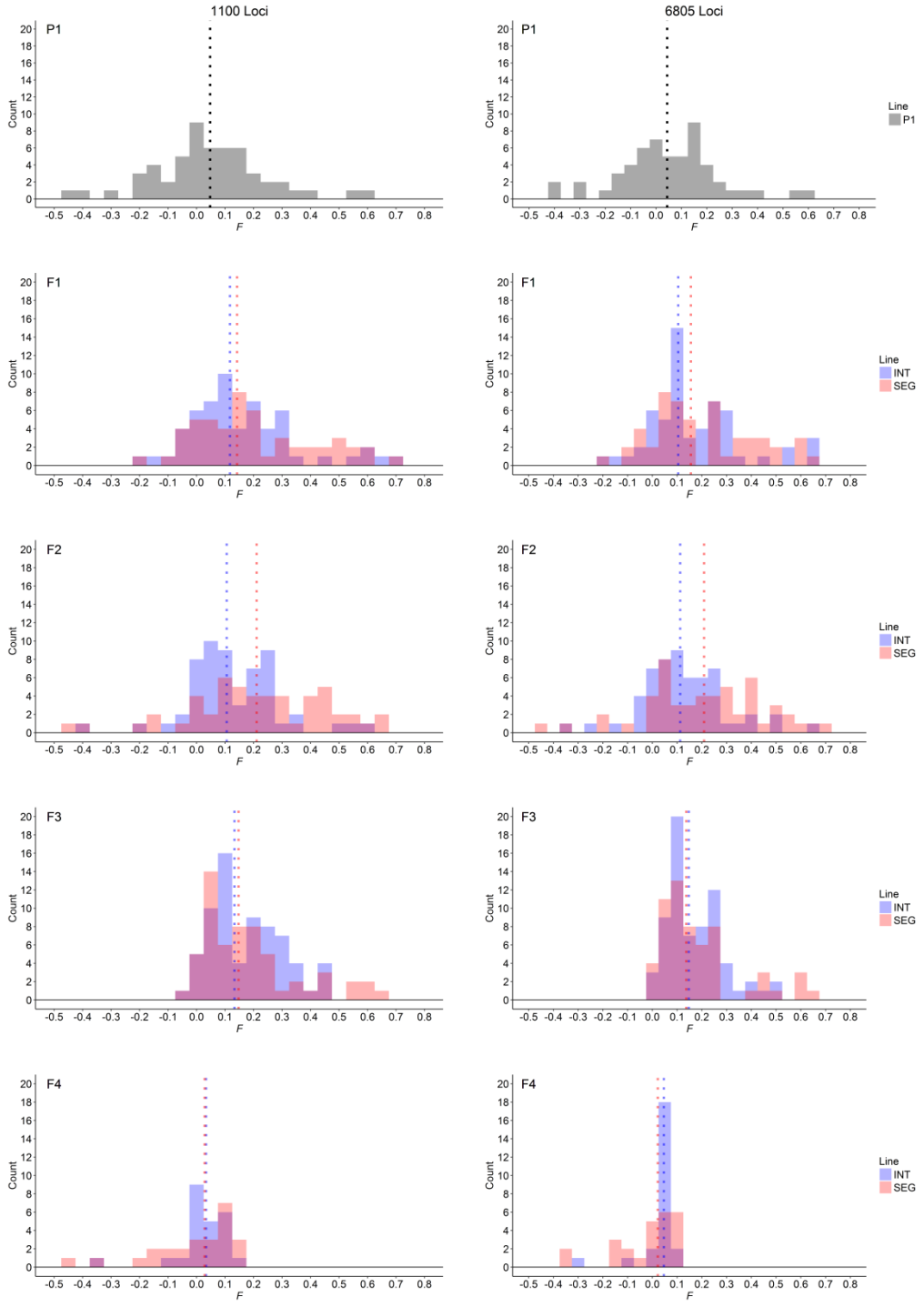

Figure 3.2 – Histograms of individual inbreeding coefficient, *F*, for the P_1 founders (gray) and the F_1, F_2, F_3, and F_4 generations of the integrated (INT, in blue) and segregated (SEG, in red)

hatchery lines estimated from 1,100 unlinked loci (left column) and all 6,805 study loci (right column). The median values for each distribution are denoted by dotted vertical lines.

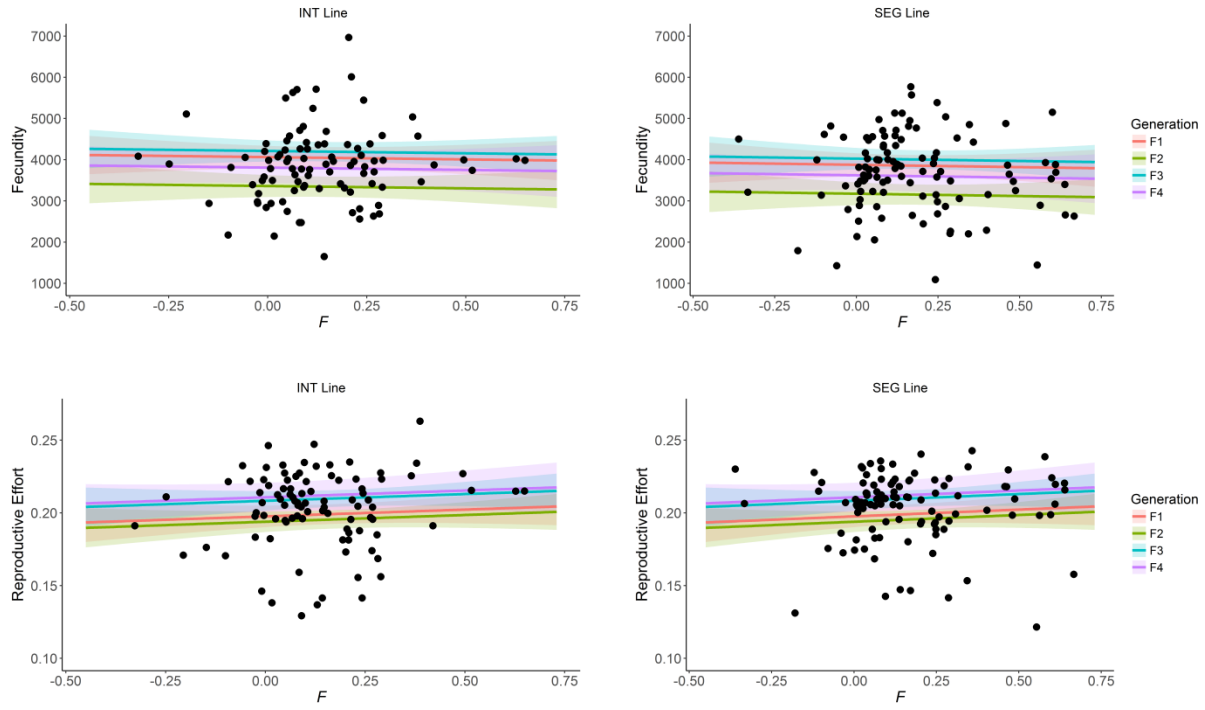

Figure 3.3 – Plots of fecundity (top row) and reproductive effort (bottom row) versus inbreeding coefficient estimated from 6,805 loci for females from the integrated (left column) and segregated (right column) hatchery lines. The solid lines represent the predicted relationship between phenotype and inbreeding coefficient for each generation based on the best fit model (Table 3.3). The shaded regions represent the 95% confidence intervals, and the black points are empirical data. Note that model predictions for the F_4 generation may lack precision due to low sample sizes.

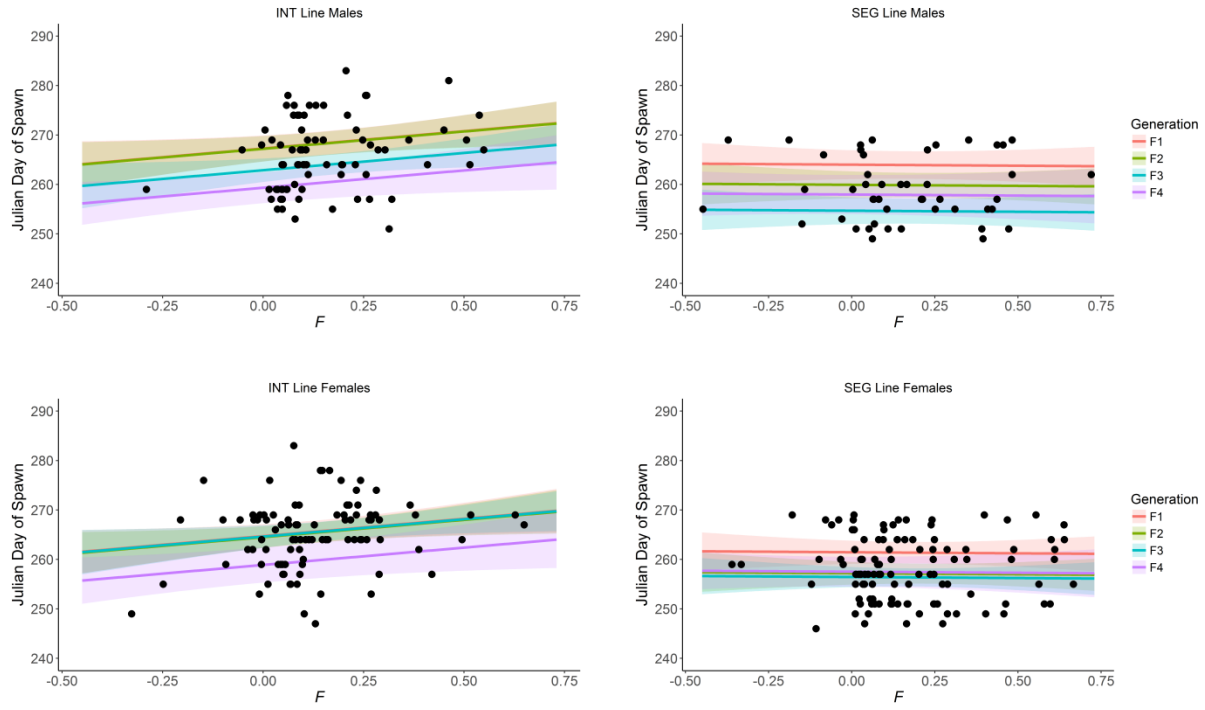

Figure 3.4 – Plots of spawn timing versus inbreeding coefficient estimated from 6,805 loci for males (top row) and females (bottom row) in the integrated (left column) and segregated (right column) hatchery lines. The solid lines represent the predicted relationship between phenotype and inbreeding coefficient for each generation based on the best fit model (Table 3.3). The shaded regions represent the 95% confidence intervals, and the black points are empirical data. Note that model predictions for the F_4 generation may lack precision due to low sample sizes.

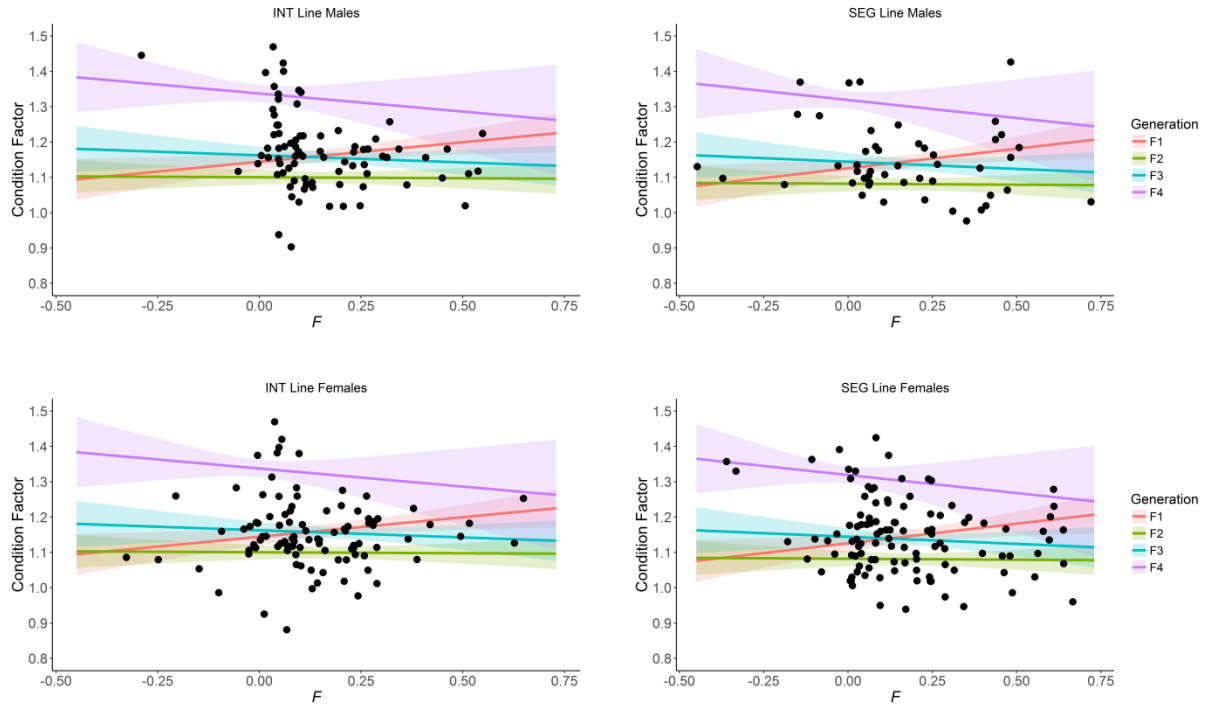

Figure 3.5 – Plots of condition factor versus inbreeding coefficient estimated from 6,805 loci for males (top row) and females (bottom row) in the integrated (left column) and segregated (right column) hatchery lines. The solid lines represent the predicted relationship between phenotype and inbreeding coefficient for each generation based on the best fit model (Table 3.3). The shaded regions represent the 95% confidence intervals, and the black points are empirical data. Note that model predictions for the F_4 generation may lack precision due to low sample sizes.

3.9 SUPPLEMENTARY MATERIAL

S3.1 – Supplementary information for results of Chapter 3

S3.1.1 – Plots of inbreeding coefficients from four probability-based estimators versus true level of inbreeding obtained from simulations implemented in COANCESTRY.

S3.1.2 – Plot of return timing versus inbreeding coefficient.

S3.1.3 – Plot of fork length versus inbreeding coefficient.

S3.1.4 – Plot of weight versus inbreeding coefficient.

S3.1.5 – Plot of daily growth coefficient versus inbreeding coefficient.

S3.1.6 – Plot of individuals from the P_1 founders and four generations of the integrated and segregated hatchery lines along the first and second discriminant functions from a discriminant analysis of principal components (DAPC).

S3.1.7 – Histogram of individual observed heterozygosity for all study individuals.

S3.2 – Supplementary tables containing metadata and results of Chapter 3

S3.2.1 – Summary of the number of fish used as broodstock, number of broodstock sampled for this study, and the number of fish retained in the final data set for each generation of each hatchery line.

S3.2.2 – Genotypes at 6,805 loci for the 452 individuals analyzed in this study.

S3.2.3 – Metadata and phenotypic measurements of eight traits for each individual.

S3.2.4 – (a) Results from the alignment of study loci to the Chinook salmon genome. (b) List of 1,100 unlinked loci identified by LD pruning in PLINK.

S3.2.5 – (a) Empirical genomic relatedness values estimated in PLINK. (b) Empirical individual inbreeding coefficients estimated in PLINK.

3.10 REFERENCES

Allendorf FW (1993) Delay of adaptation to captive breeding by equalizing family-size. *Conservation Biology*, **7**, 416-419.

Baskett ML, Waples RS (2013) Evaluating alternative strategies for minimizing unintended fitness consequences of cultured individuals on wild populations. *Conservation Biology*, **27**, 83-94.

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*, **18**, 503-511.

Bosse M, Megens HJ, Madsen O *et al.* (2015) Using genome-wide measures of coancestry to maintain diversity and fitness in endangered and domestic pig populations. *Genome Research*, **25**, 970-981.

Bowkett AE (2009) Recent captive-breeding proposals and the return of the ark concept to global species conservation. *Conservation Biology*, **23**, 773-776.

Brekke P, Bennett PM, Wang JL, Pettorelli N, Ewen JG (2010) Sensitive males: Inbreeding depression in an endangered bird. *Proceedings of the Royal Society B-Biological Sciences*, **277**, 3677-3684.

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, Weber JL (1999) Long homozygous chromosomal segments in reference families from the centre d'etude du polymorphisme humain. *American Journal of Human Genetics*, **65**, 1493-1500.

Busack C, Knudsen CM (2007) Using factorial mating designs to increase the effective number of breeders in fish hatcheries. *Aquaculture*, **273**, 24-32.

Calaprice JR (1969) Production and genetic factors in managed salmonid populations. In: *Symposium on salmon and trout in streams* (ed. Northcote TG), pp. 377-388. Institute of Fisheries, The University of British Columbia, Vancouver, British Columbia, Canada.

Cavole LM, Demko AM, Diner RE *et al.* (2016) Biological impacts of the 2013-2015 warm-water anomaly in the northeast pacific. *Oceanography*, **29**, 273-285.

Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. *Nature Reviews Genetics*, **10**, 783-796.

Cho CY (1992) Feeding systems for rainbow trout and other salmonids with reference to current estimates of energy and protein requirements. *Aquaculture*, **100**, 107-123.

Curik I, Ferencakovic M, Solkner J (2014) Inbreeding and runs of homozygosity: A possible solution to an old problem. *Livestock Science*, **166**, 26-34.

Duchesne P, Bernatchez L (2002) An analytical investigation of the dynamics of inbreeding in multi-generation supportive breeding. *Conservation Genetics*, **3**, 47-60.

Dupont-Nivet M, Karahan-Nomm B, Vergnet A *et al.* (2010) Genotype by environment interactions for growth in european seabass (*dicentrarchus labrax*) are large when growth rate rather than weight is considered. *Aquaculture*, **306**, 365-368.

Elias BA, Shipley LA, McCusker S, Sayler RD, Jobnson TR (2013) Effects of genetic management on reproduction, growth, and survival in captive endangered pygmy rabbits (*brachylagus idahoensis*). *Journal of Mammalogy*, **94**, 1282-1292.

Fast DE, Bosch WJ, Johnston MV *et al.* (2015) A synthesis of findings from an integrated hatchery program after three generations of spawning in the natural environment. *North American Journal of Aquaculture*, **77**, 377-395.
Ferenčaković M, Solkner J, Curik I (2013) Estimating autozygosity from high-throughput information: Effects of snp density and genotyping errors. *Genetics Selection Evolution*, **45**, 9.
Ford MJ (2002) Selection in captivity during supportive breeding may reduce fitness in the wild. *Conservation Biology*, **16**, 815-825.
Frankham R (2005) Genetics and extinction. *Biological Conservation*, **126**, 131-140.
Frankham R, Ballou JD, Briscoe DA (2002) *Introduction to conservation genetics* Cambridge University Press, Cambridge.
Gustafson RG, Waples RS, Myers JM *et al.* (2007) Pacific salmon extinctions: Quantifying lost and remaining diversity. *Conservation Biology*, **21**, 1009-1020.
Hammerly SC, Morrow ME, Johnson JA (2013) A comparison of pedigree- and DNA-based measures for identifying inbreeding depression in the critically endangered attwater's prairie-chicken. *Molecular Ecology*, **22**, 5313-5328.
Hard JJ, Connell L, Hershberger WK, Harrell LW (2000) Genetic variation in mortality of chinook salmon during a bloom of the marine alga *heterosigma akaskiwo*. *Journal of Fish Biology*, **56**, 1387-1397.
Hedrick PW, Garcia-Dorado A (2016) Understanding inbreeding depression, purging, and genetic rescue. *Trends in Ecology & Evolution*, **31**, 940-952.
Hoffman JI, Simpson F, David P *et al.* (2014) High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 3775-3780.
Huisman J, Kruuk LEB, Ellis PA, Clutton-Brock T, Pemberton JM (2016) Inbreeding depression across the lifespan in a wild mammal population. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 3585-3590.
Kardos M, Luikart G, Allendorf FW (2015) Measuring individual inbreeding in the age of genomics: Marker-based measures are better than pedigrees. *Heredity*, **115**, 63-72.
Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW (2016) Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, **9**, 1205-1218.
Keller LF, Waller DM (2002) Inbreeding effects in wild populations. *Trends in Ecology & Evolution*, **17**, 230-241.
Kincaid HL (1976) Inbreeding in rainbow trout (*salmo gairdneri*). *Journal of the Fisheries Research Board of Canada*, **33**, 2420-2426.
Knudsen CM, Schroder SL, Busack C, Johnston MV, Pearsons TN, Strom CR (2008) Comparison of female reproductive traits and progeny of first-generation hatchery and wild upper yakima river spring chinook salmon. *Transactions of the American Fisheries Society*, **137**, 1433-1445.
Knudsen CM, Schroder SL, Busack CA *et al.* (2006) Comparison of life history traits between first-generation hatchery and wild upper yakima river spring chinook salmon. *Transactions of the American Fisheries Society*, **135**, 1130-1144.
Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. *Nature Methods*, **9**, 357-U354.
Lynch M, O'Hely M (2001) Captive breeding and the genetic fitness of natural populations. *Conservation Genetics*, **2**, 363-378.

