SYSTEM ANALYSIS OF THE SPREAD OF INTROGRESSIVE HYBRIDIZATION IN SALMONID POPULATION THROUGH AN AGENT-BASED SIMULATION MODEL

by

Patrick Lorenzo Davide Della Croce

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

of

Doctor of Philosophy

in

Ecology and Environmental Sciences

MONTANA STATE UNIVERSITY
Bozeman, Montana

November 2015

DEDICATION

To my sons Matteo and Tobia: grazie per tutto quello che mi date ogni giorno e per come riempite tutti i giorni la mia vita di gioia e amore. Con tutto il cuore spero che voi possiate un giorno trovare qualcosa nella vita che vi appassioni come la pesca, i pesci ed i fiumi hanno sempre appassionato me.

To my wife Olivia: grazie per avermi seguito, per essermi stata sempre vicina, e per aver, tuo malgrado, letto ed ascoltato cose che mai avresti pensato di leggere o ascoltare.

To my mom, my dad, and my sister: grazie per il vostro sostegno ed il vostro amore. Se oggi sono quello che sono, lo devo anche, e magari soprattutto, a voi.

ACKNOWLEDGEMENTS

First and foremost, I want to thank my advisor and (now) colleague Dr. Geoffrey Poole (MSU) for his constant help and support during my dissertation. Geoff: you have taught me much more than I expected and probably much more than I realize. I am deeply and genuinely thankful for everything you have done for me.

Second, I want to thank Dr. Rob Payn (MSU). Rob: your friendship is one of the best "outcomes" of my time at MSU. Thanks a lot for all the discussions we have had, all the help you gave me, and all the cider you had ready during my visits to Bozeman!

Third, I am grateful for all the help and support I received from my committee members Dr. Gresswell (USGS-MSU), Dr. Cross (MSU), Dr. Baxter (Idaho SU), Dr. Luikart (UOM), and Dr. Izurieta (MSU).

I also want to thank the MSU Fluvial Landscape Lab, the Center for Computational Biology, and Dr. Todd Koel, Micheal Ruhl, and Kate Olson at the Yellowstone National Park Native Fish Conservation Program for their support.

A special thank goes also to Dr. Chris Robinson (Eawag), Dr. Tom Gonser (Eawag), and again Dr. Colden Baxter who, with their inputs during my MS were instrumental for my choice of pursuing a doctorate degree.

Last, I want to thank my Dean Dr. Sara Steinert-Borella and my colleague Dr. Brack Hale at Franklin University Switzerland for their patience and their support during these last two years of my dissertation. Thank you all!

For the financial support I am grateful to the VOEiS Group (NSF EPSCoR Track-2 EPS 0919557), and the Swiss NSF (grant PBSKP3_143944).

TABLE OF CONTENTS

TABLE OF CONTENTS-CONTINUED

TABLE OF CONTENTS-CONTINUED

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF FIGURES - CONTINUED

Figure Page

LIST OF FIGURES - CONTINUED

Figure Page

LIST OF FIGURES - CONTINUED

Figure Page

LIST OF FIGURES - CONTINUED

Figure Page

ABSTRACT

Using the specific example of introgressive hybridization between native cutthroat trout (*Oncorhynchus clarkii* sp., CTT) and introduced rainbow trout (*O. mykiss*, RBT), I have designed and developed an individual-based model that mimics the spread of introgressive hybridization in a river network inhabited by salmonids accounting for realistic population demographics, genetics, and stochasticity. As a tool capable of providing mechanistic explanations for observed patterns of introgression, the model is used in Chapter 2 to show that network bifurcations act as semi-permeable barriers to the spread of introgression across river network by decreasing (a) the rate of spread of non-native alleles and (b) the predictability of such spread. As a tool to test the effects of realistic population demographics and realistic mechanisms for the passage of alleles from one generation to the next the model is used in Chapters 3 and 4 to show that commonly used sampling and analysis methods greatly overestimate the power to detect and the precision to quantify introgression in populations.

CHAPTER 1

INTRODUCTION

Overview and Goals

Non-native species have long been recognized as one of the greatest threats to the conservation of native species worldwide, second only to habitat degradation (Fausch *et al.* 2001). In fact, countless examples have documented the adverse impacts of invasive species on the native flora and fauna all across the globe (e.g., Fritts & Rodda 1998; Holway 1999; Mack *et al.* 2000; Shepard 2004; Baxter *et al.* 2007). However, while predation and competition have long been recognized as the major mechanisms through which introduced species can affect native ones and can potentially lead to their extinction, in the last few decades non-native species have been recognized as being able to put native species at risk of genetic extinction through hybridization and introgression (Rhymer & Simberloff 1996a).

Hybridization refers to the mating of two genetically distinct organisms through which hybrids with intermediate genotypes are formed (Allendorf & Luikart 2007). In many cases, the hybrids produced by hybridizing individuals are sterile and the major consequence for the native species is the loss of reproductive opportunities (e.g., Rozhnov 1993; Kitano *et al.* 2009). While this form of hybridization can indeed affect the long term conservation of a species, especially when the species is rare (Rhymer & Simberloff 1996a; Frankham *et al.* 2010), hybridization becomes most problematic from a conservation perspective when the created hybrids are fertile and able to backcross with

the parental species. Through this phenomenon, called introgressive hybridization or introgression, gene flow between two distinct species can occur, and the gene pool of either of the parental species can be invaded – or become introgressed – with genes from the other species (Allendorf & Luikart 2007). As a result, introgression can seriously undermine the genetic purity of a species and can potentially lead to its genetic extinction (Muhlfeld *et al.* 2009c; Gozlan *et al.* 2010).

Hybridization is particularly common in fish and especially among salmonids (Muhlfeld *et al.* 2009a). In fact, due to their economical and recreational value and because they are easily reared and transported, salmonids have been widely introduced across the globe for centuries (Krueger & May 1991; Fausch 2007; Halverson 2010). Combined with the fact that salmonids have similar mating behaviors and that suitable spawning habitats in freshwater systems are often limited (Bennett & Kershner 2009; Muhlfeld *et al.* 2009a), these widespread introductions have created ample opportunities for native and non-native salmonids to hybridize. As a result, the genetic integrity of many native trout and char populations worldwide is now threatened (e.g., Leary *et al.* 1995; Yamamoto *et al.* 2006).

To date, large datasets describing the spread of introgression between native and introduced salmonids have been collected from a variety of systems. However, while such data have allowed us to monitor and describe the spread of introgression across multiple systems, we are still lacking the ability to provide mechanistic explanations of how the observed spatiotemporal patterns of non-native genes came to be. Yet, understanding the mechanisms behind observed patterns of distribution and abundance of

non-native genes is crucial to design effective management strategies for the protection of native salmonid populations. Indeed, knowing the bottlenecks to the spread of invasive genes across a system is fundamental for effective management strategies that will ensure the long term persistence of pure native salmonid populations.

Moreover, in fisheries studies, the probability to detect non-native genes in a population and the uncertainty around estimates of their abundance are still calculated – if at all – using simple analytical equations (Rasmussen *et al.* 2010; Amish *et al.* 2012) which assume idealized populations and typically do not account at all for the stochasticity involved in the passage of genes from one generation to the next. Yet, real populations are not ideal and in reality stochasticity is likely to be heavily involved in the passage of genes from parents to offspring. Therefore, estimates of the power to detect and the precision to quantify hybridization obtained from commonly used analytical equations may not be reliable. Given that managerial and protection measures are in large part driven by information on the presence or absence of non-native genes and the magnitude of hybridization in a system (e.g., Allendorf *et al.* 2001, 2004), it is arguably of paramount importance to establish when and under which circumstances traditional analytical equations cease to be reliable.

Using the specific example of introgressive hybridization between native cutthroat trout (*Oncorhynchus clarkii* sp., CTT) and introduced rainbow trout (*O. mykiss*, RBT), in collaboration with Dr. Geoffrey Poole (Fluvial Landscape Lab, Department of Land Resources and Environmental Sciences, Montana State University), Dr. Robert Payn (Fluvial Landscape Lab, Department of Land Resources and Environmental

Sciences, Montana State University), and Dr. Clemente Izurieta (Computer Science Department, Montana State University), I have designed and developed an individual-based model that mimics the spread of introgressive hybridization in a river network inhabited by salmonids accounting for realistic population demographics and stochasticity (see *Chapter 1* for a detailed model description and Figure 1). The model is ultimately aimed at quantifying, on a case-by-case basis, the relative contribution to observed introgression patterns of the drivers to the spread of introgressive hybridization across a system (see below). Therefore, the model is to be understood as a tool capable of providing mechanistic explanations for observed patterns of introgression. Moreover, the model also provides a tool to test the effects of realistic population demographics and realistic mechanisms for the passage of genes from one generation to the next on the power to detect and the precision to quantify introgression in salmonid populations.

While the ultimate goal of the model is not reached in this research project, in an attempt to show its utility, I used the model to conduct separate investigations aimed at: (a) increasing our understanding of how non-native genes move across a network and (b) providing realistic confidence in the power to detect non-native genes and the precision to quantify levels of introgression in fish populations associated with a given sampling strategy. Specifically:

Chapter 2

This chapter provides a full description of the model and then uses it to investigates how network topology and specifically the spatial distribution of network

bifurcations influences: (a) the rate at which introgressive hybridization spreads across a network and (b) the predictability of the spread of non-native genes.

Chapter 3

This chapter uses the model to show that simplifying assumptions regarding population demographics and stochasticity – that typically underpin equations used in fisheries – lead to: (a) large overestimations of the power to detect non-native genes in a population and (b) overly optimistic confidence intervals around estimates of proportion of admixture with non-native genes of a population.

Chapter 4

Using a simplified model version, this chapter: (a) defines a new equation to calculate the probability of failing to detect non-native alleles in a population through sampling; (b) quantifies the errors associated with the use and application of the commonly used equation when its underlying assumptions are violated, and; (c) proposes a new methodology to design sampling strategies that can guarantee a certain confidence to detect non-native genes in a population.

Background

Study-Case: Cutthroat and Rainbow Trout

Cutthroat trout were first encountered by Europeans in 1541, when Francisco de Coronado crossed a tributary of the Rio Grande in today's New Mexico (Behnke 2002). About three centuries later, in 1805, Meriwether Lewis and William Clark described the

westslope cutthroat trout in their account of their expedition along the Missouri River, and in honor of both explorers, westslope cutthroat trout received the scientific name of *Oncorhynchus clarkii lewisi* (Behnke 2002). CTT are native to the West Coast of the U.S. and Canada, from northern California to the Kenai Peninsula of Alaska, and also inhabit both sides of the Rocky Mountains in the U.S. and in western Canada (Figure 2). Among salmonids native to the North American continent, the historical range of CTT is second only to that of the lake trout (*Salvelinus namaycush*). However, through a combination of factors – among which are overharvesting, habitat degradation, and invasive species – to date CTT occupy just a fraction of the lakes and rivers they once used to be native to (Behnke 2002). For example, two of the most abundant subspecies of CTT, the Yellowstone CTT (O. c.bouvieri) and the westslope CTT occupy respectively just 43% and 49% of the rivers and lakes they are native to in the United States (Shepard *et al.* 2005; May *et al.* 2007). In addition, introgression with introduced RBT has exposed CTT to the threat of genetic extinction (e.g., Hitt *et al.* 2003; Muhlfeld *et al.* 2009c; Bennett *et al.* 2010).

Rainbow trout are native to the West Coast of the U.S. and Canada and to eastern Russia (Behnke 2002). On the American side of the Pacific Ocean, the historical range of RBT goes from southern California to the Alaskan Peninsula (Figure 3) thus overlapping with large parts of the historical CTT range. Ever since the late 1800s, RBT have been farmed and shipped all across the globe (Halverson 2010), thus arguably making them the best known salmonid in the world (Behnke 2002). To date, RBT have been introduced with various success to 97 countries worldwide and to all continents except for

Antarctica, including rivers and lakes occupied by CTT (Fausch 2007; Halverson 2010). This has obviously provided ample opportunity for naturally allopatric RBT and CTT populations to hybridize. While naturally sympatric RBT and CTT populations have been able to remain largely separated – through spatial and temporal reproductive segregation (Muhlfeld *et al.* 2009c) – introgressive hybridization is normally widespread and extensive where RBT have been introduced and coexistence with CTT has been forced. For example, human-mediated hybridization with RBT is thought to have compromised about 63% of the U.S. westslope CTT populations (Shepard *et al.* 2005), 80% of the Canadian populations of westslope CTT (COSEWIC 2006), and 74% of the populations of Yellowstone CTT in Wyoming, U.S. (Kruse *et al.* 2000). It is thus not surprising, that introgression with introduced RBT is considered by many as being one of the greatest threats CTT are currently facing (e.g., Dowling & Childs 1992; Keeler-Foster 2003; Meyer *et al.* 2006; Boyer *et al.* 2008).

To date, most studies focusing on introgression between CTT and RBT can be classified in two main groups: those focusing on the distribution of RBT genes across river networks and those focusing on the effects of introgression with RBT genes on the fitness and the behavior of CTT individuals. While some studies of the first group simply mapped the distribution and abundance of RBT genes across a given system (e.g., Rubidge *et al.* 2001), the majority of these investigations have taken a *landscape genetics* approach in which the distribution and abundance of RBT genes has been correlated with a set of a priori selected environmental features.

Overview of Landscape Genetics

Landscape genetics emerged about a decade ago as an interdisciplinary field from the recognition that characteristics of the landscape influence spatial and temporal genetic patterns of populations (Manel *et al.* 2003a; Holderegger & Wagner 2006). Consequently, as stated by Manel and colleagues in their seminal paper (2003), landscape genetics is an "amalgamation of molecular population genetics and landscape ecology", the goal of which is to explicitly quantify the influence of landscape features on the genetic structure of populations (Storfer *et al.* 2006). Thus, landscape genetics typically aims at: (a) detecting spatial patterns in the genetic structure of a group of organisms (e.g., clines or abrupt differences) and (b) finding correlations between the observed genetic patterns and features of the landscape hosting the organisms, such as barriers or corridors (Storfer *et al.* 2010).

Since the appearance of the term 'landscape genetics' in the scientific literature, an ever increasing number of studies have embraced the philosophy of landscape genetics and have sought correlations between genetic patterns and landscape characteristics for a variety of systems and species (Holderegger & Wagner 2006). According to an extensive review of the field published by Storfer et al. (2010), so far the identification of barriers to gene flow has been the main focus of landscape genetics studies (e.g., Epps *et al.* 2007; Dionne *et al.* 2008; Meeuwig *et al.* 2010). However, to date other recurrent topics in landscape genetics studies typically included: (a) the identification of migration corridors (e.g., Antolin *et al.* 2006; Wang *et al.* 2009); (b) the study of source-sink dynamics (Cassens *et al.* 2000), and; (c) the investigation of the spatial and temporal scales at

which landscape features influence the genetic structure of a population (Holzhauer *et al.* 2006).

As stated by Holderegger and Wagner (2006), landscape genetics is beneficial to both landscape ecology and population genetics because it "forces landscape ecologists to think more about processes rather than spatial patterns" and it reminds "population geneticists that processes are influenced by the quality of a landscape and not just by pure spatial distance". However, landscape genetics has also proven beneficial to other disciplines and one of the fields that arguably profited most from this new approach has been conservation biology (Segelbacher *et al.* 2010). Indeed, the identification of habitats critical to the conservation of a species, such as migration corridors between populations, can help prioritize protection efforts and allow a more focused and potentially more efficient management of the species of interest. For example, through the combination of genetic data on desert bighorn sheep (*Ovis Canadensis nelsoni*) and landscape models, Epps *et al.* (2007) classified man-made structures according to their potential for blocking gene flow among bighorn sheep. Applying the same strategy, Wang *et al.* (2009), identified chaparral as the least costly habitat type for the dispersal of California tiger salamanders (*Ambystoma californiense*) despite the fact that the amphibians are normally associated with ponds in grasslands in which they spawn. This and similar information can confirm or reveal unexpected interactions between organisms and their environment that can provide crucial information for the management and the protection of a species.