Mobrand LE, Barr J, Blankenship L *et al.* (2005) Hatchery reform in washington state: Principles and emerging issues. *Fisheries*, **30**, 11-23.
Naish KA, Seamons TR, Dauer MB, Hauser L, Quinn TP (2013) Relationship between effective population size, inbreeding and adult fitness-related traits in a steelhead (*oncorhynchus mykiss*) population released in the wild. *Molecular Ecology*, **22**, 1295-1309.
Naish KA, Taylor JE, Levin PS *et al.* (2008) An evaluation of the effects of conservation and fishery enhancement hatcheries on wild populations of salmon. In: *Advances in marine biology*, pp. 61-194. Elsevier Academic Press Inc, San Diego.
National Research Council (1996) *Upstream : Salmon and society in the pacific northwest* National Academy Press, Washington, D.C.
O'Grady JJ, Brook BW, Reed DH, Ballou JD, Tonkyn DW, Frankham R (2006) Realistic levels of inbreeding depression strongly affect extinction risk in wild populations. *Biological Conservation*, **133**, 42-51.
Paquet PJ, Flagg T, Appleby A *et al.* (2011) Hatcheries, conservation, and sustainable fisheries-achieving multiple goals: Results of the hatchery scientific review group's columbia river basin review. *Fisheries*, **36**, 547-561.
Purcell S, Neale B, Todd-Brown K *et al.* (2007) Plink: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559-575.
R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org, Vienna, Austria.
RASP (1992) Supplementation in the columbia river basin summary report series. In: *Final Report to Bonneville Power Administration* (ed. Planning RAoS), Portland, Oregon.
Reid JM, Keller LF, Marr AB, Nietlisbach P, Sardell RJ, Arcese P (2014) Pedigree error due to extra-pair reproduction substantially biases estimates of inbreeding depression. *Evolution*, **68**, 802-815.
Ricker WE (1975) Computation and interpretation of biological statistics of fish populations. *Bulletin of the Fisheries Research Board of Canada*, 1-382.
Ryman N (1970) A genetic analysis of recapture frequencies of released young of salmon (*salmo salar l.*). *Hereditas-Genetiskt Arkiv*, **65**, 159-&.
Saccheri I, Kuussaari M, Kankare M, Vikman P, Fortelius W, Hanski I (1998) Inbreeding and extinction in a butterfly metapopulation. *Nature*, **392**, 491-494.
Santure AW, Stapley J, Ball AD, Birkhead TR, Burke T, Slate J (2010) On the use of large marker panels to estimate inbreeding and relatedness: Empirical and simulation studies of a pedigreed zebra finch population typed at 771 snps. *Molecular Ecology*, **19**, 1439-1451.
Schroder SL, Knudsen CM, Pearsons TN *et al.* (2010) Behavior and breeding success of wild and first-generation hatchery male spring chinook salmon spawning in an artificial stream. *Transactions of the American Fisheries Society*, **139**, 989-1003.
Schroder SL, Knudsen CM, Pearsons TN *et al.* (2008) Breeding success of wild and first-generation hatchery female spring chinook salmon spawning in an artificial stream. *Transactions of the American Fisheries Society*, **137**, 1475-1489.
Snyder NFR, Derrickson SR, Beissinger SR *et al.* (1996) Limitations of captive breeding in endangered species recovery. *Conservation Biology*, **10**, 338-348.
Taylor HR (2015) The use and abuse of genetic marker-based estimates of relatedness and inbreeding. *Ecology and Evolution*, **5**, 3140-3150.

Taylor HR, Kardos MD, Ramstad KM, Allendorf FW (2015) Valid estimates of individual inbreeding coefficients from marker-based pedigrees are not feasible in wild populations with low allelic diversity. *Conservation Genetics*, **16**, 901-913.
Thrower FP, Hard JJ (2009) Effects of a single event of close inbreeding on growth and survival in steelhead. *Conservation Genetics*, **10**, 1299-1307.
Venables WN, Ripley BD (2002) *Modern applied statistics with s*, Fourth edn. Springer, New York.
Wang J (2007) Triadic ibd coefficients and applications to estimating pairwise relatedness. *Genetics Research*, **89**, 135-153.
Wang SZ, Hard J, Utter F (2002) Salmonid inbreeding: A review. *Reviews in Fish Biology and Fisheries*, **11**, 301-319.
Waters CD, Hard JJ, Brieuc MSO *et al.* (2018) Genomewide association analyses of fitness traits in captive-reared chinook salmon: Applications in evaluating conservation strategies. *Evolutionary Applications*, **00**, 1-16.
Waters CD, Hard JJ, Brieuc MSO *et al.* (2015) Effectiveness of managed gene flow in reducing genetic divergence associated with captive breeding. *Evolutionary Applications*, **8**, 956-971.
Waters CD, Hard JJ, Brieuc MSO *et al.* (2017) What can genomics tell us about the success of enhancement programs in anadromous chinook salmon? A comparative analysis across four generations. Supplementary material for bernatchez et al. (2017) harnessing the power of genomics to secure the future of seafood. *Trends in Ecology & Evolution*, 665-680.
Woodworth LM, Montgomery ME, Briscoe DA, Frankham R (2002) Rapid genetic deterioration in captive populations: Causes and conservation implications. *Conservation Genetics*, **3**, 277-288.
Zhang QQ, Calus MPL, Guldbrandtsen B, Lund MS, Sahana G (2015) Estimation of inbreeding using pedigree, 50k snp chip genotypes and full sequence data in three cattle breeds. *Bmc Genetics*, **16**, 11.

Chapter 4. Genome-wide association analyses of fitness traits in captive-reared Chinook salmon: Applications in evaluating conservation strategies[3]

4.1 ABSTRACT

A novel application of genome-wide association analyses is to use trait-associated loci to monitor the effects of conservation strategies on potentially adaptive genetic variation. Comparisons of fitness between captive- and wild-origin individuals, for example, do not reveal how captive rearing affects genetic variation underlying fitness traits or which traits are most susceptible to domestication selection. Here, we used data collected across four generations to identify loci associated with six traits in adult Chinook salmon (*Oncorhynchus tshawytscha*), and then determined how two alternative management approaches for captive rearing affected variation at these loci. Loci associated with date of return to freshwater spawning grounds (return timing), length and weight at return, age at maturity, spawn timing, and daily growth coefficient were identified using 9108 restriction site-associated markers and Random Forest, an approach suitable for polygenic traits. Mapping of trait-associated loci, gene annotations, and integration of results across multiple studies revealed candidate regions involved in several fitness-related traits. Genotypes at trait-associated loci were then compared between two hatchery populations that were derived from the same source but are now managed as separate lines, one integrated with and one segregated from the wild population. While no broad scale change was detected

[3] This chapter has been published as Genomewide association analyses of fitness traits in captive-reared Chinook salmon: Applications in evaluating conservation strategies, Waters CD, Hard JJ, Brieuc MSO, Fast DE, Warheit KI, Knudsen CM, Bosch WJ, and Naish KA (2018), *Evolutionary Applications* **00**, 1-16.

across four generations, there were numerous regions where trait-associated loci overlapped with signatures of adaptive divergence previously identified in the two lines. Many regions, primarily with loci linked to return and spawn timing, were either unique to, or more divergent in, the segregated line, suggesting that these traits may be responding to domestication selection. This study is one of the first to utilize genomic approaches to demonstrate the effectiveness of a conservation strategy, managed gene flow, on trait-associated – and potentially adaptive – loci. The results will promote the development of trait-specific tools to better monitor genetic change in captive and wild populations.

4.2 INTRODUCTION

There is considerable interest in applying the results of genome-wide association analyses to conservation and management (Funk *et al.* 2012; Harrisson *et al.* 2014; Hoffmann *et al.* 2015; Shafer *et al.* 2015; Bernatchez 2016; Garner *et al.* 2016; Pearse 2016; Bernatchez *et al.* 2017). In fact, identifying the genetic basis of fitness traits has already provided key information for these purposes, including an improved understanding of adaptive divergence (Brieuc *et al.* 2015; Hornoy *et al.* 2015), the discrimination of ecotypes within a panmictic species (Pavey *et al.* 2015), detecting polygenic selection to aquatic pollutants (Laporte *et al.* 2016), and the development of a marker panel for routine trait and population monitoring (Barson *et al.* 2015; Aykanat *et al.* 2016). As the availability of genomic resources improves, it is important to explore how markers identified by association analyses on natural populations might be applied in different contexts.

Trait-associated markers have significant potential to inform the management of captive breeding programs. Captive breeding remains one of the primary options for the conservation of

threatened populations and species (e.g. Griffiths & Pavajeau 2008; Conde *et al.* 2011; Horne *et al.* 2016; Landa *et al.* 2017). However, this approach is also controversial, because associated genetic and phenotypic changes may decrease the fitness of captive individuals when they are released into the wild and, consequently, reduce restoration success (Frankham 2008; Jule *et al.* 2008; Christie *et al.* 2014). Surveys of variation at trait-associated loci would improve our understanding of how captive breeding affects potentially adaptive genetic variation and would help to identify traits, and even specific alleles, that may drive observed fitness reductions (e.g. Bateson *et al.* 2016). Trait-based monitoring could also inform policy decisions that aim to minimize negative effects of captive breeding, including practices to reduce domestication selection.

Population supplementation using hatcheries, a form of captive breeding, is part of many recovery plans for Pacific salmon on the West Coast of North America, where numerous populations have declined or been extirpated in the last century (National Research Council 1996). In these programs, adult salmon are brought into the hatchery for reproduction, and their offspring are reared in captivity for up to two years before they are released into the wild as seaward migrants. Despite spending only a portion of their lives in captivity, hatchery-reared salmon have exhibited significant differences from their wild counterparts, including reduced reproductive success (Christie *et al.* 2014), differences in growth rate and morphology (McGinnity *et al.* 2003; Busack *et al.* 2007), and increased vulnerability to predation (Fritts *et al.* 2007). One practice that has been widely adopted to mitigate genetic and phenotypic risks of hatchery rearing is the integration of wild or natural-origin (born in the wild but may have hatchery ancestry) fish into hatchery broodstock (Mobrand *et al.* 2005; Paquet *et al.* 2011). This "integrated" approach, which we refer to as managed gene flow, contrasts with the traditional

"segregated" strategy where only hatchery-origin fish are used as broodstock. Theoretical studies based on genetic models predict that managed gene flow reduces, but does not eliminate, the effects of genetic drift, inbreeding, and domestication selection that may occur due to captive breeding (Lynch & O'Hely 2001; Duchesne & Bernatchez 2002; Ford 2002; Baskett & Waples 2013). In addition, we recently provided the first empirical genetic evidence demonstrating the benefits of managed gene flow in captive-reared fish. Specifically, genome-wide divergence over four generations of captive rearing was reduced in an integrated hatchery population of Chinook salmon when compared to a program based on broodstock segregation (Waters *et al.* 2015; Waters *et al.* 2017), signifying that the overall genetic risks of captive rearing may be reduced through gene flow. Here, we extend this work by using loci associated with several fitness-related traits to explore how the integrated and segregated management strategies affect genetic variation at trait-associated, and potentially adaptive, loci.

The spring-run Chinook salmon hatchery program at the Cle Elum Supplementation and Research Facility (CESRF) in Cle Elum, Washington, USA is unique in that it maintains an integrated hatchery line and a segregated hatchery line, both of which were derived from the same wild population at the same time (Figure 4.1). Importantly, tissue samples for DNA and phenotypic data have been collected from every adult fish used as broodstock since the inception of the program in 1997. Many of the phenotypic traits measured – length, weight, and dates of return to freshwater and maturation – are correlated with individual fitness (e.g. Thorpe *et al.* 1984; Schroder *et al.* 2010; Kodama *et al.* 2012). In addition, these traits have significant additive genetic variation on which selection can act (Hard 2004; Carlson & Seamons 2008) and can differ between hatchery and wild populations (e.g. Ford *et al.* 2006; Knudsen *et al.* 2006; Hoffnagle *et al.* 2008). Therefore, studying the genetic basis of these specific traits may provide

a better understanding of how domestication selection acts and, in turn, reveal possible mechanisms underlying the reduced fitness of hatchery-origin fish after they are released into the wild.

Here, we used individual-based data from the two hatchery lines at CESRF to 1) identify loci associated with six fitness-related traits that have been measured in returning adults across four generations and 2) determine if managed gene flow successfully limited divergence at trait-associated loci relative to a segregated management approach. Specifically, we first characterized the genomic basis of date of return to freshwater spawning grounds (return timing), fork length and weight at return, age at maturity, spawn timing, and daily growth coefficient using 9108 restriction site-associated (RAD) loci and a genome-wide association analysis suitable for polygenic traits, Random Forest (Breiman 2001). Loci associated with each trait were then annotated to provide biological context for the genotype-phenotype associations. Next, genetic variation at loci associated with each trait was compared between each generation of the integrated and segregated hatchery lines to determine if managed gene flow limited divergence. We also compared the genomic positions of trait-associated loci and highly diverged loci – interpreted as signatures of adaptive divergence – that had been previously identified in the two lines (Waters *et al.* 2015; Waters *et al.* 2017). Overlap between the two groups of loci was used to infer which traits may have responded to domestication selection. This study represents one of the first efforts to evaluate the effectiveness of a conservation strategy by examining loci linked to multiple traits within a non-model organism. The results provide molecular tools for monitoring captive populations and could inform management practices to reduce possible adverse effects of captive rearing.

4.3 METHODS

Study system

The ecological background of the study population and the initiation of the integrated and segregated hatchery lines at CESRF have been described in previous publications (Knudsen *et al.* 2006; Fast *et al.* 2015; Waters *et al.* 2015). Briefly, wild adults returning from the ocean to their freshwater spawning grounds were collected for founding broodstock (Figure 4.1) from the upper Yakima River, WA, USA population from 1997–2002 as they passed the Roza Dam Adult Monitoring Facility (RAMF), located 90 river kilometers south of CESRF. Broodstock were collected at random over the entire duration of the salmon run and transported to CESRF, where they were held in concrete raceways until maturation. Mature adults were spawned at random following a 3x3 factorial mating design (when possible). Fertilized eggs, fry, and juveniles were reared at CESRF for approximately 16 months, after which they were transferred to three acclimation sites; these sites were designed to expand the spatial influence of supplementation efforts while also enabling related research. Fish were acclimated for two months, at which point they were allowed to volitionally begin their migration to the ocean. A majority (≥75%; Knudsen *et al*. 2006) of Chinook from this population spend two years in the ocean and return at age four to reproduce.

In 2002, both wild and first-generation hatchery adults were spawned to create the integrated (INT) and segregated (SEG) hatchery lines, respectively (Figure 4.1). The integrated line uses only fish born in the wild as broodstock, and all returning adults from this line are allowed to spawn in the river. In contrast, only returning hatchery-origin fish are used as broodstock in the segregated line, and SEG adults are not allowed to reproduce naturally; fish from the two lines are differentially marked for external identification, so all SEG adults are

removed from the system at the RAMF. Therefore, the integrated line receives one generation of exposure to hatchery conditions while the segregated line is exposed every generation (Figure 4.1). Broodstock collection, spawning, and rearing procedures for both lines are conducted in the same manner as the founding generation.

Tissue sample and phenotypic data collection

Tissue samples for DNA were collected from all fish during spawning at CESRF and stored in 100% ethanol. Five generations were sub-sampled for this study: the 1998 wild founders (second founding year; P_1 Founders) and hatchery brood years 2002 (F_1 Wild and F_1 Hatchery), 2006 (F_2 INT and F_2 SEG), 2010 (F_3 INT and F_3 SEG), and 2014 (F_4 INT and F_4 SEG).

Phenotypic traits of all returning adults were measured annually upon arrival at RAMF and, for those collected as broodstock, during spawning at CESRF. Six traits were analyzed in this study (Table 4.1). Ages at maturity of hatchery-origin broodstock were determined from passive integrated transponder or coded-wire tags, which denote the brood year of each fish, while those of natural-origin broodstock and hatchery-origin fish without tags were determined from growth rings on their scales (Clutter & Whitesel 1956). Ages of fish not used as broodstock (i.e. those allowed to spawn naturally or removed from the system) were estimated with relatively high accuracy (up to 90%; C. Knudsen, *pers. comm.*) from body size measurements at RAMF. Return timing corresponds to day of arrival at RAMF. Spawn, or maturation, timing of all broodstock was estimated weekly at CESRF by manually checking for gonadal ripeness. Fork length and weight were measured at both RAMF and upon spawning at CESRF. However, only measurements from RAMF were analyzed here to minimize possible influences of the kype, a

secondary sex trait in males that develops during maturation and may increase jaw length up to 50% (Fleming 1996), and of time spent in the hatchery after collection at RAMF. Daily growth coefficient (DGC) was calculated according to Cho (1992) and Dupont-Nivet *et al*. (2010):

$$DGC = 100 \times [((\text{final weight})^{1/3} - (\text{initial weight})^{1/3})/\text{days}] \quad (1)$$

where initial weight was the weight at RAMF, final weight was weight at spawning, and days was the number of days between arrival at RAMF and spawning at CESRF. DGC has been shown to be more independent of initial weight than other measures, such as specific growth rate, and is thus better for growth rate comparisons (Cho 1992). Here, DGC provided a measure related to weight lost between RAMF and CESRF and was used to infer if the two hatchery lines responded differently to holding conditions in the tanks at CESRF (e.g. if fish from the segregated line exhibited lower daily growth coefficients and levels of stress due to holding); such differences may indicate adaptation to captivity.

DNA sequencing and genotyping

DNA from tissue samples was extracted using DNeasy Blood & Tissue kits (Qiagen, Valencia, CA) following the animal tissue protocol. RAD libraries (Baird *et al.* 2008) were prepared using the restriction enzyme *Sbf1* and by pooling 24-36 barcoded individuals per lane (28 lanes total). Single-read (100bp) sequencing was performed using an Illumina HiSeq2000.

RAD sequences were processed using *Stacks* (v. 1.09, Catchen *et al.* 2013). First, reads were demultiplexed and trimmed to 74 base pairs using *process_radtags*, as sequencing errors increased after this length. Reads were then aligned to a reference database of 48528 putatively non-duplicated RAD loci for Chinook salmon, 6350 of which are positioned on a linkage map (Brieuc *et al.* 2014), using *Bowtie* (v. 0.12.8, Langmead *et al.* 2009) with the "best" option and

allowing up to three mismatches. Loci for each individual were identified using *pstacks* in *Stacks* with the bounded-error SNP calling model, default error rates, and a minimum stack depth of 10 reads. A catalog of loci was then constructed from the five most-sequenced individuals per population using *cstacks*; a subset of individuals was used to reduce the risk of including false polymorphisms in the catalog (e.g. Waters *et al.* 2015; Hess *et al.* 2016; Nichols *et al.* 2016; Narum *et al.* 2017). Loci of individuals were matched to the catalog using *sstacks*, and genotypes were aggregated using *populations*.

After processing with *Stacks*, all biallelic loci were re-genotyped with a custom Python script to minimize potential bias in maximum likelihood genotype calls due to differences in read depth between two alleles at a locus. The custom script (Brieuc *et al.* 2014) designated a genotype as heterozygous if both alleles had a minimum depth of two and a combined read depth greater than 10. Loci were then screened and retained if they had a minor allele frequency ≥0.05 in at least one population. Individuals were removed if they had ≥50% missing genotypes across all loci. Then, individual loci were removed if they were not genotyped in at least 50% of the individuals in each population.

Missing genotypes in the final data set were subsequently imputed using *fastPhase* (Scheet & Stephens 2006), because association analyses based on Random Forest cannot process missing data. Deviations from Hardy-Weinberg equilibrium were then identified using the exact test in *Genepop* (v. 4.1, Raymond & Rousset 1995) with default parameter values. Loci that did not conform to expected Hardy-Weinberg proportions in two or more populations were removed (e.g. Benestan *et al.* 2015; Nichols *et al.* 2016), as deviations may arise from factors such as genotyping errors or the inclusion of paralogous sequence variants that remain after the whole genome duplication event in salmonids (Allendorf & Thorgaard 1984). However, we

acknowledge that this conservative approach may also remove loci of interest, such as those responding to domestication selection.

Inferring positions of unmapped loci

To increase the number of loci positioned on the Chinook linkage map, positions of unmapped loci were inferred by alignments to the rainbow trout (Berthelot *et al.* 2014) and Atlantic salmon (Lien *et al.* 2016) genomes, similar to the approach developed by Sutherland *et al.* (2016) and employed by Narum *et al.* (2017). First, all non-duplicated loci on the Chinook salmon linkage map (n=6350, Brieuc *et al.* 2014) and unmapped loci in the final data set were aligned to both genomes using *Bowtie 2* (v. 2.2.9, Langmead & Salzberg 2012) with default parameters. Next, *Bowtie 2* output files were converted from SAM to BAM format using *SAMtools* (Li *et al.* 2009), and then from BAM to BED format using *bedtools* (v. 2.26.0, Quinlan & Hall 2010). The *closest* function in *bedtools* then identified high-quality alignments (mapping quality ≥10) of mapped and unmapped Chinook loci that were within 100 kb of each other. A distance of 100 kb was chosen because mapped loci with identical linkage map positions frequently aligned to the rainbow trout or Atlantic salmon genomes within 100-300 kb of each other. Unmapped loci were assigned the same linkage map positions as the closest mapped loci.