With respect to CTT and RBT trout, landscape genetics-inspired studies revealed that levels of introgression with RBT genes generally tend to be higher in stream reaches with higher temperature, lower elevation, and lower gradient. For example, Weigel *et al.* (2003) found a negative relationship between stream elevation and levels of introgression with RBT genes in the Clearwater River Basin (Idaho, U.S.). Similarly, Muhlfeld et al. (2009c) showed that the presence of RBT genes in the North Fork of the Flathead River System (Montana, U.S.) was positively related to temperature and human disturbance. This kind of research has been very successful in mapping the distribution and abundance of RBT across various systems, and thanks to it we can now say where RBT genes are in the landscape (although, as I will explore in Chapter 2, the confidence in such results may be overly optimistic). Also, this information allows us, to some extent, even to predict, where RBT genes will: (a) move to and (b) be more or less abundant.

Effects of Introgression on Individuals' Fitness and Behavior

As stated above, in the last few decades several studies have investigated the effects of introgression with RBT on the behavior and fitness of CTT. Regarding the effects of hybridization on the behavior of the fish, the main focus of such studies has been on spawning timing. Various investigations on this topic have confirmed the tendency for pure CTT and RBT to spawn on the opposite sides of the spring – summer runoff period, though spawning periods may overlap. Indeed, RBT tend to spawn during the period of rising discharge during snow melt in late spring, while CTT tend to spawn during the subsequent period of decreasing water levels (e.g., De Rito 2004; Muhlfeld *et*

al. 2009b). However, such investigations have also shown that hybrids tend to have spawning periods intermediate to those of the parental species, thus suggesting that once hybrids are created, hybridization will become increasingly easier. In addition to modifying spawning timing, though studies formally investigating the topic have not yet been performed, hybridization between RBT and CTT is suspected to alter the homing behavior of individuals. Specifically, it is suspected that introgression with RBT genes decreases the homing instinct of hybrids, which consequently seem to have higher propensity to stray to non-natal spawning grounds to reproduce (Hitt *et al.* 2003; Boyer *et al.* 2008). This suspicion is based on the fact that hybridization between CTT and RBT is seemingly spreading at a much higher rate across river networks than what could be assumed given the very high spawning site fidelity of CTT (e.g., McCleave 1967; Gresswell & Hendricks 2007).

Hybridization with RBT also seems to affect various components of CTT fitness. According to a series of studies conducted in fish tanks, fish carrying RBT genes appear to be more aggressive (Seiler & Keeley 2007a), grow faster (Seiler & Keeley 2009), and have better swimming performance (Seiler & Keeley 2007b) than genetically pure CTT. This suggests that RBT and hybrid individuals may have better chances of surviving from one year class to the next than genetically pure CTT. On the other hand, Muhlfeld *et al.* (2009a) provided empirical evidence that, on average, the fecundity of female CTT was negatively impacted by introgression with RBT genes. However, the study also showed that F-1 hybrids (i.e., first generation hybrids: the result of a crossing between two pure

individuals) have similar fecundity to that of genetically pure CTT, thus implying that F-1 hybrids may play a crucial role in the spread of hybridization.

Current Challenge: Understanding the Spread of Hybridization

As can be inferred from the above, the effort produced so far by studies on introgression between CTT and RBT has been very successful in: (a) providing a statistical understanding of the distribution and abundance of RBT genes in a system and (b) highlighting the effects of RBT genes on individuals' behavior and fitness. Still, studies on the spread of RBT genes across river systems and studies on the effects of RBT genes on individuals' behavior and fitness have largely remained separated. Therefore, even though we can now say where non-native genes are in a system, and we can also say how hybridization with RBT genes will change the fitness and behavior of individuals, we still cannot say how interactions between genetically driven differences among individuals and landscape features influence the spread of hybridization across a system. Consequently, to date we are unable to provide a mechanistic explanation for observed patterns of hybridization. Yet, understanding the mechanisms behind the spread of RBT genes in a system is crucial if we want to identify the bottlenecks to the spread of RBT genes and hence, be able to design effective management strategies for CTT populations.

From a mechanistic perspective, regardless of the species considered, observed patterns of introgression across a system are likely to be influenced by interactions between physical and biological drivers controlling the spatiotemporal spread of non-

native genes across a system. Specifically, on the one hand, the spatial movement of non-native genes (i.e., across a system from one sub-population to the next) is likely to be driven by: (a) physical factors that determine the distribution of and the connectivity among the physical habitats hosting the species (network connectivity patterns) and; (b) the genetically driven differences among native, non-native, and hybrid individuals with respect to their propensity to disperse among habitats (*propensity to stray*). On the other hand, the temporal spread of introgression (i.e., from one generation to the next, within a single sub-population) is likely to be driven by the interactive effects of non-native genes on: (a) *propensity to crossbreed* (i.e., the likelihood of cross breeding relative to self-similar mating); (b)*fecundity* (i.e., the number of viable offspring produced per mating by an individual), and; (c)*survival ability* (i.e., the likelihood of an individual to grow to reproductive maturity). Moreover, stochastic events in the life cycle of single individuals can also affect the spatiotemporal spread of non-native genes. For example, Boyer *et al.* (2008) observed that while the spread of RBT genes in the North Fork of Flathead River (Montana, U.S.) happened largely through stepping stone movement – in which RBT moved upstream from one sub-population to the next – random and unpredictable long-distance movements of RBT genes also happened.

As a result of the above, a mechanistic understanding of the movement of non-native genes across a system requires the ability to: (a) investigate how the physical and biological drivers to the spread of hybridization influence the movement of non-native genes across a system and from one generation to the next and (b) quantify the effects of stochasticity on the spatiotemporal spread of introgression. Yet, interactions between the

physical and biological drivers to the spread of introgression are difficult – if not impossible – to study in the field, let alone the effects of stochasticity. Therefore, for the present research project, I have decided to use an individual-based simulation model to investigate how landscape features, genetically driven differences among individuals, and stochasticity may drive the movement of non-native genes across a river system.

Individual-Based Simulation Models

As the name itself reveals, *individual-or agent-based models* (IBM) are simulation models that mimic the actions and the interactions of each individual organism and evaluate how the behavior of each individual organism affects the behavior of the simulated system as a whole (Grimm *et al.* 2006a). In IBM, a simulated environment (not necessarily homogeneous) is populated by individuals with a series of characteristics (e.g., gender, age, or species). The actions of each individual and its interactions with other individuals and the environment are described by user-defined mathematical equations. As a result, the structure of IBM allows researchers to conduct simulations that are tailored to the system under study by accounting for the system's specifics: (a) population demographics (e.g., age structure and genders); (b) behavior of single individuals; (c) differences in behavior and fitness among individuals with, for example, different genotypes or genders; (d) habitat effects on simulated individuals, and; (e) stochasticity in the simulations (e.g., the behavior of simulated individuals can be programmed to be anywhere between fully deterministic to completely random).

As recognized by several authors (e.g., Balkenhol *et al.* 2009; Epperson *et al.* 2010), the structure of IBM makes them much better suited to tackle real-world complex

issues such as the spread of introgressive hybridization across river systems than the traditional analytical models currently used in landscape genetics. In fact, being an amalgamation of population genetics and landscape ecology (Manel *et al.* 2003b), landscape genetics inherited not just the methodologies typical of both "parental" disciplines but also their simulation and mathematical models (Epperson *et al.* 2010). However, traditional analytical models rely on simplistic assumptions that are likely to be violated in real systems (Balkenhol *et al.* 2009). For example, Fisher-Wright models assume idealized populations with discrete generations (Allendorf & Luikart 2007). Similarly, isolation-by-distance models used in population genetics normally assume homogeneous landscapes between two points (Balkenhol *et al.* 2009). While such models have been widely used (e.g., Rasmussen *et al.* 2010; Amish *et al.* 2012), their limitations may prevent a direct and efficient application to real systems (Balkenhol *et al.* 2009). Contrary to that, IBM are able to account for user-defined levels of realism in the simulations, and thus, as recognized by Epperson *et al.* (2010) in their essay on the utility of simulation models, they are likely to prove beneficial to landscape genetics. For example, IBM could help determine under which circumstance simple analytical models are reliable and under which they are not. Indeed, IBMs allow for the systematic relaxation of the assumptions to which analytical models abide. Through a comparison between results obtained from classical analytical models and results obtained from various IBM accounting for increasing degrees of departure from ideal conditions, it is possible to: (a) determine if and under which conditions analytical tools cease to be reliable, and (b) identify characteristics of the individuals, the populations, or the

environment that have the largest impact on the response of interest: a piece of information that could potentially be used to design field studies and allow more efficient use of resources.

Furthermore, as any other model, IBMs represent hypotheses of how a system works, and can therefore generate predictions which can then be tested and verified against field data. However, because of their ability to simulate and track single individuals across a system, IBM can be used not only to formulate predictions, but also to determine which data are necessary from the field to reliably test these predictions. Again, such information can help prioritize field efforts and allow for a more efficient use of field resources. In fact, the greatest strength of IBM arguably lies in their ability to generate outputs that are directly comparable with field data due to their individual-based nature. This ability allows IBM to generate outputs specifically tailored for the problem at hand, thus allowing highly efficient approaches to complex real-world issues (Epperson et al. 2010), such as the spread of introgressive hybridization between native and introduced salmonids in stream networks.

With respect to the study of spatiotemporal spread of introgressive hybridization, IBM provide a well-suited tool because they allow simulation of the genome – or a set of genes – for each simulated individual, thus simultaneously allowing for the tracking of both individuals and genes across a system. Moreover, as stated above, IBM can account, on a case-by-case basis, for the physical and the biological drivers to the spread of introgression in a system as well as for stochasticity, and can provide targeted information on how observed patterns of introgression came to be.

To date, spatially explicit models that can be used to study the spread of non-native genes across river systems are available (e.g., CDFISH, Landguth *et al.* 2012a). Still, I am aware of no simulation model that can explore the spread of introgressive hybridization across a system while simultaneously accounting for: (a) landscape features and quality; (b) realistic demographics of populations; (c) genetically driven differences among individuals, and; (d) realistic stochasticity. Therefore, our current ability to explain and understand observed patterns of introgression across a river network is still limited. Moreover, confidence in the power to detect and in the precision to quantify non-native genes in a fish population for a given sampling strategy is still calculated using oversimplified binomial equations. As a result, our capability to design effective management and protection strategies for populations of native salmonids undergoing – or at risk of – introgression may still be severely impaired, even in the face of the ever increasing quantity of field data available.

Figures

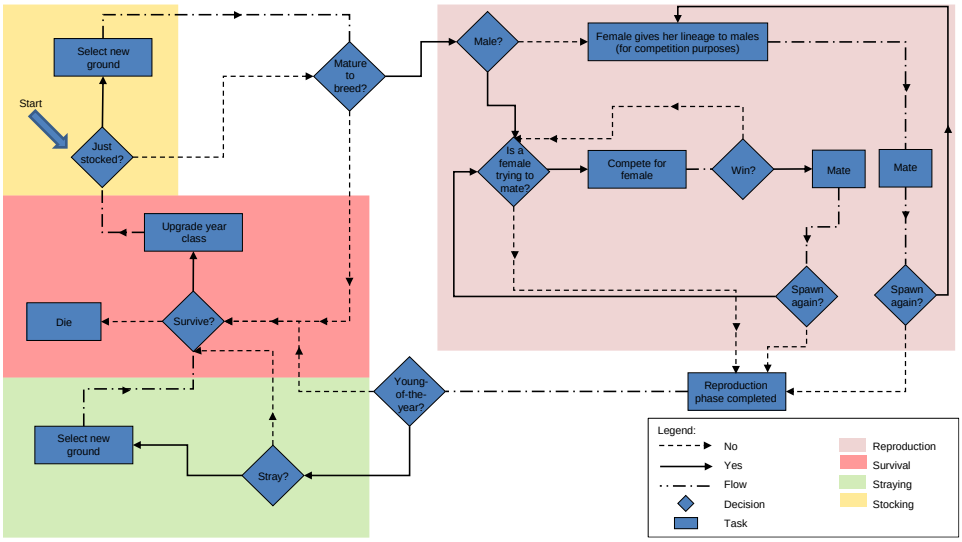

Figure 1.1: Flow chart summarizing one year for a fish in the model (a complete description of the model is provided in Chapter 2).

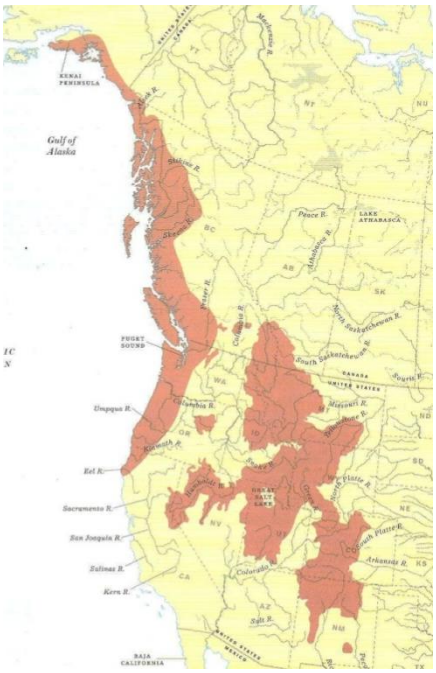

Figure 1.2:Historical range of cutthroat trout (*Oncorhynchus clarkii* sp.), from Behnke (2002).

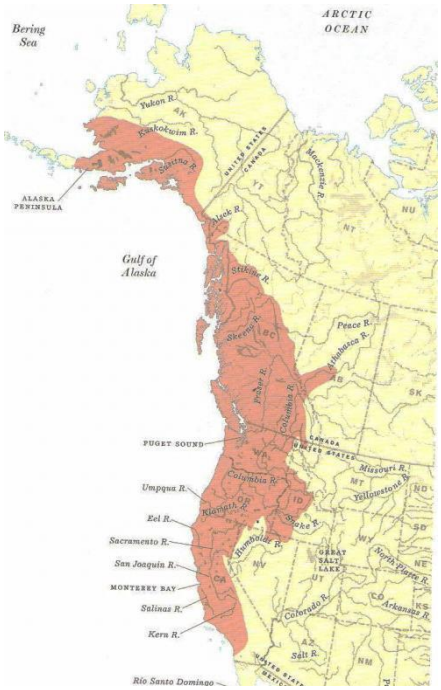

Figure 1.3: Historical range of rainbow trout (*Oncorhynchus mykiss*sp.)in North America, from Behnke (2002).

CHAPTER 2

SIMULATING THE EFFECTS OF STREAM NETWORK TOPOLOGY ON THE SPREAD OF INTROGRESSIVE HYBRIDIZATION ACROSS FISH POPULATIONS

Contribution of Authors and Co-Authors

Manuscript in Chapter 2

Author: Patrick Della Croce

Contributions: Designed the study, performed simulations and analyzed data, and wrote the manuscript.

Co-author: Dr. Geoffrey C. Poole

Contributions: Assisted with study design, analyses, results discussion, commented and edited the manuscript.

Co-author: Dr. Robert A. Payn

Contributions: Assisted with study design and edited the manuscript.

Co-author: Dr. Clemente Izurieta

Contributions: Assisted with study design and edited the manuscript.