Random Forest analyses

Loci associated with the phenotypic traits were identified by Random Forest, an approach suitable for simple and polygenic traits (Breiman 2001). Unlike traditional approaches, Random Forest provides a non-parametric framework that can simultaneously incorporate many loci, thus

enabling the identification of suites of loci that explain substantial phenotypic variation collectively but may not display significant associations individually.

Random Forest, like other methods, can be confounded by population structure and additional factors (Zhao *et al.* 2012; Stephan *et al.* 2015). Here, the approach of Zhao *et al.* (2012) was used to correct for the possible confounding factors of hatchery line, sex, age, year, and the coordinate of individuals on PC1 from a principal components analysis of the genotypes. Phenotypes and genotypes at all loci (one locus at a time) were regressed against these potential confounding factors using linear regression models in R (R Core Team 2017). The models explained any variation in the phenotypes and genotypes that may be due to differences between the two hatchery lines or any of the other factors. The residuals of the models, representing the "corrected" phenotypes and genotypes, were then analyzed by Random Forest to identify loci strictly associated with the traits of interest. When age at maturity was the response variable, genotypes were corrected for all possible confounding factors (except age), but phenotypes were not corrected to maintain their discrete (categorical) distribution, rather than transform them into continuous variables.

Random Forest analyses were conducted using the R package *randomForest* (Liaw & Wiener 2002). The P_1 founders were excluded so that each year analyzed comprised only samples from the integrated and segregated hatchery lines (i.e. paired design). Age, as a discrete trait with an unequal distribution across values, was analyzed using balanced classification trees (specified with the *strata* and *sampsize* parameters in the *randomForest* function). Regression trees with default parameter settings were employed for the five continuous traits. For all traits, three independent forests were first grown using all loci. The optimal number of trees (*ntree* parameter) per forest was determined when the correlation of locus importance values between

forests was relatively high (>0.8). Initial importance values for loci were then estimated by averaging values from the three forests. Next, three forests were grown on subsets of the most important loci (e.g. top 1%, top 2%, etc.) to determine a candidate group of loci that explained the highest proportion of phenotypic variation observed (for the five continuous traits) or had the lowest out-of-bag classification error rate (for age at maturity). The backward purging approach (Holliday *et al.* 2012) was then conducted on an expanded group of candidate loci (i.e. used top 1.5% of loci if the top 1% of loci appeared to explain the most phenotypic variation) to precisely determine groups of loci that explained the most phenotypic variation or had the lowest classification error rate. Loci identified by backward purging were deemed to be predictive of the phenotypes (i.e. trait-associated loci).

Gene annotation

As there is no reference genome for Chinook salmon, functions of genes near trait-associated loci were determined by first aligning loci to the rainbow trout genome (Berthelot *et al.* 2014) using *Bowtie 2* (v. 2.2.9, Langmead & Salzberg 2012) with default parameters. After file format conversion, the *closest* function in *bedtools* (v. 2.26.0, Quinlan & Hall 2010) was then used to identify the rainbow trout gene that was closest to each aligned locus. Coding sequences of genes were then aligned against the UniProtKB/Swiss-Prot database using a BLASTx search (Altschul *et al.* 1990) with default parameter settings and an *e*-value threshold of $1x10^{-10}$. The search identified UniProtKB/Swiss-Prot entry names and protein products associated with each gene. Gene Ontology (GO) and GO Slim terms were then identified for each entry name; this approach assumes shared functionality across species. GO Slim terms related to biological processes were summarized to provide biological context for trait-associated loci. However, to

ensure the quality of summaries, GO Slim terms were only used for loci that aligned to the rainbow trout genome with a mapping quality ≥10 and were within 100 kb of a gene. In addition, we retained all GO Slim terms for each protein but, if the same term appeared multiple times for a protein, we counted the term only once to avoid overrepresentation of individual proteins in annotation summaries.

Phenotypic differences between hatchery lines

Phenotypic differences between the integrated and segregated lines were quantified because they may provide the first indication of genetic change at trait-associated loci. For example, divergence at trait-associated loci may be more likely if the two lines also exhibit large phenotypic differences. Such divergence may indicate the action of domestication selection since the two hatchery lines experience different levels of exposure to the hatchery. Here, differences were quantified using linear or generalized linear models in R (R Core Team 2017). Return timing, spawn timing, fork length, and daily growth coefficient were modeled with a Gaussian distribution and an identity link function, while age at maturity was modeled using a Poisson distribution and a log link function. The general model form for each of these traits was:

$$y_i = \beta_0 + \beta_1 \text{Line}_i + \beta_2 \text{Sex}_i + \beta_3 \text{Line}_i * \text{Sex}_i + e_i \quad (2)$$

where y_i was the trait measurement, Line_i was the hatchery line, and Sex_i was the sex for individual *i*, respectively. The model for weight comprised the same explanatory variables as (2) but also accounted for allometric growth (e.g. Thorson 2015) by including fork length in the form of:

$$\log(w_i) = \beta_0 + \beta_1 \text{Line}_i + \beta_2 \text{Sex}_i + \beta_3 \text{Line}_i * \text{Sex}_i + \beta_4 \log(\text{fork length}) + e_i \quad (3)$$

Divergence between the lines, if present, would be more likely to occur in recent years after multiple generations of hatchery propagation. Therefore, models were analyzed using only measurements from 2010 (F_3 generation), which was the most recent year for which relatively large sample sizes (>50) existed for all traits. Models for four traits – age at maturity, return timing, fork length, and weight – were analyzed using measurements collected from all fish that returned to RAMF in 2010. Spawn timing and daily growth coefficient were analyzed using fish sampled for RAD sequencing, as these traits were only measured in the subset of adults used as broodstock. With the exception of age at maturity, traits were analyzed using only four year old fish because they represent >75% of all adults (Knudsen *et al.* 2006).

Effectiveness of managed gene flow and traits affected by domestication selection

Inferences regarding the effectiveness of managed gene flow to limit divergence at trait-associated loci, as well as specific traits potentially affected by domestication selection, were made using two comparative methods. First, genetic variation at trait-associated loci identified by Random Forest was compared between the integrated and segregated hatchery lines across four generations using principal components analyses (PCA). Temporal divergence of the segregated line from the integrated line in multivariate space, possibly resulting from domestication selection, would suggest that managed gene flow limited change at trait-associated loci.

Second, trait-associated loci were compared to loci and genomic regions that were previously identified from the same samples as exhibiting signals of adaptive divergence between the integrated and segregated hatchery lines (i.e. outlier loci and regions; Waters *et al.* 2015; Waters *et al.* 2017). Outlier loci and regions had been identified using three tests – F_{TEMP}

(Therkildsen *et al.* 2013), *Bayescan* (Foll & Gaggiotti 2008), and the sliding window approach employed by Hohenlohe et al. (2010). F_{TEMP} detects selection in a single population sampled across multiple generations by simulating genetic drift over time and identifying loci that exceed neutral expectations; we conducted F_{TEMP} separately for each hatchery line to identify outliers. *Bayescan* estimates population- and locus-specific components of F_{ST} for each locus and identifies outliers when the locus-specific component is significantly different from zero; due to the small number of populations in our analysis, this program was run with both hatchery lines combined and thus identified a single set of outliers. The sliding window approach randomly samples empirical F_{ST} values from all mapped study loci to generate a null distribution reflecting genome-wide divergence due to neutral processes. Outlier regions were identified for each generation of each hatchery line as those where the moving average of F_{ST} (compared to the P_1 Founders) exceeded the 95% confidence interval of the null distribution. Overlap between outliers identified by these three methods and trait-associated loci was interpreted as evidence that specific traits had responded to domestication selection.

4.4 RESULTS

Tissue sample and phenotypic data collection

Tissues of 753 returning adult Chinook salmon, representing the 1998 wild founders and four generations of each hatchery line, were sub-sampled (Table S4.1.1). Phenotypic traits (Table 4.1) for these and other adults that returned to the upper Yakima River were measured at RAMF (Table S4.1.2a,b). For fish used as broodstock, phenotypic data were also recorded upon spawning at CESRF (Table S4.1.2b).

DNA sequencing and genotyping

DNA from 696 of the 753 tissue samples was successfully sequenced, and 11809 biallelic loci were identified using *Stacks*. Re-genotyping of these loci using the custom Python script of Brieuc *et al*. (2014) improved the distribution of F_{IS} values based on Hardy-Weinberg expectations (e.g. median F_{IS} across loci for P_1 Founders was approximately 0.3 and 0.1 before and after the custom script, respectively). Filtering based on minor allele frequency and missing genotypes reduced the data set to 465 individuals and 9266 loci. Missing genotypes were then imputed for Random Forest analyses; the correlation of allele frequencies at each locus before and after imputation was 0.998 (Figure S4.2.1), suggesting that imputation did not significantly compromise the data. Genotypes at 158 loci significantly deviated from expected Hardy-Weinberg proportions in more than one population and were removed. Of these 158 loci, 103 and 125 significantly deviated from Hardy-Weinberg equilibrium in the P_1 Founders and F_1 Wild populations, respectively. These samples represent pure wild fish (i.e. no exposure to hatchery conditions); thus, deviations at these loci in these populations occurred from factors other than domestication selection. The final data set comprised 465 individuals genotyped at 9108 loci (Tables S4.1.3, S4.1.4).

Inferring positions of unmapped loci

The 9108 study loci included 4156 mapped loci and 4952 unmapped loci. Of the unmapped loci, 616 and 198 aligned to the rainbow trout and Atlantic salmon genomes, respectively, with mapping qualities >= 10 and were within 100 kb of a mapped locus. After accounting for those that aligned to both genomes, 700 unmapped loci were assigned the same

linkage map positions as their corresponding mapped loci (Tables S4.1.3, S4.1.5). Therefore, the positions of 4856 loci were known for downstream analyses.

Random Forest analyses

The number of individuals analyzed by Random Forest varied slightly for each trait (Table 4.2), as those with missing phenotypes were removed prior to analyses. For age at maturity, sample sizes of three-, four-, and five-year old fish were 23, 344, and 16, respectively; the use of balanced classification trees accounted for this unequal distribution. The number of trees needed to obtain high correlation (>0.8) of importance values between initial forests varied from 250,000 to 750,000 for the traits.

The backward purging approach identified 226 unique predictor (i.e. trait-associated) loci over all traits, with 53% having known or inferred positions on the Chinook salmon linkage map (Table 4.2; Table S4.1.6). The number of predictor loci per trait ranged from 26 to 68. All predictors for each of the five continuous traits explained 26.7% to 32.0% of the observed phenotypic variation, while loci predictive of age at maturity had an out-of-bag classification error rate of 24.9% (i.e. rate at which the hold-out samples are misclassified using the Random Forest model; Table 4.2).

Trait-associated loci were identified on 33 of the 34 chromosomes for Chinook salmon (Figure S4.2.3; Table S4.1.6). Notably, some loci were associated with multiple traits. For example, 12 loci were identified as predictors for both fork length and weight at return (e.g. Figures 4.2a, 4.2b). In addition, one locus was associated with both daily growth coefficient and return timing on chromosome Ots21, while an unmapped locus was shared between fork length and spawn timing. Other predictor loci did not overlap but still mapped to the same genomic

regions, both within and across traits. For instance, two loci associated with daily growth coefficient were within 3cM of each other on Ots10 (Figure 4.2d), and two loci predictive of spawn timing mapped to the same position on Ots31. Two loci for weight, two for spawn timing, and one for both fork length and weight all mapped to a 13cM region on chromosome Ots08 (Figure 4.2b). Similarly, two loci associated with fork length and one locus associated with weight mapped to a 5cM region on chromosome Ots10 (Figure 4.2d). Two maturation-related traits, spawn timing and age at maturity, had predictor loci within 0.01cM of each other on three chromosomes (Ots01, Ots02, Ots26; Figure 4.2a). Loci for daily growth coefficient and return timing mapped to the same position on chromosome Ots08 (Figure 4.2b), while loci for daily growth coefficient and spawn timing mapped to similar positions on Ots09 (Figure 4.2c). Lastly, two narrow regions on Ots04 and Ots19 (same position and 2.2cM wide, respectively) each contained one locus associated with size (fork length or weight) and one for daily growth coefficient.

Gene annotation

Annotations for 75 of the 226 trait-associated loci were obtained after filtering for loci that aligned to the rainbow trout genome with a mapping quality ≥10 and were within 100kb of a gene (Table S4.1.7a). Sixty-three of these genes had GO Slim terms associated with biological processes, the functions of which varied within and between traits. The biological processes represented by the most genes across traits included cell organization and biogenesis, developmental processes, transport, and other biological processes (Table S4.1.7b).

The low annotation rate limited exploration of gene functions for genomic regions where trait-associated loci overlapped. For example, only two of the 12 loci identified as predictors for

both fork length and weight were annotated and met quality thresholds. There were, however, a few regions of interest. The *intersectin 2* (ITSN2) gene, which is involved in cell membrane transport, was only 300 base pairs from two loci associated with return timing and daily growth coefficient that mapped to the same position on Ots08 (Figure 4.2b). The two loci linked to daily growth coefficient on Ots10 (Figure 4.2d) were within and near (13kb) the *sin3A associated protein 130* (SAP130) and *prostaglandin-endoperoxide synthase 2* (PTGS2) genes, respectively, which are involved in transcription and stress response. Another region on Ots10 contained predictors for fork length and weight (Figure 4.2d); one of these loci was within a gene that codes for phosphatase and actin regulator 1, a protein involved in cell regulation. Loci for spawn timing and age at maturity mapped to the same position on Ots02; one of the predictors was within the *ADP ribosylation factor like GTPase 6 interacting protein 5* (ARL6IP5) gene, which is involved in membrane traffic. Lastly, of the two co-mapping loci related to spawn timing on Ots31, one was within the *mitogen-activated protein kinase 15* (MAPK15) gene, which is involved in signal transduction and DNA, RNA, and protein metabolism.

Phenotypic differences between hatchery lines

After multiple generations of rearing, the effect of hatchery line on phenotype was evident for some but not all traits (Table 4.3). In the F_3 generation, four year old fish from the two lines significantly differed in weight at return, spawn timing, and daily growth coefficient. Specifically, fish from the segregated line weighed less (at a given length) than fish from the integrated line, their gonads matured earlier, and they had larger daily growth coefficients (i.e. they lost less weight during holding at the hatchery than fish from the integrated line; Table 4.3). The interaction between line and sex was significant for return timing and daily growth

coefficient, indicating that the effect of hatchery line differed between sexes for these traits. Age at maturity, fork length at return, and the main effect of return timing were not different between the lines (Table 4.3).

Phenotypic differences between the two hatchery lines were also present in other generations but varied (data not shown). Return timing, fork length, and weight significantly differed between the lines in the F_1 generation, while spawn timing differed in the F_2 generation. The interaction of hatchery line and sex was significant for fork length in the F_4 generation.

Effectiveness of managed gene flow and traits affected by domestication selection

1. *Comparisons across all trait-associated loci*

Evaluations of genetic variation at trait-associated loci using PCA showed little evidence of divergence between the hatchery lines across all traits. Each generation of the two lines overlapped extensively in multivariate space at loci associated with traits that showed significant phenotypic differentiation (spawn timing, daily growth coefficient, and weight; Figures S4.2.5 a-c; Table S4.1.10) and those that did not (return timing, age at maturity, and fork length; Figures S4.2.5 d-f; Table S4.1.10).

2. *Overlap with outlier loci and outlier regions*

Comparisons between trait-associated loci and previously identified outlier loci (Waters *et al.* 2015; Waters *et al.* 2017) revealed six loci that overlapped (Table S4.1.8a). Specifically, four loci associated with spawn timing had also been identified as outliers unique to the segregated hatchery line. In addition, one locus linked to return timing on Ots12 was identified as an outlier by *Bayescan* and, in the segregated line, by F_{TEMP} (Figure 4.3). This locus was near

two other outliers and was located within a region that exhibited significant divergence from the P_1 founders across all four generations in the segregated line (Figure 4.3). The region was also significantly divergent in the F_2 generation of the integrated line. Lastly, one unmapped locus was associated with both fork length and weight and was identified as an outlier by *Bayescan* and F_{TEMP} in both hatchery lines.

Regions of interest were also designated where trait-associated loci were in close proximity to outlier loci and regions (Tables S4.1.8b,c). Six regions in the segregated line, for example, exhibited significantly elevated divergence from the P_1 founders in at least one generation and also contained both trait-associated and outlier loci (located on Ots01, two regions on Ots09, Ots11, Ots12, Ots30; Table S8c; Figures S4.2.6 a,c). The integrated line contained three such regions (located on Ots12, Ots15, Ots30; Table S8b; Figures S4.2.6 b,c), with those on Ots12 and Ots30 coinciding with the segregated line. Other chromosomes also contained regions where trait-associated loci and outliers were in close proximity (e.g. black arrows, Figure 4.2), although the overlap was not as extensive.

Gene annotations provided functional insight for some of these loci and regions of overlap (Tables S4.1.7a,b; S4.1.8). One locus on Ots10 that was both an outlier and associated with spawn timing, for example, was located near a second outlier locus and an outlier region in the segregated line (Figure 4.2d). The nearby outlier locus was 800 base pairs from the *transcriptional adaptor 2B* (TADA2B) gene, which aids in transcription (Table S4.1.8d). The region of overlap on Ots01, which displayed elevated divergence in the F_4 SEG generation, contained a locus associated with weight (position 166.6cM; Figure 4.2a) and one outlier locus. The locus linked to weight was within a gene that codes for tyrosine-protein kinase, which has many functions, while the outlier locus was near a gene related to developmental processes, cell

organization and biogenesis, and transport. Loci within the region of overlap on Ots12 (Figure 4.3) were in genes related to signal transduction, cellular protein modification, cell morphogenesis, kinase activity, and developmental processes (Table S4.1.8d; Waters *et al*. 2015). The region on Ots15 (Figure S4.2.6b) contained a locus linked to fork length and two outlier loci. The trait-associated locus was 2kb from the *slit-robo rho GTPase activating protein 1* (SRGAP1) gene, which is related to signal transduction and other biological and metabolic processes. One of the outlier loci was also annotated; it was 13kb from the *member RAS oncogene family* (RAB32) gene, which plays a role in cell organization and biogenesis, transport, signal transduction, and other biological and metabolic processes. Finally, the segment on Ots30 contained a locus associated with fork length, a locus predictive of age at maturity, and a locus that had been identified as an outlier by *Bayescan* and F_{TEMP} in the segregated line (Figure S4.2.6c). The locus associated with fork length was within the *cut like homeobox 1* (CUX1) gene, which is involved in developmental processes, RNA metabolism, and cell organization and biogenesis. No other loci from this region could be annotated.

4.5 DISCUSSION

Here, we aimed to identify genomic regions that influence six fitness-related traits in adult Chinook salmon that are captive-reared but spend part of their life cycle in the natural environment. We then explored the use of trait-associated loci within a management context; namely, whether they could serve as tools for monitoring the effects of two different hatchery management approaches on genetic change in phenotypic traits across four generations. The integrated line uses only wild-born broodstock, and all hatchery-born fish from this line are allowed to spawn in the wild. In contrast, the segregated line uses only hatchery-born

broodstock, and all hatchery-born fish from this line are removed from the river before reproduction. By comparing these lines, our study is one of the first to utilize genomic approaches to determine the effectiveness of a conservation strategy, managed gene flow, on trait-associated – and potentially adaptive – loci.

We detected 226 unique loci associated with the six traits. Some loci were associated with multiple traits while others mapped to shared positions on the genome, results that may be due to the fact that many fitness traits are phenotypically and genetically correlated (Hard 2004; Carlson & Seamons 2008). No evidence for broad scale genetic change at trait-associated loci was observed between the integrated and segregated hatchery lines across four generations using PCA. However, numerous regions were identified where trait-associated loci overlapped with outliers. Many of these overlapping regions, primarily with loci linked to spawn timing and return timing (e.g. Figures 4.2d, 4.3), were either unique to, or more divergent in, the segregated hatchery line. Other regions were present in both hatchery lines (e.g. fork length, Figure S4.2.6c). These results highlight the role that managed gene flow plays in reducing genetic change at loci linked to important phenotypic traits. Continued monitoring of these loci will provide further insights into processes influencing polygenic traits in captive populations.