Manuscript Information Page

Patrick Della Croce[1,#], Geoffrey C. Poole[1,2,3], Robert A. Payn[1,2], Clemente, Izurieta[3,2]

Journal Name: Ecological Modelling (Elsevier)
Status of Manuscript:
____Prepared for submission to a peer-reviewed journal
____Officially submitted to a peer-reviewed journal
____Accepted by a peer-reviewed journal
__x_Published in a peer-reviewed journal
Published online 12 March 2014

[1]Dept. of Land Resources & Environmental Sciences, Montana State University, P.O. Box 173120, Bozeman, MT 59717-3120
[2]Institute on Ecosystems, Montana University System, Montana State University, 106 AJM Johnson Hall, Bozeman, MT 59717
[3]Department of Computer Science, Montana State University, 357 EPS Building, Bozeman, MT 59717

[#]Corresponding author: patrick.dellacroce@gmx.ch

Abstract

Introgressive hybridization with introduced species is threatening the genetic integrity of many native fish populations worldwide. To date, several studies have indicated a direct relationship between rates of hybridization (*H*) within a sub-population of native fish and the stream distance (*D*) separating that sub-population from the closest source of introduced non-native genes. However, the relationship between *D* and *H* is rarely quantified, and the effect of the spatial arrangement of stream bifurcations and sub-populations (i.e., the network topology) is typically ignored. For this study, we developed and applied a novel individual-based simulation model to compare the spread of non-native genes across three stream networks with differing network topologies: linear, trellis, and dendritic. The model we used simulates mating, survival, and the movements of each fish in the network, where the likelihood of a fish to move between two sub-populations is a function of both *D* and the number of bifurcations between the two sub-populations. To monitor the spatiotemporal spread of hybridization, the model keeps track of the genetic composition and the breeding location of each fish present in the network over time. Simulation results agree with past field studies that suggest that *D* is a strong predictor of *H*. However, our simulations further suggest that the nature and the strength of the *H* vs. *D* relationship are likely to be a function of the stream network topology. Specifically, in our simulations, network bifurcations act as permeable barriers to the spread of non-native genes across a network while the strength of the *H* vs. *D* relationship is inversely proportional to the complexity of the spatial arrangement of bifurcations and sub-populations. Our study shows that considering network topology

could yield a better understanding of observed patterns of hybridization across stream systems, and thus may help inform effective strategies for the management of native fish populations threatened by introgression.

Keywords

individual-based models; invasive species; introgressive hybridization; networks topology; cutthroat trout.

Introduction

Hybridization with non-native species is a well-recognized threat to the conservation of native species worldwide, especially when hybrids are fertile and able to breed with both parental species (Leary *et al.* 1995; Rhymer & Simberloff 1996a). This phenomenon, known as introgressive hybridization, allows genes to flow between two populations of parental species. The resulting lack of genetic isolation potentially leads to permanent loss of locally adapted genotypes or the genetic extinction of a native species (Rhymer & Simberloff 1996a; Gozlan *et al.* 2010). As a result, managers tasked with protecting native species from introgression require an understanding of how non-native genes spread across a system. For instance, managers need to understand how landscape characteristics influence the patterns of genetic connectivity of a system and thus the movement of non-native genes across a landscape (Cote *et al.* 2009).

Introgression of salmonid fish populations in riverine systems represent a useful case study in which to assess the fundamentals of how landscape features influence spatial patterns of hybridization. First, connections among fish sub-populations in stream

networks are limited to the wetted channel, hence eliminating uncertainty associated with the more diffuse movement of individuals across terrestrial systems. Second, riverine systems are often invaded from known sources of non-native individuals (e.g., Boyer *et al.* 2008; Bennett *et al.* 2010), which is relevant because connectivity to non-native source habitats influences the susceptibility of a site to invasion (e.g., Adams *et al.* 2001; Benjamin *et al.* 2007). Third, patterns of network connectivity affect gene flow between fish populations inhabiting stream networks (Wofford *et al.* 2005; Guy *et al.* 2008). Finally, hybridization and introgression are particularly common and extensively studied in salmonids (Bennett & Kershner 2009; Muhlfeld *et al.* 2009a).

In the western U.S., native cutthroat trout (CTT; *Oncorhynchus clarkii*) are threatened with genetic extinction by widespread introgression with introduced rainbow trout (RBT; *O. mykiss*) (Dowling & Childs 1992; e.g., Kruse *et al.* 2000; Keeler-Foster 2003; Shepard *et al.* 2005; Meyer *et al.* 2006). To date, patterns of introgression between CTT and RBT have been monitored across several stream networks (e.g., Dowling & Childs 1992; Kruse *et al.* 2000; Rubidge & Taylor 2005). Several studies show the importance of connectivity and proximity to sources of RBT genes in determining the susceptibility of CTT populations to introgression (e.g., Rubidge & Taylor 2005; Bennett *et al.* 2010). Other work has studied the influence of physical features such as waterfalls, dams, or other barriers that are known to influence the movement of non-native genes across a river network (e.g., Adams *et al.* 2000; Wofford *et al.* 2005). However, we are aware of no detailed investigations of the effect of stream bifurcations on the spread of

non-native genes, despite evidence that stream network topology and complexity affects fish movement and migration (e.g., Olsen *et al.* 2008).

Simulation modeling provides a flexible and powerful approach to investigating problems in landscape genetics (Epperson *et al.* 2010; Hoban *et al.* 2012). For instance, clinal models can be used to interpret observed spatial distributions and abundances of the alleles of two species undergoing introgression (e.g., Blumler 2003; Bímová *et al.* 2011). Recently, individual-based simulation models (*sensu* Grimm *et al.* 2006b) have specifically received increased recognition for their contribution in the management and the conservation of native species, including studies that investigate patterns of connectivity across stream networks. For example, using an individual-based model, Jager *et al.* (2001) have demonstrated that river impoundments influence the genetic diversity and the likelihood of extinction of white sturgeon (*Acipenser transmontanus*). Additionally, the individual-based simulation program CDFISH (Landguth *et al.* 2012b) was used to predict and compare the likelihood of different measures aimed at protecting native westslope cutthroat trout (*O. c. lewisi*) in Glacier National Park, USA (Muhlfeld *et al.* 2012). Individual-based approaches allow for relaxation of assumptions that constrain more conventional models and are therefore ideally suited to study complex issues in non-idealized settings (Epperson *et al.* 2010). For instance, an individual-based approach can account for fitness and behavioral differences among individuals based on their age, gender, or genetic composition – which all may influence simulated genetic distributions.

Here, we introduce an individual-based model that integrates both landscape connectivity patterns and life-cycle characteristics of individual fish to simulate the

spread of hybridization between native and non-native species across river networks. We also present results from modeling experiments that explicitly investigate the effects of network topology on the spread of introgressive hybridization across river networks. Specifically, we analyzed how the simulated rate of non-native gene movement was influenced by the number and arrangement of network bifurcations in 3 hypothetical stream network topologies. Our simulation model assumes that bifurcations force individual fish to select between two possible directions when moving between interconnected sub-populations. We hypothesized that the aggregated genetic effect of these selections at bifurcations will differ among topologies, but generally will slow the spread of non-native genes across differing network topologies. We expected: (1) the susceptibility of native fish populations to hybridization to decrease with increasing bifurcation density in stream networks; and (2) a less predictable spread in hybridization across networks with more irregular arrangement of bifurcations.

Methods

We developed an individual-based simulation model of fish movement within stream networks and then applied the model to three theoretical stream networks with differing topology. We first present the structure and assumptions of the model (*Simulation Model Description*) followed by the methods employed to apply the model, including parameterizations and model scenarios descriptions (*Simulations and Analyses*).

Simulation Model Description

We developed an individual-based model that tracks the genetic composition and spawning location of each fish living in a stream network. The modeled stream network is composed of discrete and interconnected spawning grounds (Figure 1), where each spawning ground is inhabited by a sub-population of fish. Thus, when describing the model, we refer to the collection of fish associated with a particular spawning ground as a "sub-population" within the total network population.

The model simulates reproduction, survival, and fidelity (or lack thereof) to the natal spawning ground for each fish in the population. The model represents a simplified system that incorporates critical characteristics of salmonid fish populations – such as age structure and genders differentiation – that influence the spatiotemporal spread of introgressive hybridization across a river network (i.e., among spawning grounds and from one generation to the next within a spawning ground).

We developed the model using the Java programming language and runtime environment (Oracle, Redwood Shores, California, USA). Model code is available upon request, from the corresponding author.

Stream Network and Patterns of Connectivity. The model requires a user-defined stream network topology, comprised of spawning grounds, stocking sites, bifurcations, and stream segments. Spawning grounds are the locations in the stream network that host interbreeding sub-populations of fish, while stocking sites represent the point of entry to the network for non-native individuals. Bifurcations and stream segments are used to

connect spawning grounds and stocking sites to one another and thus to determine the topology of the network (Figure 1).

Bennett *et al.* (2010) showed that the probability of fish moving between two spawning sites in a stream network was determined by the distance between spawn sites, as follows:

$$P_{xy} = e^{-\lambda_1 D_{xy}} \qquad (1)$$

where P_{xy} represents the probability of a fish inhabiting spawning or stocking ground x to move to spawning ground y, D_{xy} is the distance along the stream channel between spawning or stocking grounds x and y (in km), and λ_1 is an exponential decay constant determined from empirical data by Bennett *et al.* (2010) to be 0.05.

Equation 1, however, does not account for bifurcations in the stream network. We accounted for intervening bifurcations along the path of a given fish by modifying equation 1 as follows:

$$P_{xy} = \frac{1}{2^{b_{xy}}} e^{-\lambda_2 D_{xy}} \qquad (2)$$

where b_{xy} represents the number of bifurcations between spawning or stocking grounds x and y. Equation 2 assumes that the stream segments offer no additional resistance to movement other than their length. Also, equation 2 effectively assumes that each moving individual arriving at a bifurcation has an equal chance of selecting either of the 2 stream segments available, such that upstream and downstream movements are equally likely. The probability of a fish moving between two spawning grounds is thus effectively reduced by half for each bifurcation between the two spawning grounds. We consider only bifurcations that have spawning sites in both directions because we

presume that a fish seeking a spawning ground might temporarily explore a fork containing no spawning ground, but would move back into the rest of the stream network to find one.

Fish Population and Life-Cycle Simulation. Once the stream network topology is described within the model, the model populates each spawning ground with a user-defined number of genetically pure native fish equally divided between males and females. The age structure of each sub-population (i.e., number of fish per year class) can be adjusted according to the simulation needs for each individual spawning ground. However, the maximum age a fish can reach and the age of sexual maturity must be the same for all sub-populations.

During execution, the model simulates annually the introduction and dispersal of non-native fish along with reproduction, spawning site selection, and survival of each fish (native, non-native, or hybrid) present in the network. At the beginning of any simulation year, a user-defined number of non-native fish are introduced assuming an equal sex ratio. When multiple stocking events over the course of a simulation are required, then each single event has to have the same number of fish, with the same age structure and gender ratio. Non-native fish enter the network at the stocking site. Right after they have entered the network, stocked non-native fish are moved to one of the network's spawning grounds according to equation 2. During a simulation, no fish is allowed to move back to- or linger on the stocking site, and thus no simulated mating or survival occurs in the stocking site.

To distinguish between pure native, pure non-native, and hybrid individuals, the model calculates the lineage of each fish present in the network. We defined the genetic lineage of each fish (L_i) as the fraction of non-native ancestry of each individual, where $L_i = 0$ indicates pure native, $L_i = 1$ indicates pure non-native, and $0 < L_i < 1$ indicates a hybridized individual. L_i for any individual is the arithmetic mean of the L_i values from its parents. The use of L_i as the measure of non-native ancestry assumes perfect knowledge of the genome of each simulated fish. By tracking individuals' L_i, the model results are unaffected by uncertainty arising from the use of a finite number of species-diagnostic markers (e.g., Hohenlohe *et al.* 2011; Kalinowski *et al.* 2011), such as: (a) imperfect information on the fish genome and (b) confounded investigations of the independent influence of a given model variable on the genetic response of interest (see also Sections 4 and 6).

During reproduction, each mature female randomly selects a mating partner among all the mature males available in the same spawning ground. The model assumes that each mature female mates exactly once per year, whereas males can be selected by multiple females and thus mate multiple times each year. The number of offspring resulting from each mating event is a user-defined model parameter. The sex of each offspring is determined stochastically with an equal chance of being male or female. Thus an equal sex ratio is not necessarily maintained through a given simulation, and multiple model executions with the same inputs will generate differing sex ratio dynamics.

Prior to reaching sexual maturity, each juvenile fish has a user-defined chance to stray to a non-natal spawning ground. Based on data for salmonids (Hendry & Stearns

2003) the model assumes that once a fish has spawned at a particular spawning ground, it returns there to spawn for the rest of its life. As a consequence, fish in the network have only one opportunity to stray.

Each year, a user-defined portion of the fish in each year class is randomly selected from the entire network population and advanced to the next year class. The remaining fish, including all the fish of the oldest year class, are assumed to have died and are removed from the simulated population. The number of fish selected to survive is a constant for each year class, effectively maintaining a constant age structure of the overall population in the network (see Figure 3 for an example).

The assumption of a constant age structure in the overall population means that introduced non-native fish have the same chance of growing to the next year class as any other fish in the network. As can be inferred from equation 2, stocked fish have higher chance of clustering among the spawning grounds closest to the stocking sites, rather than dispersing evenly across the network. Thus, the populations on the most heavily invaded spawning grounds would be subjected to an artificially increased, density-driven mortality if the age structure of each sub-population was to be maintained constant over time. To avoid this indirect penalization of the stocked non-native fish in this study, we chose to maintain constant age structure over the whole network and allow age structure in individual sub-populations to vary.

For each year of a simulation, the model output includes the gender, age, lineage, natal location, and spawning location of each fish present in the network.

Simulations and Analyses

Stream Networks. We created three models of stream networks that differed solely with respect to their topology and juxtaposition of spawning grounds, including: a linear network, a trellis network, and a dendritic network (Figure 1). The three networks had the same total stream length (260 km), the same number of spawning grounds (11), and a single non-native stocking site located 20 km downstream of the spawning ground at the network outlet. The linear network has no bifurcations and is the highest mean distance between spawning sites (Figure 2a). The trellis network has a regular, alternating pattern of bifurcations (Figure 2b) along a single main stem stream, while the dendritic network represents an irregular pattern of bifurcations (Figure 2b) in a very compact network with a shorter average distance between spawning sites (Figure 2a). The trellis and dendritic network models we used are also analogous to contrasting fluvial network topologies found in nature, as governed by bedrock weathering and fracture patterns (see for example Guy *et al.* 2008).

We determined a value for λ_2 that would produce a probability of fish moving between spawning grounds in the dendritic network similar to the empirical observations of Bennett *et al.* (2010). To determine an appropriate value for λ_2, we created a set of D_{xy} and b_{xy}pairs for every unique pair of spawning and/or stocking sites in the dendritic network (Figure 1). We used Equation 1 (λ_1= 0.05 km^{-1}; Bennett *et al.* 2010) to calculate P_{xy} for each D_{xy} value in the set of D_{xy} and b_{xy}pairs. We then used Equation 2, along with the derived set of D_{xy} and b_{xy} values, to fit λ_2 to the P_{xy} from Equation 1, by minimizing the summed absolute error. This yielded λ_2 = 0.025 km^{-1} as a value (rounded to the

nearest 0.005 km^{-1}) that would allow the model to incorporate the same overall pattern observed by Bennett *et al.* (2010), but account for variation in b_{xy}.

Fish Population Parameter Selection. For this study, we populated each modeled spawning ground with an initial native population of 2000 fish divided among eight year classes according to data for salmonid age class structures from Langeland and L'Abee-Lund (1996) and Behnke (2002) (Figure 3). The maximum age any fish could reach during a simulation was 8 years and we assumed sexual maturity at age 4 based on Behnke (2002). Given 11 spawning sites per network, these assumptions yielded a network-wide population size of 22,000 fish, which is a reasonable population density (84 fish per km) for CTT (Hilderbrand & Kershner 2000). Furthermore, by comparing the mean variance of the spread of hybridization across our dendritic network among runs with 200, 500, 1000, or 2000 individuals per spawning ground (Figure 4), we determined that 2000 fish per spawning ground was a sufficient number of effective samples to minimize random disproportionate effects of individual fish on the genetic composition of future generations.

We simulated stocking regimes of 3 different durations, where 1100 non-native fish (5% of the network-wide total initial native population) were added in each of 5, 10, or 20 consecutive years. All stocked fish were in year class 1 with an equal number of males and females.

We set the number of offspring per mating to 50, regardless of the female's age. This number was sufficient to produce enough young of the year fish to maintain a stable network-wide age structure over time. Although this number may be low relative to

actual fecundity, a larger value for this parameter would have increased the computational time for simulations without changing the model results. Based on data for Yellowstone CTT (McCleave 1967), we assumed that each juvenile fish had only a 5% chance of straying to a non-natal spawning ground prior to its first mating.