Detecting trait-associated loci in natural populations

A key challenge in applying genomic-based approaches to conservation is the ability to detect loci linked to polygenic traits in natural populations (Pritchard & Di Rienzo 2010; Olson-Manning *et al.* 2012). Notably, experimental power to detect trait-associated loci may be limited by sample size, marker coverage across the genome, and effect sizes of loci on the traits of interest (Korte & Farlow 2013). While the current study may have been limited by such factors,

the data set analyzed here is typical for natural populations and comparable with those from other recent association studies with an evolutionary emphasis (Brieuc *et al*. 2015; Laporte *et al.* 2015; Pavey *et al.* 2015; Hess *et al.* 2016; Nichols *et al.* 2016). In addition, Random Forest is predicted to perform well in natural populations when linkage disequilibrium between quantitative trait loci and study markers may be reduced (Holliday *et al.* 2012) or sample size is small (Rokach 2016), which are typical conditions for many conservation scenarios.

Power to detect genotype-phenotype associations may also have been reduced by the amount of phenotypic variation within the six traits that were analyzed, as the study was restricted to variability found within a single river system. For example, the standard deviations of return timing and spawn timing across all individuals were 22 and 8 days, respectively (Table S4.1.2b). Similarly, over 75% of the adult Chinook salmon in the study population mature at age four (Knudsen *et al.* 2006), leaving few mature adults of other ages. In contrast, other studies employing Random Forest have identified associations in traits with greater variance (e.g. seasonal salmonid migration timing; Brieuc *et al*. 2015; Hess *et al.* 2016). Incorporating other populations of Chinook salmon for which extensive genetic and phenotypic data exist, though rare, will improve our ability to identify the genetic basis of key traits in the future.

This study is one of the first to use the same genomic dataset to analyze associations with several covarying traits in adult Chinook salmon. In doing so, we observed that the percent phenotypic variation explained by Random Forest was similar across all of the traits (25-32%), despite the fact that the number of predictor loci varied for each trait. The similarity of the results is likely due to the power of the study design. Each tree in Random Forest is grown to its maximal depth, or until a designated stopping criterion is reached (Liaw & Wiener 2002; Goldstein *et al.* 2011). The depth of the trees and the predictive power of the algorithm are

partially dependent on the number of individuals and number of loci analyzed, which were nearly identical across traits. Likewise, the number of individuals analyzed directly influences the calculation of percent variation explained (Liaw & Wiener 2002). Therefore, the similar percentages of variance explained across traits may reflect the size of the study. Nevertheless, we considered the designation of predictor loci as preliminary and placed an emphasis on loci that were verified with additional evidence.

Mapping and annotation of trait-associated loci

Further support for candidate loci identified by Random Forest was obtained by comparing the genome map positions of loci across all fitness-related traits. Specifically, we interpreted sites that contained multiple predictors, including loci that were associated with more than one trait, as candidates for regions underlying fitness. For instance, the 12 loci that were associated with both fork length and weight, traits that are phenotypically (Pearson's r=0.92 for fork length and weight in this study) and genetically correlated in Chinook salmon (Hard 2004; Carlson & Seamons 2008), may impact growth and, in turn, survival. One of these loci was 875 base pairs from a gene involved in microtubule bundle formation, while another locus was 5kb from the *DnaJ heat shock protein family (Hsp40) member C5* (DNAJC5) gene, which regulates the exocytosis of insulin (Lang 1999), a hormone critical for processing glucose. The 13cM region on Ots08 and the 5cM segment on Ots10, which each contained three loci linked to length or weight, may also affect metabolic processes and growth. One locus from the region on Ots10 was within *phosphatase and actin regulator 1* (PHACTR1), which is a key regulator of endothelial cells (Jarray *et al.* 2011) and is also significantly associated with coronary artery calcification in humans (van Setten *et al.* 2013).

Similarly, regions that may be linked to maturation were identified on Ots31, where two loci associated with spawn timing mapped to the same position, and on three other chromosomes (Ots01, Ots02, Ots26), which each had predictor loci for spawn timing and age at maturity within 0.1cM of each other. The two annotated loci from these regions were within the ARL6IP5 and MAPK15 genes on Ots02 and Ots31, respectively. ARL6IP1, a gene that associates with ARL6IP5, affects neurotransmission through Na+-dependent neural glutamate transport activity (Akiduki & Ikemoto 2008). ARL6IP5 also positively regulates the MAPK pathway (Safran *et al.* 2010), to which MAPK15 belongs. The MAPK pathway regulates numerous processes, including cell growth, differentiation, and apoptosis (Qi & Elion 2005), and can be activated by stressful conditions such as heat shock (Dorion & Landry 2002). The association of these two genes with maturation traits in salmon may be related to the transition from marine to freshwater environments that occurs prior to spawning (change in Na+ concentration) and its concomitant stressors (e.g. cessation of feeding, exposure to warmer temperatures, physiological costs associated with upriver migration, development of secondary sex characteristics, and senescence).

Lastly, regions containing multiple loci associated with return timing or daily growth coefficient (e.g. Ots08, Figure 4.2b; 3cM segment on Ots10, Figure 4.2d) may also be involved in the marine to freshwater transition and maturation process. Annotations of these regions identified the ITSN2 (Ots08), SAP130 (Ots10), and PTGS2 (Ots10) genes as potential candidates. ITSN2 regulates cell membrane traffic via clathrin-mediated endocytosis (Pucharcos *et al.* 2000), SAP130 is a transcriptional repressor (Fleischer *et al.* 2003), and PTGS2 expression is related to inflammation, blood vessel formation, estrogen synthesis, and reduced apoptosis (Marshall *et al.* 2005). However, the functional importance of these and other regions requires

further exploration and will improve with the development of additional genomic resources for Chinook salmon.

Comparative analyses across studies

Comparison of our results to those from other studies of Chinook salmon provided additional evidence for certain genomic regions underlying fitness. Notably, a circadian clock gene that was divergent between spring- and fall-migrating populations of Chinook (OmyFbxw11, O'Malley *et al.* 2013) was 2cM from the region on Ots08 in which we identified five predictor loci for spawn timing, length, and weight (Figure 4.2b; Table S4.1.9). Similarly, Brieuc *et al*. (2015) identified two predictors of adult seasonal migration time on Ots12 and Ots17 that were 2.3cM and 0.6cM, respectively, from loci that we linked to spawn timing (Table S4.1.9). A locus that we linked to fork length on Ots10 was in the same 3cM region where Brieuc *et al*. (2015) identified five predictors of migration time. Lastly, two thermotolerance QTL (quantitative trait loci; Everett & Seeb 2014; McKinney *et al.* 2016) mapped to the same positions as loci associated with age at maturity and return timing on Ots11 and Ots27, respectively. Migration and spawn timings in Chinook salmon are known to exhibit strong phenotypic and genetic correlations in some populations (e.g. Quinn *et al.* 2000), and body size may also be correlated with these traits as it can influence access to spawning grounds and breeding success (Schroder *et al.* 2008; Lin *et al.* 2016). The colocation of thermotolerance QTL with loci linked to age and return timing is also supportive, as the migration of spring Chinook salmon is correlated with water temperature (Keefer *et al.* 2008). These regions should be specifically targeted by future investigations that aim to identify the specific genes underlying fitness-related traits.

Traits potentially affected by domestication selection

Although significant phenotypic differences were observed between the integrated and segregated hatchery lines for weight, spawn timing, and daily growth coefficient, no trait exhibited genetic change over time when all associated loci were examined using PCA. This result may be due to the effectiveness of recent hatchery reforms (Mobrand *et al.* 2005; Paquet *et al.* 2011) and the management practices of CESRF (Fast *et al.* 2015), which aim to minimize potential negative ecological and genetic effects of captive rearing. However, null results may also reflect insufficient power to detect genetic divergence at trait-associated loci. Adaptation of complex quantitative traits likely involves selection on standing genetic variation (Fu & Akey 2013; Bernatchez 2016) and is predicted to result in minor allele frequency changes at many loci, or an increased degree of covariance across loci (Pritchard & Di Rienzo 2010; Le Corre & Kremer 2012). Therefore, domestication selection may not produce discernible genetic change across many trait-associated loci, particularly after just four generations of captive rearing.

Yet, specific candidates for where domestication selection may be affecting variation underlying certain traits were identified through comparisons between trait-associated loci and signatures of adaptive divergence from previous studies (Waters *et al.* 2015; Waters *et al.* 2017). Spawn timing, for example, is a fitness trait that exhibited significant phenotypic divergence between the two hatchery lines and also showed the most overlap between trait-associated and outlier loci (four loci). Each of these four overlapping loci were outliers in the segregated line, which is exposed to hatchery conditions every generation, but not in the integrated line, which is only exposed to the hatchery for one generation (Figure 4.2d; Table S4.1.8a). There was also a fifth locus associated with spawn timing on Ots09 that was located within an outlier region in the F_3 and F_4 generations of the segregated line but not in the integrated line (Tables S4.1.8b,c).

Similarly, phenotypic comparisons of return timing revealed a significant interaction of hatchery line and sex (Table 4.3), suggesting that hatchery rearing may be disproportionately affecting this trait in males than females, although the mechanisms of such an effect remain unclear. Genetic comparisons identified one locus linked to return timing on Ots12 that exhibited signals of adaptive divergence by *Bayescan* and, in the segregated line, by F_{TEMP} (Figure 4.3). Notably, divergence of the region containing this locus was consistent across all four generations of the segregated line, compared to just one generation in the integrated line.

It should be noted that the greater levels of overlap observed between trait-associated and outlier loci in the segregated line may, in part, be due to an increased ability to detect outliers in that line, which has a smaller effective population size (Waters *et al.* 2015; Waters *et al.* 2017) and thus potentially higher levels of linkage disequilibrium between study loci and loci under selection. In addition, this study lacks a wild "control" population and thus cannot fully discriminate between processes that occur in the hatchery and those that occur in the natural environment, such as natural selection. Nevertheless, multiple lines of evidence – phenotypic divergence, greater overlap with outliers in the segregated line than in the integrated line, and temporal consistency – suggest that these regions may be responding to domestication selection on return and spawn timing. Our results also support those from other systems, where phenotypic differences in return and spawn timing between wild and hatchery-reared Chinook salmon have been observed (Hoffnagle *et al.* 2008; Williamson *et al.* 2010).

Conclusions and conservation implications

The utility of trait-linked markers in conservation genetics is being actively discussed and explored (Shafer *et al.* 2015; Garner *et al.* 2016; Pearse 2016; Bernatchez *et al.* 2017). For

example, markers associated with key phenotypic traits can provide further insights into the mechanisms by which adaptive variation is maintained in populations (Bernatchez 2016) and may also assist the delineation of conservation (Funk *et al.* 2012; Brieuc *et al.* 2015; Garner *et al.* 2016) and fisheries management units (Bernatchez *et al.* 2017). This study applied trait-associated loci in a novel way – to evaluate the effects of alternative management approaches in captive breeding on genetic variation underlying several fitness traits. Our findings demonstrate the future utility of genomic-based approaches in conservation monitoring.

The identification of loci associated with six key traits by Random Forest is a first step towards characterizing the functional genetic basis of fitness in Chinook salmon (Macqueen *et al.* 2017). The trait-associated loci were supported by genome mapping, gene annotations, and the integration of results across multiple studies. In the future, these loci may contribute to the development of trait-specific tools to monitor genetic change in captive and wild populations, and to better understand the responses of populations to conservation actions and environmental variability (e.g. climate change). The regions where trait-associated and outlier loci overlapped will provide useful starting points for future sequencing efforts that aim to identify the specific genes responding to domestication selection. The observed phenotypic and genomic divergence in certain traits, most notably spawn timing, may also have more immediate impacts on specific management practices at CESRF to further reduce possible effects of captive rearing, since the program is adaptively managed (Fast *et al.* 2015). Lastly, the results support previous work demonstrating the effectiveness of managed gene flow in conservation-focused breeding programs (Waters *et al.* 2015; Waters *et al.* 2017) and will provide additional information that managers can use to assess the relative advantages and disadvantages of different captive rearing approaches.

4.6 ACKNOWLEDGEMENTS

We thank the following individuals for project development, broodstock collection and sampling, and laboratory and analytical assistance: Michael Ford, all Yakama Nation and Washington Department of Fish and Wildlife personnel at the Roza Dam Adult Monitoring Facility and CESRF, Isadora Jimenez-Hidalgo, Daniel Drinan, Kotaro Ono, and Mackenzie Gavery. We also thank everyone who was involved in establishing CESRF and shaping its research direction, including Levi George, Melvin Sampson, Steve Schroder, Craig Busack, past and present members of the Independent Scientific Review Panel, and the Yakama Nation Tribal Council. We are grateful to Krista Nichols for her helpful review of the manuscript. Finally, we thank anonymous reviewers and Associate Editor Maren Wellenreuther for their insightful comments. Funding for this study was provided by Washington Sea Grant (Award R/HCE-4 to K.A.N) and the Hall Conservation Genetics Research Award from the University of Washington (to C.D.W.).

4.7 TABLES

Table 4.1 – Individual traits measured in adult Chinook salmon returning to the upper Yakima River. Fish are measured at the Roza Dam Adult Monitoring Facility (RAMF) and again if used as broodstock at the Cle Elum Supplementation and Research Facility (CESRF). Measurements from RAMF were used in the analyses for fork length and weight.

Trait Category	Traits (units)	Locations Measured	
Life history	age at maturity (years)	RAMF	CESRF
	return timing (day of year)	RAMF	
	spawn timing (day of year)		CESRF
Morphometric	forklength (cm)	RAMF	CESRF
	weight (kg)	RAMF	CESRF
Growth	daily growth coefficient (no units)	RAMF to CESRF	

Table 4.2 – Results of Random Forest association analyses for six phenotypic traits. For each trait, the number of individuals analyzed, the total number of predictor loci identified by the backward purging approach and the number of those that were mapped, and the percent trait variation explained by the predictor loci are given. * denotes the out-of-bag classification error rate for age at maturity, rather than variation explained, since the trait was analyzed using classification trees. Here, classification error rate refers to the percentage of hold-out samples whose ages were misclassified using the tree.

Trait	**Individuals**	**Predictor loci (mapped)**	**Percent variation explained**
age at maturity	383	30 (20)	24.9*
return timing	383	26 (15)	29.2
spawn timing	379	68 (36)	26.7
fork length	380	44 (20)	32.0
weight	381	37 (15)	31.5
daily growth coefficient	372	35 (20)	31.8

Table 4.3 – Sample sizes from each hatchery line, regression coefficients (standard errors) of terms, and test statistics with *p*-values based on linear or generalized linear models for each phenotypic trait. Significant *p*-values are in bold font. The reference level for the intercept is integrated females. Due to the form of the models, coefficients for age at maturity and weight were exponentiated (except for $\beta_{\text{fork length}}$) and refer to the proportionate response compared to the reference level. $\beta_{\text{fork length}}$ is the allometric coefficient.

	age at maturity (years)	**return timing (day of year)**	**fork length (cm)**	**weight (kg)**	**spawn timing (day of year)**	**DGC**
N_{INT}	1135	984	984	984	60	60
N_{SEG}	247	200	200	200	52	48
$\beta_{Intercept}$	4.00 (0.078) *Z*=71.39 ***p*<0.001**	156.57 (0.77) *t*=204.42 ***p*<0.001**	72.34 (0.17) *t*=425.42 ***p*<0.001**	$1.28e^{-05}$ ($1.98e^{-06}$) t=-77.85 ***p*<0.001**	265.75 (1.01) *t*=263.78 ***p*<0.001**	-0.05 (0.003) *t*=-16.49 ***p*<0.001**
β_{Line}	1.00 (0.047) *Z*=0.00 *p*=1.000	3.35 (1.86) *t*=1.80 *p*=0.072	-0.70 (0.41) *t*=-1.70 *p*=0.089	0.98 (0.01) *t*=-2.54 ***p*=0.011**	-9.26 (1.36) *t*=-6.81 ***p*<0.001**	0.01 (0.004) *t*=2.59 ***p*=0.011**
β_{Sex}	0.92 (0.028) *Z*=-2.56 ***p*=0.011**	-0.87 (1.33) *t*=-0.65 *p*=0.514	0.44 (0.30) *t*=1.50 *p*=0.135	0.99 (0.005) *t*=-1.49 *p*=0.138	-1.46 (1.47) *t*=-0.99 *p*=0.323	-0.03 (0.004) *t*=-5.87 ***p*<0.001**
$\beta_{Line*Sex}$	0.98 (0.071) *Z*=-0.31 *p*=0.760	7.34 (3.25) *t*=2.25 ***p*=0.024**	0.74 (0.72) *t*=1.02 *p*=0.31	0.98 (0.01) *t*=-1.42 *p*=0.157	-0.79 (2.35) *t*=-0.34 *p*=0.736	0.02 (0.007) *t*=2.24 ***p*=0.027**
$\beta_{fork\ length}$	N/A	N/A	N/A	2.98 (0.03) *t*=88.02 ***p*<0.001**	N/A	N/A

4.8 FIGURES

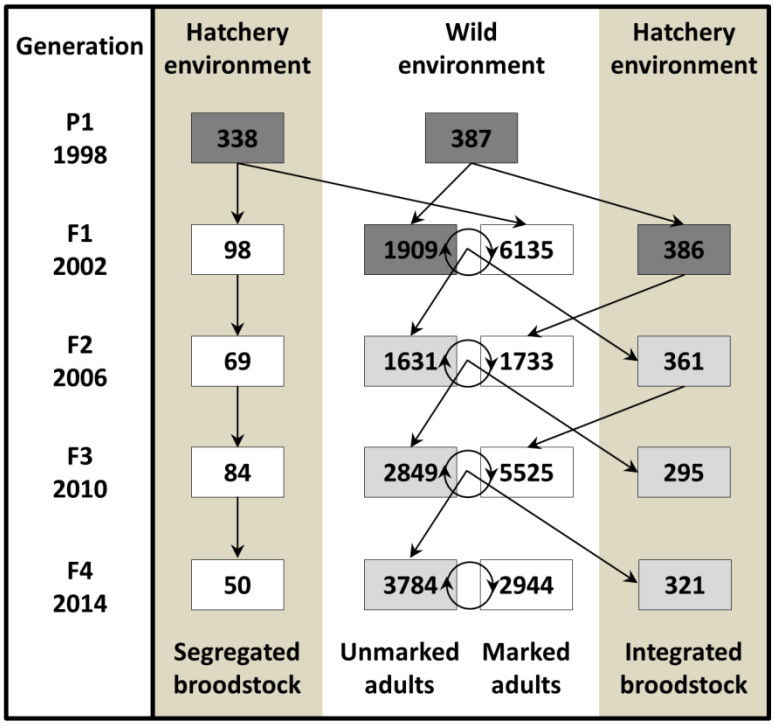

Figure 4.1 – Schematic illustrating the initiation (the founding P_1 generation) and subsequent propagation (F_1-F_4 generations) of the integrated and segregated hatchery lines of anadromous Chinook salmon at the Cle Elum Supplementation and Research Facility (modified from Waters *et al*. 2015). Each box denotes the number of spawners (wild environment) and the number of broodstock (hatchery environment) for each year surveyed. Linear arrows indicate the contribution of wild spawners or hatchery broodstock to the subsequent generation. Circular arrows represent unobserved mating between wild-born (unmarked) and hatchery-born (marked) spawners in the wild environment. Fish from the two lines are differentially marked, so only hatchery-born fish from the integrated line are permitted to spawn in the wild. Dark gray boxes represent wild adults, light gray boxes represent natural origin adults with hatchery, wild, or hybrid ancestry, and white boxes represent adults born in the hatchery.

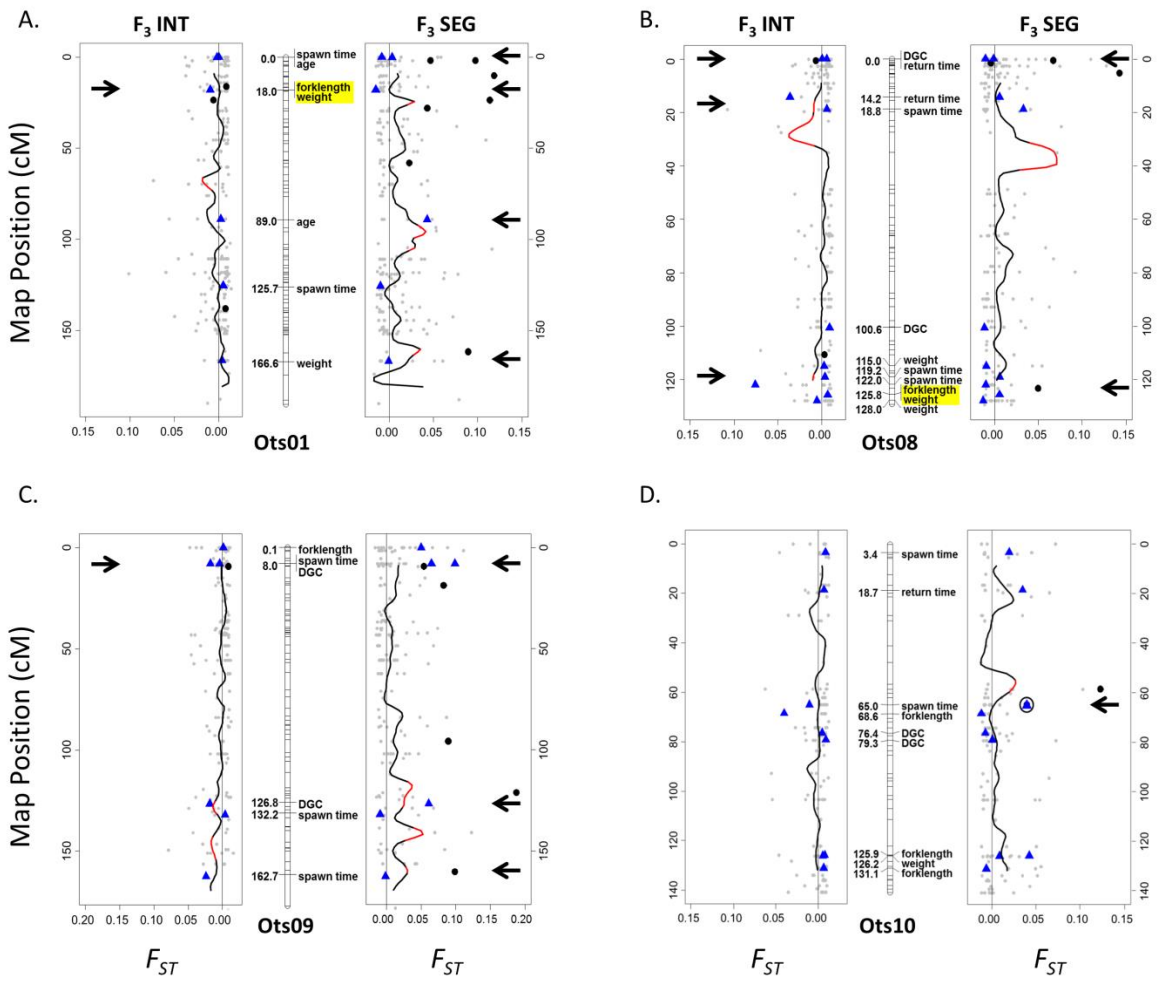

Figure 4.2 – Graphical representation of four Chinook salmon chromosomes (center panels, a-d) showing the map positions (cM) of loci associated with six fitness-related traits, as identified by Random Forest analyses. Loci associated with different traits mapped to the same regions, including loci on Ots01 and Ots08 that were associated with both fork length and weight (highlighted in yellow). Divergence (F_{ST}) of the F_3 INT and F_3 SEG hatchery lines when compared to the P_1 founders is displayed in the left and right panels of each figure, respectively. The F_3 generation is shown because it is the most recent hatchery generation for which there are relatively large sample sizes (>50), and thus has greater statistical support for all outlier tests. The black line denotes the moving average of F_{ST} across the chromosome, with regions

exhibiting significant levels of divergence (i.e. outlier regions) from the P_1 Founders in red (Waters *et al.* 2015; Waters *et al.* 2017). The centromere of each chromosome is shaded with diagonal black lines. Black circles represent outlier loci previously identified by F_{TEMP} and *Bayescan*, blue triangles correspond to trait-associated loci, and gray points are all other study loci. Locations where trait-associated loci are in close proximity to outlier loci or regions are marked with black arrows, including one outlier locus on Ots10 that was also associated with spawn timing (circled).

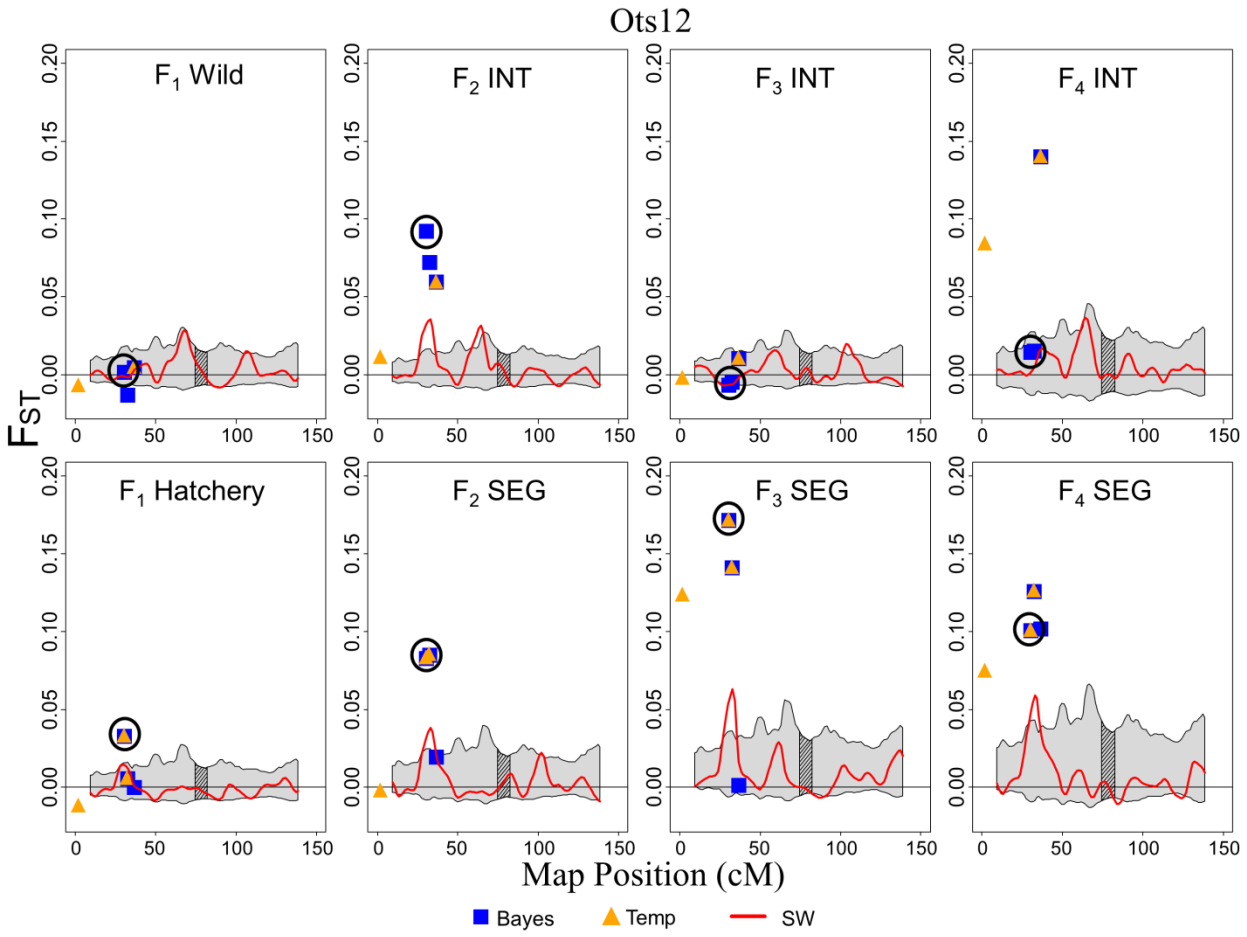

Figure 4.3 – Loci and regions of chromosome Ots12 showing signatures of adaptive divergence based on measures of pairwise F_{ST} between each generation of each line and the P_1 founders. The results are given for the integrated (top panel) and segregated (bottom panel) hatchery lines through the F_1, F_2, F_3, and F_4 generations. Blue squares are loci that were identified as outliers by *Bayescan* (Foll & Gaggiotti 2008) and orange triangles are outliers identified by F_{TEMP}, a method designed to detect selection in a single population over time (Therkildsen *et al.* 2013). The red line represents the kernel smoothed moving average of F_{ST} and the gray shaded area is the 95% confidence interval. Genomic regions exhibiting significant levels of divergence (i.e. outlier regions) from the P_1 founders occur where the moving average of F_{ST} exceeds the 95% confidence intervals. The centromere of the chromosome is shaded with diagonal black lines.

The black circle designates a locus predictive of return timing, Ot005185_Ots12p, which was also identified as an outlier by *Bayescan* and, in the segregated line, by F_{TEMP}. Negative F_{ST} values occur due to finite sample sizes and slight variance in sample sizes between populations (Weir & Cockerham 1984).

4.9 SUPPLEMENTARY MATERIAL

S4.1 – Supplementary tables containing metadata and results of Chapter 4

S4.1.1 – Numbers of adults used as broodstock, the number of adults sampled in this study, the final sample size retained after filtering, the observed and expected heterozygosity for each population, and the number of loci that deviated from Hardy-Weinberg equilibrium.

S4.1.2 – (a) Phenotypic data for all adult Chinook returning to the upper Yakima River in 2010. (b) Phenotypic data for adult Chinook salmon sequenced in this study.

S4.1.3 – Lists of the 9108 RAD loci and 465 individuals analyzed in this study.

S4.1.4 – Genotypes at the 9108 RAD loci for each individual.

S4.1.5 – Results from inferring genomic positions of unmapped study loci.

S4.1.6 – Predictor loci identified by Random Forest for each phenotypic trait.

S4.1.7 – (a) Gene annotation results for all trait-associated loci. (b) Summary of GO Slim terms associated with biological processes for genes near trait-associated loci.

S4.1.8 – (a) Comparison of trait-associated loci with previously-identified outlier loci. (b) Overlap of regions of elevated divergence in the integrated line, trait-associated loci, and outlier loci. (c) Overlap of regions of elevated divergence in the segregated line, trait-associated loci, and outlier loci. (d) Annotation results of outlier loci.

S4.1.9 – Comparison of trait-associated loci with trait-associated loci identified in other studies of Chinook salmon.

S4.1.10 – Loadings of trait-associated loci on the first two PCs from a principal components analysis for each phenotypic trait.

S4.1.11 – Sequencing barcode and RAD lane identifier for each fish analyzed in this study.

S4.2 – Supplementary information for results of Chapter 4

S4.2.1 – Imputation of missing genotypes.

S4.2.2 – Inferring positions of unmapped loci.

S4.2.3 – Predictor loci identified by Random Forest for each phenotypic trait.

S4.2.4 – Gene annotation methods.

S4.2.5 – Plots of principal components analyses comparing the integrated and segregated hatchery lines at predictor loci for each phenotypic trait.

S4.2.6 – Plots showing overlap between trait-associated loci and outlier loci and genomic regions.

4.10 REFERENCES

Akiduki S, Ikemoto MJ (2008) Modulation of the neural glutamate transporter eaac1 by the addicsin-interacting protein arl6ip1. *Journal of Biological Chemistry*, **283**, 31323-31332.

Allendorf FW, Thorgaard GH (1984) Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary genetics of fishes* (ed. Turner BJ). Plenum Press, New York.

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

Aykanat T, Lindqvist M, Pritchard VL, Primmer CR (2016) From population genomics to conservation and management: A workflow for targeted analysis of markers identified using genome-wide approaches in atlantic salmon *salmo salar*. *Journal of Fish Biology*, **89**, 2658-2679.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid snp discovery and genetic mapping using sequenced rad markers. *PLoS ONE*, **3**, e3376.

Barson NJ, Aykanat T, Hindar K *et al.* (2015) Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, **528**, 405-408.

Baskett ML, Waples RS (2013) Evaluating alternative strategies for minimizing unintended fitness consequences of cultured individuals on wild populations. *Conservation Biology*, **27**, 83-94.

Bateson ZW, Hammerly SC, Johnson JA, Morrow ME, Whittingham LA, Dunn PO (2016) Specific alleles at immune genes, rather than genome-wide heterozygosity, are related to immunity and survival in the critically endangered attwater's prairie-chicken. *Molecular Ecology*, **25**, 4730-4744.

Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L (2015) Rad genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the american lobster (homarus americanus). *Molecular Ecology*, **24**, 3299-3315.

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

Bernatchez L, Wellenreuther M, Cristián A *et al.* (2017) Harnessing the power of genomics to secure the future of seafood. *Trends in Ecology and Evolution*, **32**, 665-680.

Berthelot C, Brunet F, Chalopin D *et al.* (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, **5**, 10.

Breiman L (2001) Random forests. *Machine Learning*, **45**, 5-32.

Brieuc MSO, Ono K, Drinan DP, 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.

Busack C, Knudsen CM, Hart G, Huffman P (2007) Morphological differences between adult wild and first-generation hatchery upper yakima river spring chinook salmon. *Transactions of the American Fisheries Society*, **136**, 1076-1087.

Carlson SM, Seamons TR (2008) A review of quantitative genetic components of fitness in salmonids: Implications for adaptation to future change. *Evolutionary Applications*, **1**, 222-238.
Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: An analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124-3140.
Cho CY (1992) Feeding systems for rainbow trout and other salmonids with reference to current estimates of energy and protein requirements. *Aquaculture*, **100**, 107-123.
Christie MR, Ford MJ, Blouin MS (2014) On the reproductive success of early-generation hatchery fish in the wild. *Evolutionary Applications*, **7**, 883-896.
Clutter RI, Whitesel LE (1956) *Collection and interpretation of sockeye salmon scales* International Pacific Salmon Fisheries Commission.
Conde DA, Flesness N, Colchero F, Jones OR, Scheuerlein A (2011) An emerging role of zoos to conserve biodiversity. *Science*, **331**, 1390-1391.
Dorion S, Landry J (2002) Activation of the mitogen-activated protein kinase pathways by heat shock. *Cell Stress & Chaperones*, **7**, 200-206.
Duchesne P, Bernatchez L (2002) An analytical investigation of the dynamics of inbreeding in multi-generation supportive breeding. *Conservation Genetics*, **3**, 47-60.
Dupont-Nivet M, Karahan-Nomm B, Vergnet A *et al.* (2010) Genotype by environment interactions for growth in european seabass (*dicentrarchus labrax*) are large when growth rate rather than weight is considered. *Aquaculture*, **306**, 365-368.
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.
Fast DE, Bosch WJ, Johnston MV *et al.* (2015) A synthesis of findings from an integrated hatchery program after three generations of spawning in the natural environment. *North American Journal of Aquaculture*, **77**, 377-395.
Fleischer TC, Yun UJ, Ayer DE (2003) Identification and characterization of three new components of the msin3a corepressor complex. *Molecular and Cellular Biology*, **23**, 3456-3467.
Fleming IA (1996) Reproductive strategies of atlantic salmon: Ecology and evolution. *Reviews in Fish Biology and Fisheries*, **6**, 379-416.
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.
Ford MJ (2002) Selection in captivity during supportive breeding may reduce fitness in the wild. *Conservation Biology*, **16**, 815-825.
Ford MJ, Fuss H, Boelts B, LaHood E, Hard J, Miller J (2006) Changes in run timing and natural smolt production in a naturally spawning coho salmon (*Oncorhynchus kisutch*) population after 60 years of intensive hatchery supplementation. *Canadian Journal of Fisheries and Aquatic Sciences*, **63**, 2343-2355.
Frankham R (2008) Genetic adaptation to captivity in species conservation programs. *Molecular Ecology*, **17**, 325-333.
Fritts AL, Scott JL, Pearsons TN (2007) The effects of domestication on the relative vulnerability of hatchery and wild origin spring chinook salmon (*Oncorhynchus tshawytscha*) to predation. *Canadian Journal of Fisheries and Aquatic Sciences*, **64**, 813-818.

Fu WQ, Akey JM (2013) Selection and adaptation in the human genome. In: *Annual review of genomics and human genetics, vol 14* (eds. Chakravarti A, Green E), pp. 467-489. Annual Reviews, Palo Alto.

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

Garner BA, Hand BK, Amish SJ *et al.* (2016) Genomics in conservation: Case studies and bridging the gap between data and application. *Trends in Ecology & Evolution*, **31**, 81-83.

Goldstein BA, Polley EC, Briggs FBS (2011) Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, **10**, 35.

Griffiths RA, Pavajeau L (2008) Captive breeding, reintroduction, and the conservation of amphibians. *Conservation Biology*, **22**, 852-861.

Hard JJ (2004) Evolution of chinook salmon life history under size-selective harvest. In: *Evolution illuminated: Salmon and their relatives* (eds. Hendry A, Stearns S), pp. 315-337. Oxford University Press.

Harrisson KA, Pavlova A, Telonis-Scott M, Sunnucks P (2014) Using genomics to characterize evolutionary potential for conservation of wild populations. *Evolutionary Applications*, **7**, 1008-1025.

Hess JE, Zendt JS, Matala AR, Narum SR (2016) Genetic basis of adult migration timing in anadromous steelhead discovered through multivariate association testing. *Proceedings of the Royal Society B-Biological Sciences*, **283**, 10.

Hoffmann A, Griffin P, Dillon S *et al.* (2015) A framework for incorporating evolutionary genomics into biodiversity conservation and management. *Climate Change Responses*, **2**, 1.

Hoffnagle TL, Carmichael RW, Frenyea KA, Keniry PJ (2008) Run timing, spawn timing, and spawning distribution of hatchery- and natural-origin spring chinook salmon in the imnaha river, oregon. *North American Journal of Fisheries Management*, **28**, 148-164.

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**, 23.

Holliday JA, Wang TL, Aitken S (2012) Predicting adaptive phenotypes from multilocus genotypes in sitka spruce (*picea sitchensis*) using random forest. *G3-Genes Genomes Genetics*, **2**, 1085-1093.

Horne JS, Hervert JJ, Woodruff SP, Mills LS (2016) Evaluating the benefit of captive breeding and reintroductions to endangered sonoran pronghorn. *Biological Conservation*, **196**, 133-146.

Hornoy B, Pavy N, Gerardi S, Beaulieu J, Bousquet J (2015) Genetic adaptation to climate in white spruce involves small to moderate allele frequency shifts in functionally diverse genes. *Genome Biology and Evolution*, **7**, 3269-3285.

Jarray R, Allain B, Borriello L *et al.* (2011) Depletion of the novel protein phactr-1 from human endothelial cells abolishes tube formation and induces cell death receptor apoptosis. *Biochimie*, **93**, 1668-1675.

Jule KR, Leaver LA, Lea SEG (2008) The effects of captive experience on reintroduction survival in carnivores: A review and analysis. *Biological Conservation*, **141**, 355-363.

Keefer ML, Peery CA, Caudill CC (2008) Migration timing of columbia river spring chinook salmon: Effects of temperature, river discharge, anti ocean environment. *Transactions of the American Fisheries Society*, **137**, 1120-1133.

Knudsen CM, Schroder SL, Busack CA *et al.* (2006) Comparison of life history traits between first-generation hatchery and wild upper yakima river spring chinook salmon. *Transactions of the American Fisheries Society*, **135**, 1130-1144.

Kodama M, Hard JJ, Naish KA (2012) Temporal variation in selection on body length and date of return in a wild population of coho salmon, *oncorhynchus kisutch*. *Bmc Evolutionary Biology*, **12**, 12.

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with gwas: A review. *Plant Methods*, **9**, 9.

Landa A, Flagstad O, Areskoug V *et al.* (2017) The endangered arctic fox in norway-the failure and success of captive breeding and reintroduction. *Polar Research*, **36**, 14.

Lang JC (1999) Molecular mechanisms and regulation of insulin exocytosis as a paradigm of endocrine secretion. *European Journal of Biochemistry*, **259**, 3-17.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. *Nature Methods*, **9**, 357-U354.

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**, 10.

Laporte M, Pavey SA, Rougeux C *et al.* (2016) Rad sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in north atlantic eels. *Molecular Ecology*, **25**, 219-237.

Laporte M, Rogers SM, Dion-Cote AM *et al.* (2015) Rad-qtl mapping reveals both genome-level parallelism and different genetic architecture underlying the evolution of body shape in lake whitefish (*coregonus clupeaformis*) species pairs. *G3-Genes Genomes Genetics*, **5**, 1481-1491.

Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, **21**, 1548-1566.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078-2079.

Liaw A, Wiener M (2002) Classification and regression by randomforest. *R News*, **2**, 18-22.

Lien S, Koop BF, Sandve SR *et al.* (2016) The atlantic salmon genome provides insights into rediploidization. *Nature*, **533**, 200-+.

Lin JE, Hard JJ, Naish KA, Peterson D, Hilborn R, Hauser L (2016) It's a bear market: Evolutionary and ecological effects of predation on two wild sockeye salmon populations. *Heredity*, **116**, 447-457.

Lynch M, O'Hely M (2001) Captive breeding and the genetic fitness of natural populations. *Conservation Genetics*, **2**, 363-378.

Macqueen DJ, Primmer CR, Houston RD *et al.* (2017) Functional annotation of all salmonid genomes (faasg): An international initiative supporting future salmonid research, conservation and aquaculture. *Bmc Genomics*, **18**.

Marshall SF, Bernstein L, Anton-Culver H *et al.* (2005) Nonsteroidal anti-inflammatory drug use and breast cancer risk by stage and hormone receptor status. *Journal of the National Cancer Institute*, **97**, 805-812.