Simulations and Analyses Conducted. We simulated 9 scenarios based on fully factorial combinations of 3 network topologies and 3 stocking regimes. Each scenario was run over 100 years of simulated time. In order to quantify the variation in model output due to stochasticity in mate selection, reproduction, straying, and survival we simulated each scenario 25 times and analyzed the ensemble results. To verify that 25 simulations effectively captured variance in model behavior due to stochasticity, we compared ensemble results based on 25, 50, and 100 executions of the dendritic model and found little difference in the mean variance of the spread of hybridization among the ensemble results (Figure 4).

We quantified the rate of the spread of introgressive hybridization across our networks by deriving the *year of complete hybridization* (YCH) for each spawning ground. We defined the YCH of a spawning ground as the number of years since the first stocking event at which all the offspring produced by that sub-population had some non-native ancestry (e.g., all offspring had $L_i > 0$). Consequently, we defined the year of complete hybridization of the network (YCH_{max}) as the highest YCH among the spawning grounds within a network. For a spawning ground to produce only offspring with non-native ancestry, non-native genes must: (a) arrive on the spawning ground; (b) be carried by a fish that survives to the age of reproductive maturity, and; (c) be

successfully passed to subsequent generations and spread among individuals. Thus, YCH provides an objective measure of the rate at which these steps are successfully occurring in each spawning ground. Exploration of dynamics in the spatial distribution of YCH across spawning grounds in the various scenarios allows us to assess the relative influence of network structure and stocking rates on the relative rate of non-native gene movement across the network and the spread of non-native genes within subpopulations.

For each spawning ground, we performed linear regressions on the ensemble mean simulated YCH vs. the stream distance (*D*) between the spawning ground and the source of non-native genes (the stocking site). We interpreted the slopes of these regressions as a relative metric of the rate of the spread of hybridization across the network. These slopes can be interpreted as the distance-normalized travel time of non-native genes to spawning grounds (yr km^{-1}) and are thus inversely related to the speed of the movement of non-native genes from the source of hybridization (km yr^{-1}). Therefore, smaller slopes indicate a faster spread of hybridization while larger slopes indicate a slower spread of hybridization.

For each stocking scenario, we tested for statistically significant differences ($\alpha = 0.01$) in the slopes of the regressions among the three simulated networks, in order to test our expectation that network bifurcations would slow the spread of hybridization for a given D. Additionally, we used YCH_{max} to compare the overall rates of hybridization among different network topologies. Moreover, we compared the distributions of *D* among our three networks to illustrate the effects of bifurcation on the mean and maximum D_{xy} (Figure 2a). Finally, for the trellis and dendritic networks, we performed a

linear regression on the number of bifurcations vs. D_{xy} to illustrate the effect of differences in bifurcation density (i.e., number of bifurcations km^{-1} of stream channel) among networks (Figure 2b).

From the regressions of YCH vs. D_{xy}, we used the coefficient of determination (R^2) obtained from all the 25 replicates of each simulated scenario as a measure of the sensitivity of results from individual networks to stochastic processes within the model. We then compared the distribution of R^2 among network types using an ANOVA with a Tukey HSD multiple comparison test (Ott 1997; R Core Team 2015), in order to test the expectation that more complex arrangements of bifurcations would result in less predictability of YCH as a function of D.

All statistical tests were conducted using R (version 2.13.2, R Core Team 2012).

Results

YCH was generally lowest in the dendritic network, intermediate in the linear network, and highest in the trellis network (Figure 5a). Thus, complete network hybridization (YCH_{max}) occurred first in the dendritic network (YCH_{max} = 39), followed in order by the linear network (YCH_{max} = 50) and the trellis network (YCH_{max} = 69). Variation in *D* explained simulated YCH across spawning grounds, regardless of the network or stocking duration simulated (Figure 6). The YCH for each spawning ground decreased somewhat with increasing stocking durations, but was largely insensitive to stocking duration. Thus, we report remaining results using the simulations with the longest stocking regimes.

Across the 25 repetitions of each scenario, the mean of R^2 of the linear regression of YCH vs. *D* for the dendritic network was significantly lower than that of the linear and trellis networks ($p > 0.001$, Figure 7), likely because *D* was less correlated with the number of bifurcations between spawning sites in the dendritic network (Figure 2b). The variation in R^2 among runs was also highest for the dendritic network (Figure 7), suggesting that YCH is more sensitive to stochastic events within the dendritic network model than in the trellis or linear network models.

While D explains almost all of the variation in YCH within a network, network topology significantly influenced the rate of spread for hybridization within the network (slope of lines in Figure 5b; all slopes are different at $p<0.001$). The linear network yielded the fastest spread of hybridization from the stocking ground (or shortest transport time of non-native genes), with the lowest slope of YCH vs. *D* of 0.137 yr km^{-1} (95% confidence interval between 0.124 and 0.150 yr km^{-1}); next fastest was the dendritic network with a slope of 0.250 (95% confidence interval between 0.214 and 0.286); and the slowest spread of hybridization was in the trellis network with a slope of 0.430 (95% confidence interval between 0.402 and 0.457).

Discussion

Effects of Network Topology on the Spread of Non-Native Genes

Model results are consistent with previous work that suggests *D* is a primary control on the rate of the spread of hybridization (e.g., Wofford *et al.* 2005; Bennett *et al.* 2010), even within complex network topologies. However, *sensu* Moritz *et al.* (2013),

our work also indicates that network characteristics other than *D* (e.g., bifurcations) influence patterns of connectivity among spawning sites. Failure to consider topology may lead to inaccurate expectations regarding the rate of non-native gene spread. Indeed, in agreement with Olsen *et al.* (2008), our results suggest that network topology and complexity may have a strong influence on how *D* and hybridization rates are related. Specifically, network bifurcations can decrease the rate of non-native gene movement and the spread of hybridization across a network (Figure 5b) because each intervening bifurcation halves the probability of fish moving between spawning sites for a given *D*. Yet, bifurcations also tend to compact the network, thus reducing the mean and the maximum *D* among sites in the network (Figure 2a). In the case of the dendritic network, the compacting effect of bifurcations seems to override their slowing effect, and the network has the lowest simulated YCH_{max}, even though the rate of the spread of hybridization is lower than that of the linear network. On the other hand, the trellis network has the highest YCH_{max} and the slowest rate of spread of hybridization, which suggests that the slowing effect of bifurcations is stronger than their compacting effect. Differences between dendritic and trellis networks suggest that the relative magnitude of increasing or decreasing the spread of hybridization varies depending on the number and the spatial arrangement of bifurcations. We posit that the maximum *D* and bifurcation density of the network might be useful statistical predictors of the change in YCH_{max} relative to a linear network; though quantifying such a relationship is beyond the scope of this paper.

The spread of hybridization across an irregular distribution of bifurcations (i.e., the dendritic network) was significantly less predictable from the stream channel distance than the spread of hybridization across either the trellis or linear networks (Figure 7). The linear network has no effect of bifurcations and thus *D* dominates as a predictor of the spread of hybridization. The trellis network is influenced by bifurcations, but the number of bifurcations is strongly correlated with *D* (Figure 2b), such that the mechanistic effect of bifurcations on the spread of hybridization can be accounted for empirically by altering the relationship with *D*. In the dendritic network, the number of bifurcations is less correlated with *D* (Figure 2b), which results in less predictability of the spread of hybridization based on *D* alone.

Our comparisons of the spread of hybridization among different network topologies suggest that the relationship between hybridization rates and *D* is apt to vary with network topology. Consequently, our simulations show that differences in network topologies could at least partially account for different hybridization vs. *D* relationships observed across various river systems (Weigel *et al.* 2003; Boyer *et al.* 2008; Gunnell *et al.* 2008). Moreover, as suggested by the variance in the predictability of YCH from *D* in the dendritic network (Figure 7), the stochasticity involved in fish movement across networks with complex topologies may explain some of the scatter observed in real systems around fitted hybridization vs. *D* curves.

While variations in stocking duration influenced the YCH of single spawning grounds, they surprisingly did not affect the YCH vs. *D* relationship for the entire stream network (Figure 6). This suggests that, given our model's structure, the speed at which

hybridization will move through the network may be relatively fixed, whether or not the system continuously receives newly introduced non-native fish. However, the slope of YCH vs. *D* only measures the rate of spread of hybridization, not the magnitude of hybridization, which is undoubtedly influenced by the stocking intensity and duration.