McGinnity P, Prodohl P, Ferguson K *et al.* (2003) Fitness reduction and potential extinction of wild populations of atlantic salmon, *salmo salar*, as a result of interactions with escaped farm salmon. *Proceedings of the Royal Society B-Biological Sciences*, **270**, 2443-2450.
McKinney GJ, Seeb LW, Larson WA *et al.* (2016) An integrated linkage map reveals candidate genes underlying adaptive variation in chinook salmon (*oncorhynchus tshawytscha*). *Molecular Ecology Resources*, **16**, 769-783.
Mobrand LE, Barr J, Blankenship L *et al.* (2005) Hatchery reform in washington state: Principles and emerging issues. *Fisheries*, **30**, 11-23.
Narum SR, Gallardo P, Correa C *et al.* (2017) Genomic patterns of diversity and divergence of two introduced salmonid species in patagonia, south america. *Evolutionary Applications*, **10**, 402-416.
National_Research_Council (1996) *Upstream: Salmon and society in the pacific northwest* National Academy Press, Washington, D.C.
Nichols KM, Kozfkay CC, Narum SR (2016) Genomic signatures among *oncorhynchus nerka* ecotypes to inform conservation and management of endangered sockeye salmon. *Evolutionary Applications*, **9**, 1285-1300.
O'Malley KG, Jacobson DP, Kurth R, Dill AJ, Banks MA (2013) Adaptive genetic markers discriminate migratory runs of chinook salmon (*oncorhynchus tshawytscha*) amid continued gene flow. *Evolutionary Applications*, **6**, 1184-1194.
Olson-Manning CF, Wagner MR, Mitchell-Olds T (2012) Adaptive evolution: Evaluating empirical support for theoretical predictions. *Nature Reviews Genetics*, **13**, 867-877.
Paquet PJ, Flagg T, Appleby A *et al.* (2011) Hatcheries, conservation, and sustainable fisheries-achieving multiple goals: Results of the hatchery scientific review group's columbia river basin review. *Fisheries*, **36**, 547-561.
Pavey SA, Gaudin J, Normandeau E *et al.* (2015) Rad sequencing highlights polygenic discrimination of habitat ecotypes in the panmictic american eel. *Current Biology*, **25**, 1666-1671.
Pearse DE (2016) Saving the spandrels? Adaptive genomic variation in conservation and fisheries management. *Journal of Fish Biology*, **89**, 2697-2716.
Pritchard JK, Di Rienzo A (2010) Adaptation - not by sweeps alone. *Nature Reviews Genetics*, **11**, 665-667.
Pucharcos C, Estivill X, de la Luna S (2000) Intersectin 2, a new multimodular protein involved in clathrin-mediated endocytosis. *Febs Letters*, **478**, 43-51.
Qi MS, Elion EA (2005) Map kinase pathways. *Journal of Cell Science*, **118**, 3569-3572.
Quinlan AR, Hall IM (2010) Bedtools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
Quinn TP, Unwin MJ, Kinnison MT (2000) Evolution of temporal isolation in the wild: Genetic divergence in timing of migration and breeding by introduced chinook salmon populations. *Evolution*, **54**, 1372-1385.
R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org, Vienna, Austria.
Raymond M, Rousset F (1995) Genepop (version 1.2) - population-genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248-249.
Rokach L (2016) Decision forest: Twenty years of research. *Information Fusion*, **27**, 111-125.
Safran M, Dalah I, Alexander J *et al.* (2010) Genecards version 3: The human gene integrator. *Database-the Journal of Biological Databases and Curation*, 16.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, **78**, 629-644.
Schroder SL, Knudsen CM, Pearsons TN *et al.* (2010) Behavior and breeding success of wild and first-generation hatchery male spring chinook salmon spawning in an artificial stream. *Transactions of the American Fisheries Society*, **139**, 989-1003.
Schroder SL, Knudsen CM, Pearsons TN *et al.* (2008) Breeding success of wild and first-generation hatchery female spring chinook salmon spawning in an artificial stream. *Transactions of the American Fisheries Society*, **137**, 1475-1489.
Shafer ABA, Wolf JBW, Alves PC *et al.* (2015) Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, **30**, 78-87.
Stephan J, Stegle O, Beyer A (2015) A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, **6**, 10.
Sutherland BJG, Gosselin T, Normandeau E *et al.* (2016) Salmonid chromosome evolution as revealed by a novel method for comparing radseq linkage maps. *Genome Biology and Evolution*, **8**, 3600-3617.
Therkildsen NO, Hemmer-Hansen J, Als TD *et al.* (2013) Microevolution in time and space: Snp analysis of historical DNA reveals dynamic signatures of selection in atlantic cod. *Molecular Ecology*, **22**, 2424-2440.
Thorpe JE, Miles MS, Keay DS (1984) Developmental rate, fecundity and egg size in atlantic salmon, *salmo salar*. *Aquaculture*, **43**, 289-305.
Thorson JT (2015) Spatio-temporal variation in fish condition is not consistently explained by density, temperature, or season for california current groundfishes. *Marine Ecology Progress Series*, **526**, 101-112.
van Setten J, Isgum I, Smolonska J *et al.* (2013) Genome-wide association study of coronary and aortic calcification implicates risk loci for coronary artery disease and myocardial infarction. *Atherosclerosis*, **228**, 400-405.
Waters CD, Hard JJ, Brieuc MSO *et al.* (2015) Effectiveness of managed gene flow in reducing genetic divergence associated with captive breeding. *Evolutionary Applications*, **8**, 956-971.
Waters CD, Hard JJ, Brieuc MSO *et al.* (2017) What can genomics tell us about the success of enhancement programs in anadromous chinook salmon? A comparative analysis across four generations. Supplementary material for bernatchez et al. (2017) harnessing the power of genomics to secure the future of seafood. *Trends in Ecology & Evolution*, 665-680.
Weir BS, Cockerham CC (1984) Estimating f-statistics for the analysis of population structure. *Evolution*, **38**, 1358-1370.
Williamson KS, Murdoch AR, Pearsons TN, Ward EJ, Ford MJ (2010) Factors influencing the relative fitness of hatchery and wild spring chinook salmon (*oncorhynchus tshawytscha*) in the wenatchee river, washington, USA. *Canadian Journal of Fisheries and Aquatic Sciences*, **67**, 1840-1851.
Zhao Y, Chen F, Zhai RH *et al.* (2012) Correction for population stratification in random forest analysis. *International Journal of Epidemiology*, **41**, 1798-1806.

Chapter 5. A practical introduction to Random Forest for genetic association studies in ecology and evolution[4]

5.1 ABSTRACT

Large genomic studies are becoming increasingly common with advances in sequencing technology, and our ability to understand how genomic variation influences phenotypic variation between individuals has never been greater. The exploration of such relationships first requires the identification of associations between molecular markers and phenotypes. Here we explore the use of Random Forest (RF), a powerful machine learning algorithm, in genomic studies to discern loci underlying both discrete and quantitative traits, particularly when studying wild or non-model organisms. RF is becoming increasingly used in ecological and population genetics because, unlike traditional methods, it can efficiently analyze thousands of loci simultaneously and account for non-additive interactions. However, understanding both the power and limitations of Random Forest is important for its proper implementation and the interpretation of results. We therefore provide a practical introduction to the algorithm and its use for identifying associations between molecular markers and phenotypes, discussing such topics as data limitations, algorithm initiation and optimization, as well as interpretation. We also provide short R tutorials as examples, with the aim of providing a guide to the implementation of the algorithm. Topics discussed here are intended to serve as an entry point for molecular ecologists interested in employing Random Forest to identify trait associations in genomic data sets.

[4] This chapter has been published as A practical introduction to Random Forest for genetic association studies in ecology and evolution, Brieuc MSO*, Waters CD*, Drinan DP, and Naish KA (2018*), Molecular Ecology Resources* **00**, 1-12. *These authors contributed equally.

5.2 INTRODUCTION

Identifying the genetic basis of phenotypic traits and the specific genes underlying adaptation can improve our understanding of processes that shape variation in natural populations (Savolainen *et al.* 2013; Barson *et al.* 2015; Brieuc *et al.* 2015). Such information, in turn, can assist conservation and management efforts (Funk *et al.* 2012; Shafer *et al.* 2015; Bernatchez *et al.* 2017) and help predict the responses of populations to future environmental variability. However, many phenotypic traits are polygenic (Roff 2007; Le Corre & Kremer 2012; Santure *et al.* 2013; Silva *et al.* 2017). Genome-wide association studies (GWAS) that survey polymorphisms across the genome are often used to determine the genetic architecture underlying these traits. A typical approach to perform such association studies is to use linear mixed effects models to detect single-locus associations between traits and molecular markers. However, polygenic traits may be influenced by non-additive interactions (Yang *et al.* 2010; Mackay 2014) and adaptation of these traits through selection on standing genetic variation is predicted to result in minor allele frequency changes at many loci, or an increased degree of covariance across loci (Pritchard & Di Rienzo 2010; Le Corre & Kremer 2012). One way to increase the power of association studies to detect many interacting loci is to extend linear mixed effects approaches to multivariate models that accommodate all loci as explanatory variables simultaneously to calculate genome-wide breeding values, an approach known as genomic selection (GS; Jannink *et al.* 2010). However, while conventional GS analyses provide an estimate of the percentage trait variance explained, they do not identify the best predictors of a trait (Goddard & Hayes 2007; Desta & Ortiz 2014; Barabaschi *et al.* 2016), and they rarely accommodate non-additive epistatic effects (Stephan *et al.* 2015). Further, they have greater power to detect associations in populations with higher linkage disequilibrium (LD), such as

agricultural populations, but they may have less power in populations typical of evolutionary studies where LD may decay rapidly (Holliday *et al.* 2012). Finally, the utility of markers detected using GS diminishes as populations become more divergent (Desta & Ortiz 2014; Barabaschi *et al.* 2016), possibly because they reflect pedigree-based breeding values within specific populations. Therefore, alternative analytical approaches that survey multiple loci simultaneously for trait associations, incorporate all interactions between loci, and identify specific predictors of a trait relevant for the identification of causal variants, are necessary for detecting complex genetic architectures across a broader range of populations.

Random Forest (RF; Breiman 2001), a machine learning algorithm, is a promising method to identify loci underlying polygenic traits. The algorithm can analyze large genomic data sets, which typically comprise many more markers than samples. RF also provides a non-parametric framework that can account for dominance and epistatic interactions, enabling the identification of suites of loci that explain substantial phenotypic variation collectively but may not display significant changes in allele frequencies individually. Importantly, the method may prove useful in identifying relevant trait-associated markers that segregate in natural populations, where linkage disequilibrium between markers and traits may be lower, and where there are fewer genomic resources available (Holliday *et al.* 2012).

Studies utilizing RF for genetic analyses in evolutionary studies have increased in recent years. As examples, RF was used to investigate epistatic interactions between loci underlying budset timing and cold hardiness in Sitka Spruce (Holliday *et al.* 2012). Similarly, loci underlying genetic adaptation to temperature and precipitation in white spruce trees were detected using RF (Hornoy *et al.* 2015). Brieuc *et al.* (2015) combined RF with F_{ST} outlier analyses to examine patterns of adaptive divergence of a key life history trait – run timing –

between populations of Chinook salmon. In the same species, Waters *et al*. (2018) used trait-linked markers, detected using RF, to investigate the efficacy of gene flow in reducing domestication selection in a population enhancement program. Markers linked to American eel ecotypes were used to demonstrate repeated within-generation selection in a panmictic species (Pavey *et al.* 2015), and loci discriminating eels in polluted and unpolluted environments were shown to covary with specific contaminants (Laporte *et al.* 2016). RF and GWAS based on mixed effects models can also serve complementary purposes, often identifying loci in common (Holliday *et al.* 2010; Holliday *et al.* 2012; Hess *et al.* 2016; Nichols *et al.* 2016). RF has also found a role in related studies, such as identifying cryptic diversity in cross-species phylogeographic studies (Espindola *et al.* 2016), tracing invasion routes (Fraimout *et al.* 2017), and identifying marker panels for population assignment (Sylvester *et al.* 2017). Given these myriad applications of RF, it is not surprising that numerous methods exist by which the algorithm can be run, and the results analyzed and interpreted. This range of possibilities, coupled with the popularity and relative ease of running RF, may contribute to inconsistent and potentially flawed results if the data and algorithm are not treated appropriately.

Here, we combine lessons learned from our own analyses with information from other studies to provide a simple, introductory guide to facilitate the use of RF to identify genotype-phenotype associations in non-model organisms. This paper and the accompanying tutorials are intended to introduce users in ecological and evolutionary genomics to considerations in data treatment, approaches, and possible pitfalls when conducing RF, and therefore serve as a starting point for deeper investigations, such as estimating the extent and magnitude of non-linear marker effects (e.g. Holliday *et al*. 2012) and monitoring processes that influence polygenic traits in natural populations. We take as inspiration the paper by Wilson *et al*. (2010), which introduced

ecologists to the Animal Model in quantitative genetic studies. More complete descriptions of the theory and implementation of the RF algorithm have been published elsewhere (e.g. Breiman 2001; Breiman 2002; Goldstein *et al.* 2011; Chen & Ishwaran 2012; Rokach 2016). For the sake of brevity, we focus on the implementation of the original Random Forest algorithm using the *randomForest* package (Liaw & Wiener 2002) in R (R Core Team 2017) and its associated parameters, although we encourage the exploration of other programs. We first provide an overview of the Random Forest algorithm. Steps are then described to prepare data for RF, including initial data exploration and the identification of important covariates and possible confounding factors. We next provide guidance on the initiation of RF and the optimization of the algorithm parameters for classification and regression. Last, we summarize methods for interpreting the results of RF and identifying trait-associated, or predictor, loci. The suggested workflow is presented in Figure 5.1. However, analyses of genomic data sets are evolving. Therefore, we remain neutral on "best practices," but rather point the reader to potential avenues to explore in the implementation of RF in their own studies.

5.3 ALGORITHM OVERVIEW

Random Forest was first introduced by Breiman (2001), extending previous research on random decision trees (Ho 1995) and bagging of predictors (Breiman 1996). Briefly, Random Forest builds a series, or "forest," of classification or regression trees (CART; Table 5.1) that recursively partitions a group of explanatory variables in order to predict a categorical response, such as disease state (Figure 5.2a), or the value of a continuous response variable, such as body weight. The algorithm first randomly samples, with replacement, a subset of samples to form the training data set (*sampsize* parameter; Table 5.2), from which the tree is "grown." RF then randomly selects a subset of predictors (*mtry* parameter; Table 5.2) and searches for the predictor

that best partitions the response variable in the training data set. Partitioning is based on identifying the predictor that minimizes either the within group variance (categorical response variable) or error when regressed against the response (continuous response variable). The optimal predictor becomes the first node in the tree and splits the data (Figures 5.2a,b). The algorithm continues to randomly select a subset of predictors at each node and then partition the data until a full tree is grown (Figure 5.2b). The random selection of samples and predictors reduces bias and variance within and between trees, improving the predictive power (Goldstein *et al.* 2011; Rokach 2016). For classification trees based on categorical data, samples not included in the training data (out-of-bag [OOB] data; Table 5.1) are then classified using the tree, and the misclassification of those individuals provides an estimate of the predictive error rate (OOB-ER; Table 5.1). The power of a regression tree based on continuous variables is determined by the proportion of variation of the response variable explained by the tree (proportion variation explained [PVE]; Table 5.1). The importance of each predictor is estimated as the change in predictive power of the tree after the predictor is randomly permuted amongst individuals. The predictive power of a tree will be reduced when an important predictor is permuted, while those with little effect on the response variable will not change tree accuracy. Many trees are grown to build a forest (*ntree* parameter; Table 5.2); the importance values of predictors are averaged across trees and analyzed to ultimately determine those that best explain variation in the response variable.

5.4 DATA EXPLORATION

Type of data

Random Forest is flexible in the types of data that can be used as response and predictor variables. Response variables can be either continuous or categorical, and no practical limitation

exists for the number of levels allowed within a categorical response (e.g. Liaw & Wiener (2002) simulated up to 10,000 unique categories in the *randomForest* package in R). However, the efficacy of the algorithm may be limited or biased if there are few observations within categories or unequal numbers of observations between categories. Predictor variables can also be either continuous or categorical. However, in the *randomForest* package, categorical variables are limited to 53 or fewer levels, an arbitrary threshold likely set for computational efficiency. In theory, this means that all but extremely polymorphic genetic markers are suitable for use in RF, including single nucleotide polymorphisms (Brieuc *et al.* 2015), microsatellites (Del Carpio *et al.* 2011), and amplified fragment length polymorphisms (Del Carpio *et al.* 2011).

In addition, ecological and environmental data can be analyzed by RF in conjunction with genetic analyses (Mager *et al.* 2014), or even included as complementary predictors to genetic data when their inclusion makes biological sense. However, combining environmental and genetic data may result in predictors with different variances (if continuous) or numbers of classes (if categorical). Such imbalances have been shown to produce spurious results, because variables with higher variances or more categories will appear to be better at splitting the data and thus could have upwardly biased importance values (Japkowicz & Stephen 2002; Strobl *et al.* 2007). Therefore, caution must be taken when including environmental and genetic predictors in the same analysis, and imbalances should be corrected (Japkowicz & Stephen, 2002; Strobl *et al.* 2007).

Missing values

While nearly every type of molecular marker can be used in Random Forest, the algorithm does not allow missing values for the predictors or the response variable. However,

there are methods to impute missing values for the predictors. One convenient tool is *rfImpute*, a built-in imputation function within the *randomForest* package (Liaw & Wiener 2002) in R (R Core Team 2017), that uses RF to impute missing values (see proximity analysis in Tang & Ishwaran 2017). For continuous predictors, this function replaces missing values by a weighted average of the observations for that predictor (Breiman & Cutler 2007). Missing values for categorical predictors are replaced with the category from the observation that is most similar to it across all other predictors. Although the accuracy of this method was assessed for ecological data (Breiman 2002), its performance has not been evaluated for genetic data. There are, however, several imputation methods specifically designed for genetic data that can be used, such as *fastPhase* (Scheet & Stephens 2006) or *Beagle* (Browning & Browning 2007), which rely on haplotypes of various lengths for imputation, rather than on genotype frequencies at a single locus.

Confounding effects

Random Forest, like traditional GWAS approaches, can be confounded by genetic structure between populations and relatedness within populations (Zhao *et al.* 2012; Stephan *et al.* 2015). For example, if two populations differ phenotypically and genetically, then analyses may identify correlations reflecting this stratification, rather than true genotype-phenotype associations. Analyses could also be confounded by imbalances within the response and explanatory variables (e.g. sample sizes or variances), high levels of linkage disequilibrium between study loci, and other factors such as repeated phenotypic measurements for individuals or samples from different years or those of different age classes. Confounding factors should be identified prior to conducting RF, as there are various methods to account for their effects while implementing the algorithm.

5.5 INITIATION OF ANALYSIS

Program selection

Although we have chosen to focus mainly on the *randomForest* package in R, the original Random Forest algorithm can be implemented using a number of other programming languages, such as Fortran (Breiman 2002) and MATLAB (The Mathworks, Inc.), and software packages (Table S5.1.1). In addition, modifications of the algorithm have been developed for various purposes, including different methods of tree construction that aim to improve accuracy (e.g. Geurts *et al.* 2006; Bernard *et al.* 2012), extensions that explicitly account for population structure and other genetic effects (Stephan *et al.* 2015), and variations of RF designed for specific types of response data (e.g. right-censored survival data; Ishwaran *et al.* 2008). We encourage users to explore these other implementations and choose the one that best suits their data.

Correction for confounding effects

1. *Population stratification and relatedness*

Various methods to account for correlations between individuals (e.g. population structure and relatedness) and reduce false positive results in association studies have been developed (Devlin & Roeder 1999; Price *et al.* 2006; Yu *et al.* 2006; Zhang *et al.* 2010; Zhao *et al.* 2012; Stephan *et al.* 2015). One approach specifically developed for Random Forest analyses was introduced by Holliday *et al*. (2012) and extended by Zhao *et al*. (2012). This two-step approach first entails regressing (one at a time) the phenotypes and the genotypes at each locus against a measure of population or family membership, such as the coordinates of individuals on axes from a principal components analysis, using linear or generalized linear models. The

residuals of the models, representing the "adjusted" phenotypes and genotypes, are then analyzed by Random Forest. This method is easy to implement, as it relies on common regression techniques, and can be extended to account for other possible confounding factors, such as age and sex (e.g. Waters *et al*. 2018). Furthermore, its effectiveness in accounting for population stratification has been verified by both Zhao *et al*. (2012) and our own simulations (Table S5.1.2). However, correcting for population structure may also reduce statistical power of association and other genomic analyses (Marchini *et al.* 2004; Lotterhos & Whitlock 2015). That said, newer, more flexible methods are being developed, including correction through best linear unbiased prediction or covariance eigenvectors (Azevedo *et al.* 2017 and references therein), and are worth further investigation.

Mixed Random Forest is another method that was recently developed to account for both population structure and relatedness (Stephan *et al.* 2015). This approach is an extension of linear mixed model techniques that have been used for traditional genome-wide association analyses (e.g. Yu *et al.* 2006; Zhang *et al.* 2010), which rely on the inclusion of a random effect term directly in the model to account for correlations between individuals. Stephan *et al*. (2015) modified the original Random Forest algorithm to incorporate a random effect term at each node in the tree, thus eliminating the need to correct for confounding effects prior to conducting Random Forest. While this single-step approach may have increased statistical power to detect genotype-phenotype associations (Stephan *et al.* 2015), implementation of Mixed Random Forest requires knowledge of the Python programming language.

2. *Unbalanced design: impacts on classification based approaches*

Data sets in which unequal numbers of observations exist across phenotypic classes are common and may arise from a variety of sources, including the experimental design or through inherent properties of the experimental population, such as rare versus common phenotypes. This unbalanced design is not a major issue in regression-based RF analyses because the phenotypes are continuous. However, in classification-based RF analyses, unbalanced data can greatly affect performance (Japkowicz & Stephen 2002; Strobl *et al.* 2007; Galar *et al.* 2012). For example, the RF algorithm aims at minimizing the overall classification error rate, rather than the error rate for each phenotypic class. When some classes of categorical phenotypes are under-represented, RF may preferentially assign individuals to the majority phenotype, resulting in a low classification error rate overall but (potentially) a high error rate in the minority class (Table 5.3). Similarly, variable importance measures may be inaccurate if the predictors vary in the number of discrete classes (e.g. the number of alleles) or in their numeric scale (e.g. after correction for population structure; Strobl *et al.* 2007), because variables with more classes or larger variances could appear to be better at splitting the data. Such imbalances must be accounted for when conducting RF, as they can lead to inaccurate conclusions regarding influential predictors.

Problems arising from unbalanced data can be overcome using algorithm- and data-based approaches (Lin & Chen 2013). While algorithm-based approaches aim at modifying the classification algorithm itself, data-based approaches use sampling methods to correct for unbalanced design with no modification of the Random Forest algorithm, and can therefore be easily implemented. Stratified sampling, or over-sampling the minority class and under-sampling the majority classes (Chen *et al.* 2004), is one data-based strategy that has proven effective at correcting for class imbalance, provided the imbalance is not too severe (Blagus & Lusa 2010). Stratified sampling can be implemented using the *strata* and *sampsize* options when running the

randomForest function (full details in *randomForest* R package manual; Liaw & Wiener 2002). The s*trata* option specifies the variable that is unbalanced, which may be the response or a variable that is not directly incorporated in the RF analysis. For example, in situations where a categorical phenotype is not corrected for underlying family structure, *strata* could be used to ensure equal representation of families. *Sampsize* is then used to specify the sample size for each class of *strata* variable. Generally, *sampsize* is not set to more than 2/3 of the smallest phenotypic class to ensure that this class is not under-represented in the validation (OOB) data set. However, the power of the Random Forest algorithm could be affected if the sample size of the smallest phenotypic class is too low. The efficacy of stratified sampling can be assessed by comparing the OOB-ER for the minority class (or classes) before and after using the method. If the correction for the imbalance is successful, the OOB-ER for the class should be reduced. It should also be noted that if more than one variable must be considered in order to balance the data, an aggregated variable can be manually created to capture the full diversity and stratification.

3. Linkage disequilibrium among study loci

Linkage disequilibrium between predictors (i.e. study loci) may be problematic for RF analyses because it can downwardly bias the importance values for predictors that are truly associated with a trait (Meng *et al.* 2009; Goldstein *et al.* 2011). As previously discussed, the importance value of each predictor is estimated as the change in predictive power of a tree after the predictor is randomly permuted among the out-of-bag samples. The predictive power of a tree will be reduced when an important predictor is permuted, while those with little effect on the response variable will not change tree accuracy. For two loci that are tightly linked and included

in the same tree, the predictive power of the tree will not be significantly affected when one of the predictors is permuted because the values of the second predictor remain unchanged. Therefore, the importance of a truly associated predictor may be reduced if it is linked with one or more other predictors, and Random Forest may fail to detect the association. Methods to account for LD in Random Forest analyses have been developed (Meng *et al*. 2009) and include pruning the data set for LD prior to RF, using haplotypes as predictors instead of individual SNPs, and modifying the calculation of variable importance measures.

4. *Repeated phenotypic measures*

Repeated phenotypic measurements of individuals, which may occur in long-term studies of wild populations, must be considered prior to association analyses for the same reason as population structure and relatedness: the measurements are not independent. If utilized properly, however, repeated measures can improve the power of association analyses. For example, Rönnegård *et al*. (2016) introduced a method for GWAS that uses a linear mixed effects model to account for repeated measures of individuals and corresponding environmental variation that may affect the phenotype. Notably, their method had higher power to detect QTL than an existing method where the average phenotype across repeated measures is used as the response, particularly in scenarios where the phenotype varied substantially across years. The use of Random Forest for ecological and evolutionary studies in natural populations is still relatively new, and – to our knowledge – no similar method has been developed to utilize repeated measurements in RF. The Mixed Random Forest method of Stephan *et al*. (2015) may be suitable for repeated measures since it incorporates a random effect term to account for correlations between observations. However, such an application has yet to be evaluated, although we encourage its exploration.

5.6 OPTIMIZATION

Like most other analyses, the Random Forest algorithm has limitations that can lead to spurious results if not appropriately considered and optimized (Huang & Boutros 2016). The main objectives when optimizing RF are to 1) minimize the OOB error rate (for analyses with a categorical response) or maximize the proportion variation explained (PVE, for analyses with a continuous response) and to 2) ensure the results are accurate and repeatable. Optimization of RF can be achieved by adjusting the parameters of the algorithm.

Optimizing algorithm parameters

The two parameters that have the most influence on the OOB-ER or PVE are the number of trees grown (*ntree*) per forest and the number of predictors to randomly sample at each node (*mtry*; Goldstein *et al.* 2010). Increasing the values of *ntree* and *mtry* will usually improve the accuracy of Random Forest until the OOB-ER or PVE reaches a plateau (Figure 5.3; Goldstein *et al.* 2010). Once a plateau is reached, increasing values of *ntree* and *mtry* do not improve predictive power but instead become computationally costly. Therefore, the aim should be to minimize these parameters while maximizing the accuracy of the algorithm.

Identification of the optimal values for *ntree* and *mtry* should be conducted simultaneously (Goldstein *et al.* 2010). As an example, Brieuc *et al*. (2015) sampled 414 individuals representing 14 populations of anadromous Chinook salmon from the Columbia River and Puget Sound and genotyped them at 9107 loci. A Random Forest analysis was then conducted to identify loci associated with adult return to freshwater (run timing). Here, we ran several iterations of RF with varying values of *mtry* and *ntree* using the data set from Brieuc *et*

al. (2015). Similar to Goldstein *et al*. (2010), values of *mtry* ranged from $\sqrt{p}$ to p ($\sqrt{p}$, $2\sqrt{p}$, $0.1p$, $0.2p$, $p/3$, p), where p is the number of predictors (loci), and *ntree* varied initially from 10 to 1000. In this data set, accuracy was maximized for *ntree* = 400 and *mtry* = $0.2p$ (Figure 5.3). Increasing *ntree* and *mtry* above these values did not increase the accuracy of the regression. It should be noted that Figure 5.3 illustrates optimization for the data set from Brieuc *et al*. (2015), and that such plots will vary across studies, as optimal parameter values depend on many factors, including the number of loci, number of samples, and the genetic architecture of the response variable.

Achieving repeatability

Although very important, minimizing the OOB-ER or maximizing the PVE should not be the only goal of optimization. It is also essential to ensure repeatability between forests, specifically with regard to importance values of the predictors, so that genotype-phenotype associations are reliably identified. For example, we showed for the data set from Brieuc *et al*. (2015) that convergence of the PVE was obtained for *ntree* = 400 and *mtry* = $0.2p$ (Figure 5.3). However, with those parameter settings, the importance values for the predictors varied substantially between forests (correlation of 0.2). In fact, convergence of importance values (with a correlation > 0.8) was only achieved after approximately 10,000 trees (Figure 5.4). This discrepancy highlights the significance of verifying the plateau of the OOB-ER or PVE as well as the convergence of the importance values between Random Forest runs.

5.7 INTERPRETATION

Overview of marker selection approaches

Accurately identifying a set of trait-linked markers is a main goal of association studies. Unlike typical GWAS methods, Random Forest does not rely on significance values (i.e. *p*-values) that are associated with individual predictors. The algorithm instead assigns an importance value to each predictor, the actual significance of which is not known. There have been many approaches proposed for identifying a set of predictors that are reliably trait-linked, all of which have advantages and limitations, whether biological, statistical, or practical.

1. *Cut-off approaches and thresholds.*

A simple approach is to use a cutoff for the number of markers deemed important (e.g. Botta *et al.* 2014) or rely on a threshold importance value above which markers might be considered important. In such instances, it may be preferable to examine relative importance values between markers rather than the absolute importance values. This is because the absolute importance value for a locus is not only influenced by the locus itself, but by all the other loci included in the analysis, as well as other parameters such as the number of trees. Accordingly, the "elbow method" has also been proposed as a way to identify a set of important markers. In this method, the cutoff for the importance values is determined from the "elbow" on the plot of the important values, where the differences between importance values start to decrease (Figure 5.5; Goldstein *et al.* 2010; Laporte *et al.* 2016). Cut-off approaches are advantageous because they are easily implemented. However, even in the absence of associations between predictors and the response, these methods will identify "important" loci, as some predictors will always be ranked higher than others. Therefore, one could argue that cut-off methods might be subjective (Goldstein *et al.* 2011) and lack statistical or biological justification.

2. *Recurrent relative variable importance measure (r2VIM).*

A method for variable selection that relies on the variable importance measure (VIM) of each locus relative to the observed minimal VIM across repetitive runs of RF has recently been proposed (r2VIM; Szymczak *et al.* 2016). In brief, several iterations of Random Forest are conducted. Loci with a large VIM across all runs, relative to the lowest VIM in each run (ratio >1), are considered important, as these VIM values exceed those expected from the distribution of uninformative loci. This method can be easily implemented and has been shown to minimize the number of false positive associations (Szymczak *et al.* 2016). However, special attention should be paid to the number of trees grown per forest to ensure that reliable variable importance measures are obtained (i.e. enough trees are grown to achieve convergence).

3. *p- values for predictors.*

Calculating *p*-values for each predictor through a permutation process has been proposed (see the *rfPermute* (Archer 2016) or *pRF* (Chakravarthy 2015) packages in R). This method has the advantage of using a statistical cut-off familiar to most scientists. However, one drawback is that a *p*-value is calculated for each locus individually but not for a set of loci. The method therefore does not account for interactions between predictors during the estimation process.

4. *Purging approaches.*

Several variations of purging approaches have been developed (e.g. Jiang *et al.* 2004; Diaz-Uriarte & de Andres 2006), and each involves iteratively removing non-informative markers from data sets to identify a group of loci that best predicts the response variable. Here,

we present a generalization of the backward purging approach that was developed by Holliday *et al*. (2012) and used in Brieuc *et al*. (2015) and Waters *et al*. (2018). First, an initial Random Forest is run with the entire data set to estimate importance values of all loci. Then, the performance (measured using OOB-ER or PVE) of forests from various subsets of the most important loci (e.g. top 1%, top 5%, top 10% etc.) is estimated in order to determine a candidate group of loci that minimizes the OOB-ER or maximizes the PVE (Figure 5.6a). Backward purging is then conducted on an expanded group of candidate loci (e.g. use top 1.5% of loci if the top 1% of loci appeared to have the highest PVE) to account for any uncertainty in the ranking of the loci. For example, in Brieuc *et al*. (2015), the top 70 loci appeared to maximize the PVE in the initial Random Forest runs (Figure 5.6a), so backward purging was initiated with the best 150 loci (Figure 5.6b). Each step of the backward purging approach consists of conducting multiple Random Forest runs and recording the average performance of the forests (OOB-ER or PVE). The locus with the lowest average importance value is then removed, and the process is repeated iteratively until there are only two loci left. The best set of loci, for which the OOB-ER is minimized or the PVE is maximized, is then identified and deemed to be predictive of the phenotype (Figure 5.6b). One advantage of the backward purging approach is that the marker selection process does not rely on a subjective decision from the user, such as a threshold importance value. However, the method may require extended computational time if purging includes a large number of loci and many trees per forest (e.g. a complete RF analysis that included the generalized purging approach described here for 400 individuals, 9000 loci, and a continuous trait took approximately three weeks to run on a desktop computer with 24 GB of RAM and a processor speed of 3.47 GHz). In addition, backward purging may increase the risk of overfitting (e.g. Jiang *et al.* 2004; Svetnik *et al.* 2004; Diaz-Uriarte & de Andres 2006), which

occurs when RF fits noise within the training data set. The risk of overfitting increases with the number of loci included in association analyses (Wray *et al.* 2013), and it may lead to the inclusion of false positive associations within the final set of predictor loci and a reduction in predictive accuracy of the model for new samples. Tests to check for overfitting should be conducted to quantify and minimize these risks. This could be done by performing a cross validation analysis (similar to Roberts & Hamann 2012), where a fraction of the data set (training data, usually 70-90% of the samples) is used to conduct RF and the remaining data (test data, usually 10-30% of samples that are omitted from the RF analysis entirely) are subsequently classified by the RF model to assess its performance (for an example, see Brieuc *et al*. 2015). This validation step should be repeated multiple times and can be performed manually or by using available packages, such as the *caret* (Kuhn 2008) or *rfUtilities* (Evans *et al.* 2011) packages in R. Poor classification accuracy of the test data by the RF model may be interpreted as evidence of overfitting. One possible way to reduce overfitting would be to prune the number of markers prior to conducting a final RF analysis, based on importance values from an initial RF analysis or a relaxed cutoff from a traditional GWAS analysis (J. Holliday, *pers. comm.*). However, the risk of overfitting and the effectiveness of possible solutions likely vary for every study design (e.g. Figure S5.1.3).

5.8 CONCLUSION

The utility of Random Forest to characterize the genetic basis of polygenic traits and address a broad range of ecological and evolutionary questions has been increasingly recognized in recent years. The approach can incorporate thousands of loci simultaneously in a non-parametric framework and thus serves as a powerful alternative to traditional GWAS methods,

many of which test each marker individually or are unable to account for non-additive epistatic interactions. Yet, the myriad applications of Random Forest, along with its accessibility to users, have led to extensive variation in how the algorithm is run, and the results analyzed and interpreted. For this reason, we developed a simple guide that aims to inform potential users of key factors that should be considered before, during, and after conducting Random Forest analyses. The guide is not intended to be comprehensive, but rather introduces a variety of practices that can be subsequently investigated by each user. We also provide tutorials for conducting regression-based and classification-based analyses using the *randomForest* package in R (Liaw & Wiener 2002; R Core Team 2017) to illustrate some of the recommendations discussed in the text. Lastly, we hope this guide contributes to the reporting of accurate and reproducible results and facilitates the use of Random Forest in future ecological and evolutionary studies. We advocate the full reporting of decisions made in data standardization, optimization, and interpretation in studies that implement Random Forest, so that future users can explore the utility and outcomes of this promising approach in ecological and evolutionary studies.

5.9 ACKNOWLEDGEMENTS

This work was funded by a grant from Washington Sea Grant (Award R/HCE-4), and by the Agriculture and Food Research Initiative competitive grant no. 2012-67015-19960 of the USDA National Institute of Food and Agriculture (both to K.A.N). We thank Kotaro Ono for very helpful comments. We also thank the Editor, Anna Santure, and Jason Holliday and two anonymous referees for their very useful suggestions during the review process.

5.10 TABLES

Table 5.1 – Glossary of terms commonly used in Random Forest analyses.

Term	Description
Backward purging	A method developed by Holliday *et al*. (2012) to identify a subset of loci that either minimizes the out-of-bag error rate (classification analyses of categorical response variables) or maximizes the proportion of variation explained (regression analyses of continuous variables). Backward purging is performed in two steps. First, a Random Forest analysis is performed with all data to order predictors based on their importance. Second, loci are removed in an iterative process based on their importance value (removing the least important) until a final set remains that minimizes the out-of-bag error rate (classification analyses) or maximizes the variation explained (regression analyses).
Classification and regression trees (CART)	A general term describing the use of decision trees to identify explanatory variables that predict either categorical or continuous response variables.
Out-of-bag (OOB)	Samples not included in the training data set (a hold-out set). Out-of-bag samples are used to estimate the predictive power of each tree within the forest.
Out-of-bag error rate (OOB-ER)	An estimate of the misclassification of out-of-bag samples when using a classification Random Forest model.
Proportion variation explained (PVE)	Proportion of variance in the out-of-bag response variable that is explained by the regression-based Random Forest model.
Training data set	A bootstrapped subset of samples (generally 2/3 of samples) used to create a decision tree.
Variable importance	A measure of the influence of each predictor on the response variable, estimated by permuting predictor values and calculating the change in predictive power of the entire tree.

Table 5.2 – Description of key parameters utilized by the *randomForest* function in R

Parameter	Description
mtry	Number of predictors to be randomly sampled at each node in a tree, used to search for the predictor that best partitions samples in the training data set (defaults = square root of the number of predictors for classification trees ($\sqrt{p}$), and number of predictors/3 for regression trees).
ntree	Number of trees "grown" in the forest.
sampsize	Number of samples to be randomly drawn, with replacement, for the training data set (default = 2/3 of samples). For classification trees, this parameter can be used in conjunction with *strata* to ensure equal representation of all strata in unbalanced experimental designs.
strata	Used to define a stratified variable (a variable that has unbalanced representation in the data set, such as a response variable or a variable that influences the study design).

Table 5.3 – Example of low overall OOB-ER, but high OOB-ER for the minority class that may arise when imbalances exist between phenotypic classes (A and B). Here, the overall OOB-ER is relatively low (0.15), despite the fact that the OOB-ER for the minority class (B) is high (0.80).

	Predicted A	**Predicted B**	**OOB-ER per class**
True A (n=100)	98	2	0.02
True B (n=20)	16	4	0.80
Overall OOB-ER			0.15

5.11 FIGURES

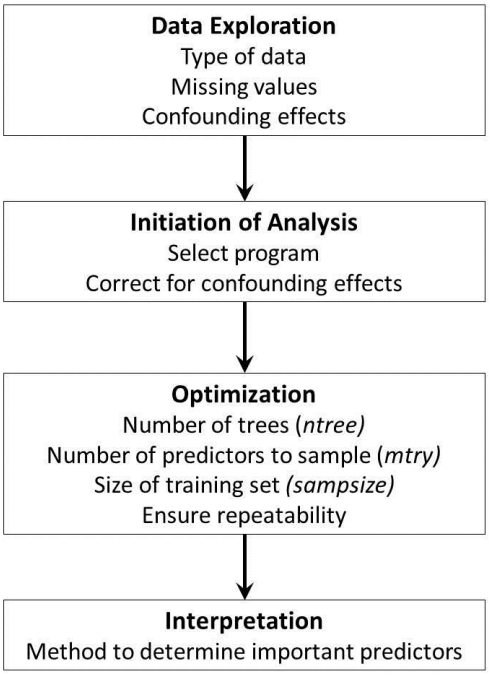

Figure 5.1 – Key steps (bold) and analytical considerations at each step for conducting Random Forest analyses.

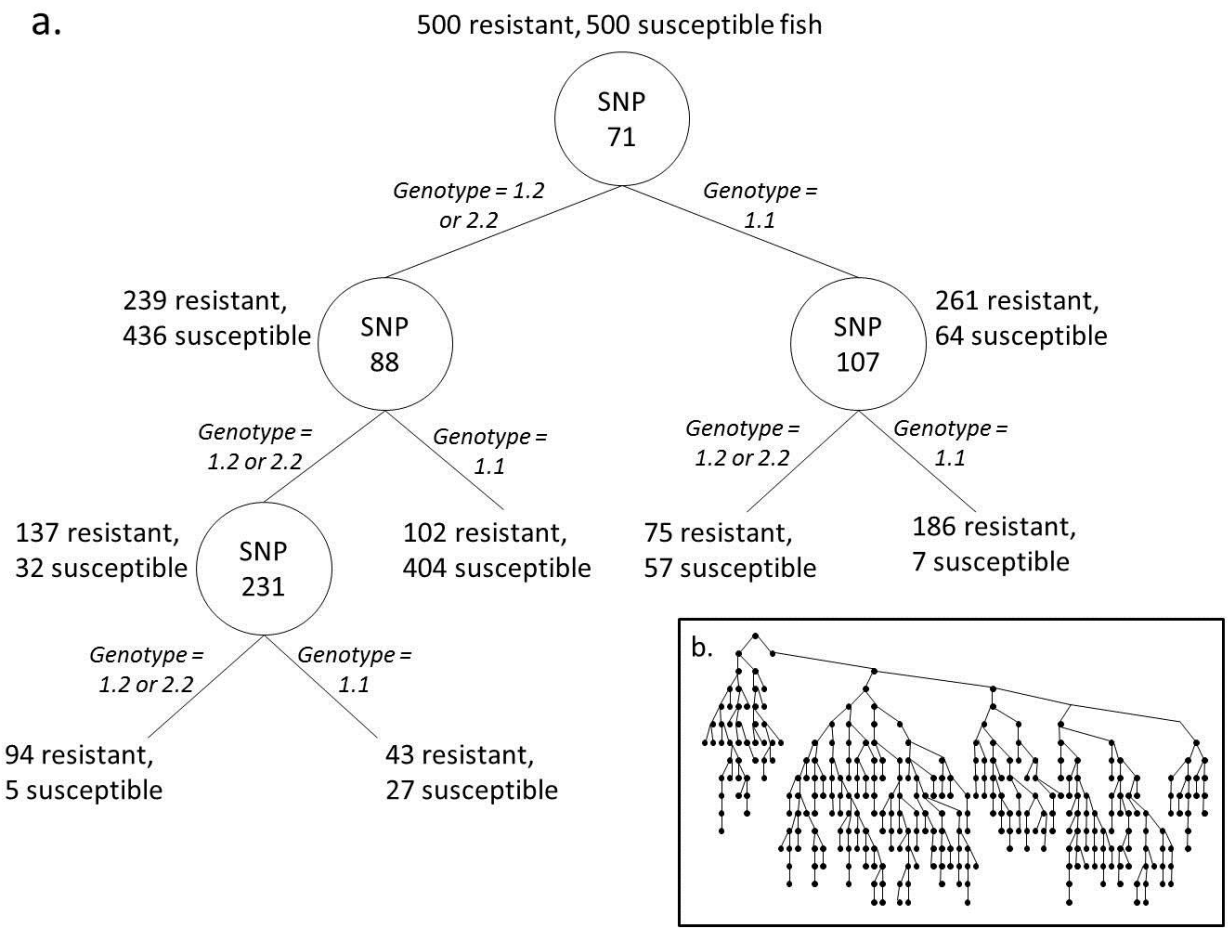

Figure 5.2 – a) Diagram illustrating the first few nodes of a hypothetical classification-based Random Forest tree to identify loci associated with a binary trait, disease resistance (survival or mortality; figure modified from Cordell 2009). First, a subset of loci (*mtry* parameter) is chosen from the data, and the locus that best partitions the resistant and susceptible individuals (here, SNP 71) becomes the first node of the tree. The algorithm continues to partition the resistant and susceptible individuals based on their genotypes until a node contains individuals of one phenotype, or if a pre-determined stopping criterion is reached. Note that this tree is only partially grown for illustrative purposes. b) Example of a small but fully grown Random Forest tree, where each point represents a node that splits the data.

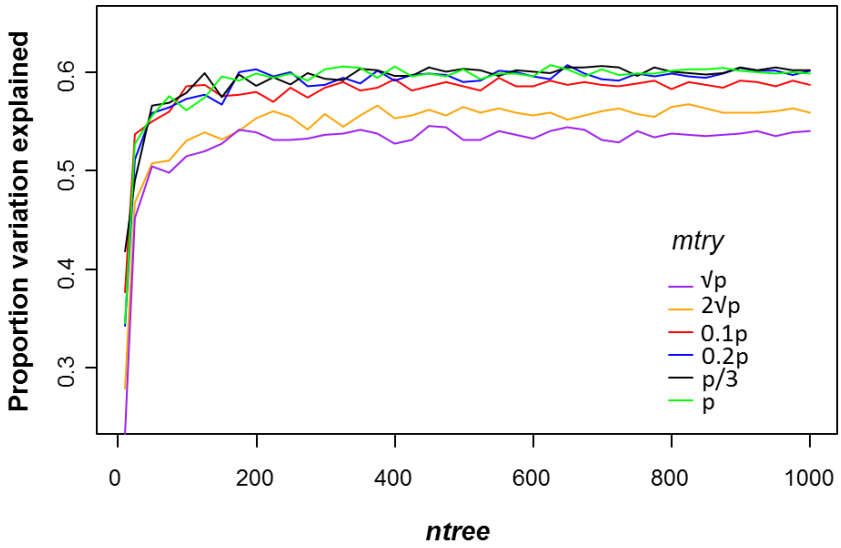

Figure 5.3 – Convergence of the proportion of variation explained (PVE) across different *mtry* values with increasing *ntree* using the data set from Brieuc *et al*. (2015). Here, PVE is maximized for *ntree*=400 and *mtry* = 0.2p.

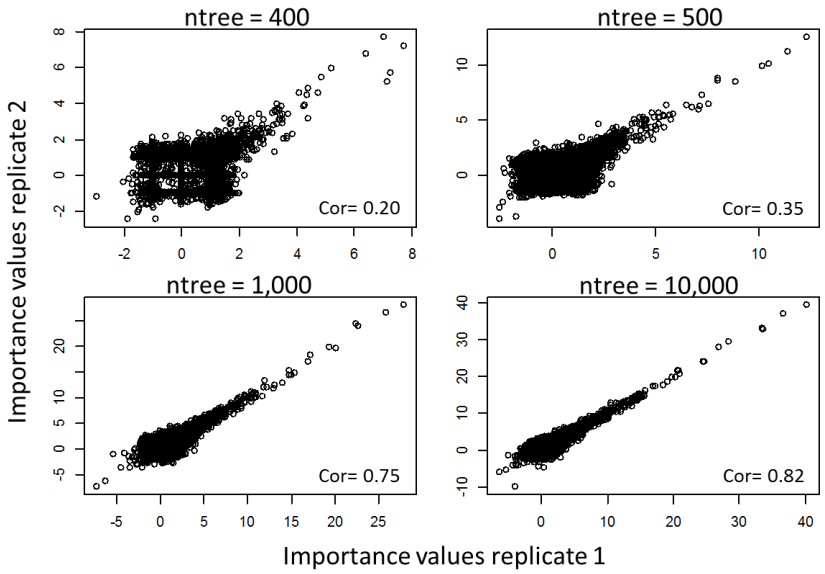

Figure 5.4 – Correlation of importance values between replicate Random Forest runs with increasing number of trees using the data set from Brieuc *et al*. (2015).

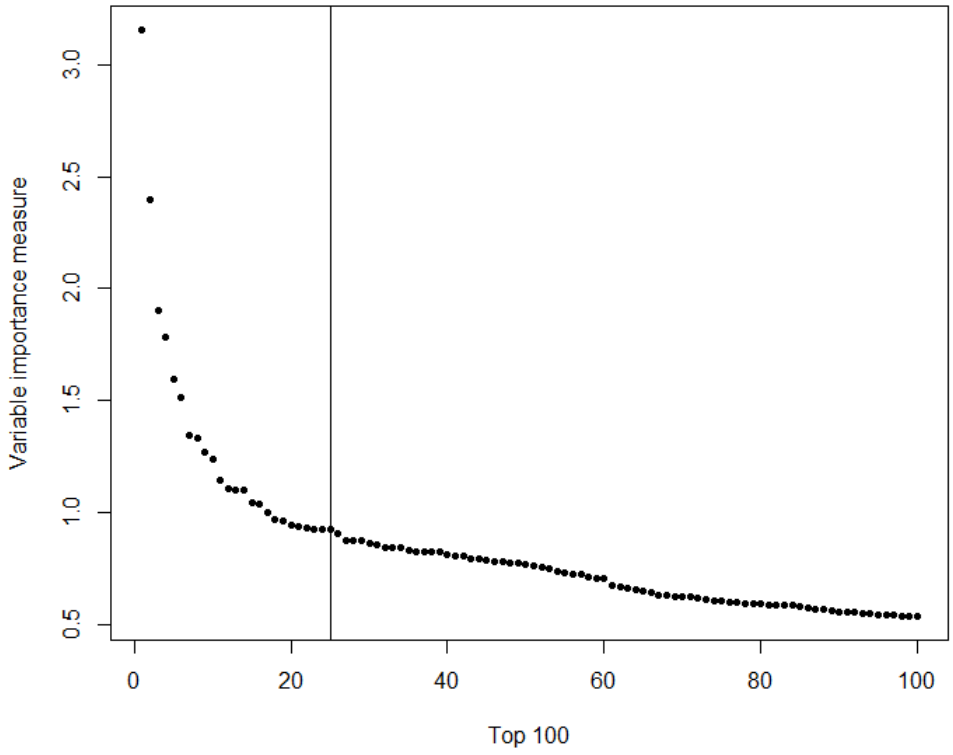

Figure 5.5 – Illustration of the "elbow" method described in Goldstein *et al*. (2010). Here the "elbow" is obtained at 25 loci; these are therefore considered as important loci (modified from Goldstein *et al.* 2010).

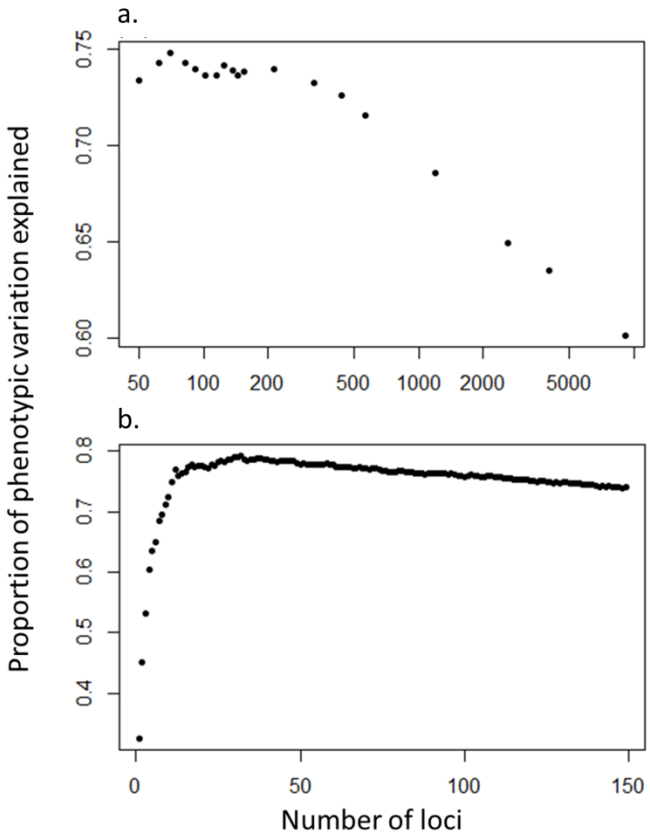

Figure 5.6 – Two-step backward purging analysis for run-timing in Chinook salmon from Brieuc *et al*. (2015). (a) Performances of initial Random Forest analyses on all 9107 loci and subsets of loci with the highest importance values, which indicate that the top 70 loci maximized the proportion of phenotypic variation explained. (b) The backward purging approach was conducted on the top 150 loci; the top 33 loci maximized the PVE and were considered as predictor loci for run timing.

5.12 SUPPLEMENTARY MATERIAL

S5.1 – Supplementary tables and figures for Random Forest implementation and optimization

S5.1.1 - Subset of programs that implement Random Forest algorithms.

S5.1.2 – Results of RF analyses on nine scenarios to determine the effectiveness of a procedure to correct data sets for population structure.

S5.1.3 – Box and whisker plots summarizing the distributions of root mean squared error values from 50 bootstrap iterations of cross validation analyses for the 1,000 locus and 10,000 locus data sets.

S5.2 – R scripts for classification and regression Random Forest tutorials.

S5.3 – Input data files for the classification and regression Random Forest tutorials.

5.13 REFERENCES

Archer E (2016) Package 'rfpermute'.

Azevedo CF, de Resende MDV, Silva FFE, Nascimento M, Viana JMS, Valente MSF (2017) Population structure correction for genomic selection through eigenvector covariates. *Crop Breeding and Applied Biotechnology*, **17**, 350-358.

Barabaschi D, Tondelli A, Desiderio F *et al.* (2016) Next generation breeding. *Plant Science*, **242**, 3-13.

Barson NJ, Aykanat T, Hindar K *et al.* (2015) Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, **528**, 405-408.

Bernard S, Adam S, Heutte L (2012) Dynamic random forests. *Pattern Recognition Letters*, **33**, 1580-1586.

Bernatchez L, Wellenreuther M, Cristián A *et al.* (2017) Harnessing the power of genomics to secure the future of seafood. *Trends in Ecology and Evolution*, **32**, 665-680.

Blagus R, Lusa L (2010) Class prediction for high-dimensional class-imbalanced data. *Bmc Bioinformatics*, **11**, 17.

Botta V, Louppe G, Geurts P, Wehenkel L (2014) Exploiting snp correlations within random forest for genome-wide association studies. *PLoS ONE*, **9**, 11.

Breiman L (1996) Bagging predictors. *Machine Learning*, **24**, 123-140.

Breiman L (2001) Random forests. *Machine Learning*, **45**, 5-32.

Breiman L (2002) Manual--setting up, using, and understanding random forests v4.0.

Breiman L, Cutler A (2007) Random forests.

Brieuc MSO, Ono K, Drinan DP, 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.

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, **81**, 1084-1097.

Chakravarthy A (2015) Prf: Permutation significance for random forests.

Chen C, Liaw A, Breiman L (2004) Using random forest to learn unbalanced data. In: *Technical Report 666, Statistics Department, University of California at Berkeley*.

Chen X, Ishwaran H (2012) Random forests for genomic data analysis. *Genomics*, **99**, 323-329.

Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, **10**, 392-404.

Del Carpio DP, Basnet RK, De Vos RCH, Maliepaard C, Paulo MJ, Bonnema G (2011) Comparative methods for association studies: A case study on metabolite variation in a *brassica rapa* core collection. *PLoS ONE*, **6**, 10.

Desta ZA, Ortiz R (2014) Genomic selection: Genome-wide prediction in plant improvement. *Trends in Plant Science*, **19**, 592-601.

Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**, 997-1004.

Diaz-Uriarte R, de Andres SA (2006) Gene selection and classification of microarray data using random forest. *Bmc Bioinformatics*, **7**, 13.

Espindola A, Ruffley M, Smith ML, Carstens BC, Tank DC, Sullivan J (2016) Identifying cryptic diversity with predictive phylogeography. *Proceedings of the Royal Society B-Biological Sciences*, **283**.

Evans J, Murphy M, Holden Z, Cushman S (2011) Modeling species distribution and change using random forests. In: *Predictive species and habitat modeling in landscape ecology: Concepts and applications* (eds. Drew C, Wiersma Y, Huettmann F), pp. 139-159. Springer, New York.

Fraimout A, Debat V, Fellous S *et al.* (2017) Deciphering the routes of invasion of drosophila suzukii by means of abc random forest. *Molecular Biology and Evolution*, **34**, 980-996.

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

Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2012) A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *Ieee Transactions on Systems Man and Cybernetics Part C-Applications and Reviews*, **42**, 463-484.

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Machine Learning*, **63**, 3-42.

Goddard ME, Hayes BJ (2007) Genomic selection. *Journal of Animal Breeding and Genetics*, **124**, 323-330.

Goldstein BA, Hubbard AE, Cutler A, Barcellos LF (2010) An application of random forests to a genome-wide association dataset: Methodological considerations & new findings. *Bmc Genetics*, **11**.

Goldstein BA, Polley EC, Briggs FBS (2011) Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, **10**, 35.

Hess JE, Zendt JS, Matala AR, Narum SR (2016) Genetic basis of adult migration timing in anadromous steelhead discovered through multivariate association testing. *Proceedings of the Royal Society B-Biological Sciences*, **283**.

Ho TK (1995) Random decision forsts. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pp. 278-282, Montreal, QC.

Holliday JA, Ritland K, Aitken SN (2010) Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in sitka spruce (picea sitchensis). *New Phytologist*, **188**, 501-514.

Holliday JA, Wang TL, Aitken S (2012) Predicting adaptive phenotypes from multilocus genotypes in sitka spruce (*picea sitchensis*) using random forest. *G3-Genes Genomes Genetics*, **2**, 1085-1093.

Hornoy B, Pavy N, Gerardi S, Beaulieu J, Bousquet J (2015) Genetic adaptation to climate in white spruce involves small to moderate allele frequency shifts in functionally diverse genes. *Genome Biology and Evolution*, **7**, 3269-3285.

Huang BFF, Boutros PC (2016) The parameter sensitivity of random forests. *Bmc Bioinformatics*, **17**.

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. *Annals of Applied Statistics*, **2**, 841-860.

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics*, **9**, 166-177.

Japkowicz N, Stephen S (2002) The class imbalance problem: A systematic study. *Intelligent Data Analysis*, **6**, 429-449.

Jiang HY, Deng YP, Chen HS *et al.* (2004) Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *Bmc Bioinformatics*, **5**, 12.

Kuhn M (2008) Caret package. *Journal of Statistical Software*, **28**.

Laporte M, Pavey SA, Rougeux C *et al.* (2016) Rad sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in north atlantic eels. *Molecular Ecology*, **25**, 219-237.
Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, **21**, 1548-1566.
Liaw A, Wiener M (2002) Classification and regression by randomforest. *R News*, **2**, 18-22.
Lin WJ, Chen JJ (2013) Class-imbalanced classifiers for high-dimensional data. *Briefings in Bioinformatics*, **14**, 13-26.
Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031-1046.
Mackay TFC (2014) Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, **15**, 22-33.
Mager KH, Colson KE, Groves P, Hundertmark KJ (2014) Population structure over a broad spatial scale driven by nonanthropogenic factors in a wide-ranging migratory mammal, alaskan caribou. *Molecular Ecology*, **23**, 6045-6057.
Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. *Nature Genetics*, **36**, 512-517.
MATLAB and Statistics Toolbox Release (2016). The Mathworks, Inc., Natick, Massachusetts, United States.
Meng YA, Yu Y, Cupples LA, Farrer LA, Lunetta KL (2009) Performance of random forest when snps are in linkage disequilibrium. *Bmc Bioinformatics*, **10**, 17.
Nichols KM, Kozfkay CC, Narum SR (2016) Genomic signatures among *oncorhynchus nerka* ecotypes to inform conservation and management of endangered sockeye salmon. *Evolutionary Applications*, **9**, 1285-1300.
Pavey Scott A, Gaudin J, Normandeau E *et al.* (2015) Rad sequencing highlights polygenic discrimination of habitat ecotypes in the panmictic american eel. *Current Biology*, **25**, 1666-1671.
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904-909.
Pritchard JK, Di Rienzo A (2010) Adaptation - not by sweeps alone. *Nature Reviews Genetics*, **11**, 665-667.
R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org, Vienna, Austria.
Roberts DR, Hamann A (2012) Method selection for species distribution modelling: Are temporally or spatially independent evaluations necessary? *Ecography*, **35**, 792-802.
Roff DA (2007) A centennial celebration for quantitative genetics. *Evolution*, **61**, 1017-1032.
Rokach L (2016) Decision forest: Twenty years of research. *Information Fusion*, **27**, 111-125.
Rönnegård L, McFarlane SE, Husby A, Kawakami T, Ellegren H, Qvarnström A (2016) Increasing the power of genome wide association studies in natural populations using repeated measures - evaluation and implementation. *Methods in Ecology and Evolution*, **7**, 792-799.
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.

Savolainen O, Lascoux M, Merila J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807-820.
Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, **78**, 629-644.
Shafer ABA, Wolf JBW, Alves PC *et al.* (2015) Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, **30**, 78-87.
Silva CNS, McFarlane SE, Hagen IJ *et al.* (2017) Insights into the genetic architecture of morphological traits in two passerine bird species. *Heredity*, **119**, 197-205.
Stephan J, Stegle O, Beyer A (2015) A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, **6**, 10.
Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *Bmc Bioinformatics*, **8**, 21.
Svetnik V, Liaw A, Tong C, Wang T (2004) Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. *Multiple Classifier Systems, Proceedings*, **3077**, 334-343.
Sylvester EV, Bentzen P, Bradbury IR *et al.* (2017) Applications of random forest feature selection for fine-scale genetic population assignment. *Evolutionary Applications*.
Szymczak S, Holzinger E, Dasgupta A *et al.* (2016) R2vim: A new variable selection method for random forests in genome-wide association studies. *Biodata Mining*, **9**, 15.
Tang F, Ishwaran H (2017) Random forest missing data algorithms. *Statistical Analysis and Data Mining*, **2017**, 1-15.
Waters CD, Hard JJ, Brieuc MSO *et al.* (2018) Genomewide association analyses of fitness traits in captive-reared chinook salmon: Applications in evaluating conservation strategies. *Evolutionary Applications*, **00**, 1-16.
Wilson AJ, Reale D, Clements MN *et al.* (2010) An ecologist's guide to the animal model. *Journal of Animal Ecology*, **79**, 13-26.
Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM (2013) Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics*, **14**, 507-515.
Yang JA, Benyamin B, McEvoy BP *et al.* (2010) Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, **42**, 565-U131.
Yu JM, Pressoir G, Briggs WH *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, **38**, 203-208.
Zhang ZW, Ersoz E, Lai CQ *et al.* (2010) Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, **42**, 355-U118.
Zhao Y, Chen F, Zhai RH *et al.* (2012) Correction for population stratification in random forest analysis. *International Journal of Epidemiology*, **41**, 1798-1806.