We caution that our simulations assume that all stream segments have the same resistance to fish movement and that there is no difference in the habitat quality or carrying capacity among spawning grounds. We expect substantive changes in our model predictions should the model be changed to simulate such variability in the system. For instance, we expect that *D* would become a much poorer predictor of YCH if we simulated variability in dispersal resistance or preferred dispersal direction in stream segments across a network. Also, a decrease in the power of *D* to predict the spread of introgression across a network is expected when the fitness of fish carrying non-native genes varies with location in the network. In fact, spatial heterogeneity in fitness might yield sinks and sources of non-native genes within the network which may reduce the influence of *D* as a predictor of YCH.

Finally, for this study we limited our sample of networks to three different topologies. Investigating a larger set of networks – for example, including a broader representation of different dendritic formations – may yield new insights on the role of number and distribution of network bifurcations on the spread of introgression.

Future Directions

Regardless of their lineage (native or non-native), all modeled fish in this study were simulated with the same parameters controlling mate selection, fecundity, spawning

site fidelity, and survival. While many studies have demonstrated that this assumption is not strictly realistic (for the specific CTT – RBT case see for example Hitt *et al.* (2003), De Rito (2004), Seiler and Keeley (2007a), Boyer *et al.* (2008), Muhlfeld *et al.* (2009a), Seiler and Keeley (2009)), we maintained the assumption for this study to remove the possibility of confounding variability and because of the current lack of adequate field data that quantifies fitness and behavioral differences between native, non-native and hybrid individuals. However, our model is capable of simulating variability in fish behavior and fitness as a function of genetic composition. Specifically, mate selection, fecundity, straying propensity, and survival rates can all be simulated as a function of each fish's genotype (Section 6). These features facilitate the use of our model for future work investigating how genetically driven differences among individuals would influence the spread of hybridization. For example, we hypothesize that increased straying rates for individuals with non-native ancestry ($L_i > 0$) (e.g., Boyer *et al.* 2008) might lower the spatial variation in YCH across the networks. Under certain circumstances, for example when only few stocking events of non-native occur and individuals with $L_i > 0$ have very high straying rates, this may reduce differences in the slope of the YCH vs. D relationship among different network topologies.

Results presented here assume perfect knowledge of the non-native ancestry of each simulated fish. This assumption, however, is not strictly realistic as the genetic composition of fish – and thus their non-native ancestry – is measured in real systems using a select number of species-diagnostic markers (see for example Hohenlohe *et al.* 2011, Kalinowski *et al.* 2011). These limited genetic markers are: (a) only a sample of the

fish genome and (b) subject to Mendelian genetics and can thus under- or over-represent the non-native ancestry of a fish. As described in greater detail in Section 6 (Appendix), our model is able to simulate a user-defined number of neutral species diagnostic markers in each fish thus providing a more realistic way to describe the genome of the fish. While use of this model feature was beyond the scope of this study, it would be possible to evaluate the consequences of the number of diagnostic markers analyzed through a comparison between lineage-based model results and diagnostic-markers-based model results.

For our simulations, we used data from CTT and RBT studies to parameterize our model and investigate how hybridization spreads across a stream network. However, the three life-cycle phases described in our model (reproduction, straying, and survival) are applicable to many organisms, aquatic or terrestrial, that reproduce sexually. Similarly, the spread of introgression across a system is likely to be driven by the interactions between landscape connectivity patterns and genetically driven differences among individuals. Therefore the model we present here may provide a general species- and system-independent template for understanding gene movement and the spread of introgression across landscapes, where patterns of connection between sub-populations can be isolated and quantified.

Conclusions

Network structure and patterns of connectivity influence the nature of the spread of hybridization. Our simulations suggest that stream network topology is likely to affect

the relationship between hybridization rates and distance to the source of non-native genes. Individual sites with more interceding bifurcations along the distance to the non-native gene source are less susceptible to hybridization with non-native genes. However, among networks with similar overall stream length, bifurcations compact the network, yielding shorter average distances of sites to non-native gene sources. Model results suggest this may allow non-native genes to cover the entire network faster in networks with higher bifurcation densities. These results have implications for the study and management of introgressive hybridization in river ecosystems, and may help to explain observed differences in rates of hybridization that are based on distances alone. Integration of the concepts presented here into field studies may yield an improved understanding of how stream network context influences rates of introgressive hybridization in fish.

Acknowledgements

We thank the Center for Computational Biology at Montana State University (MSU), and especially Ivan Judson, Pol Llovet, and Sean Cleveland for their continuous support during this research. Also, we are grateful to Dr. C. Baxter, Dr. W. Cross, Dr. B. Gresswell, and Dr. G. Luikart along with two anonymous reviewers for their comments and suggestions on this manuscript. This work was supported by the U.S. NSF EPSCoR Track-2 EPS-0919557 and by the Swiss NSF grant PBSKP3_143944. (Disclaimer: “Any opinions, findings and conclusions or recommendations expressed in this material are

those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation").

Figures

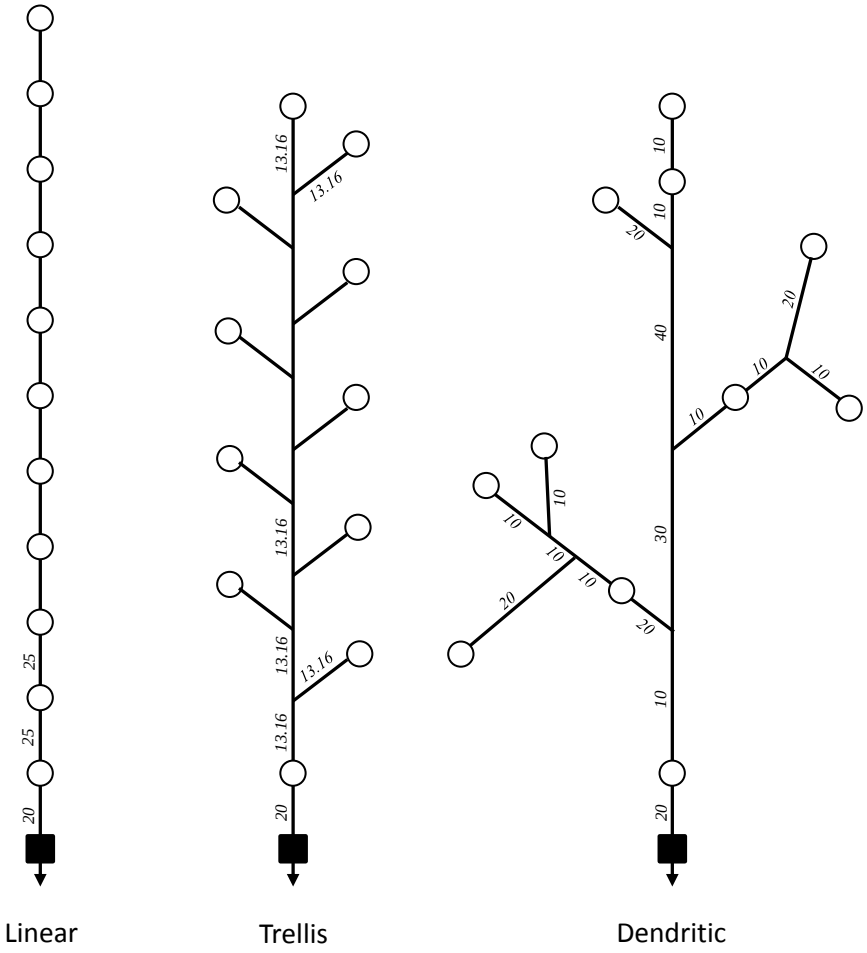

Figure 2.1: Three simulated river networks (sketches not to scale). All networks contain the same number of spawning sites and total length of stream. Spawning grounds are represented with circles, stocking site with filled squares, and stream segments with black lines. The length (km) of each stream segment is given in italics. For the linear network all stream segments are 25 km long, except for the segment connecting the stocking ground to the spawning ground at the outlet of the network which is 20 km long. For the trellis network all stream segments are 13.16 km long except for the segment connecting the stocking site to the spawning ground at the outlet of the network which is 20 km long. Stream flow direction is indicated by the arrow.

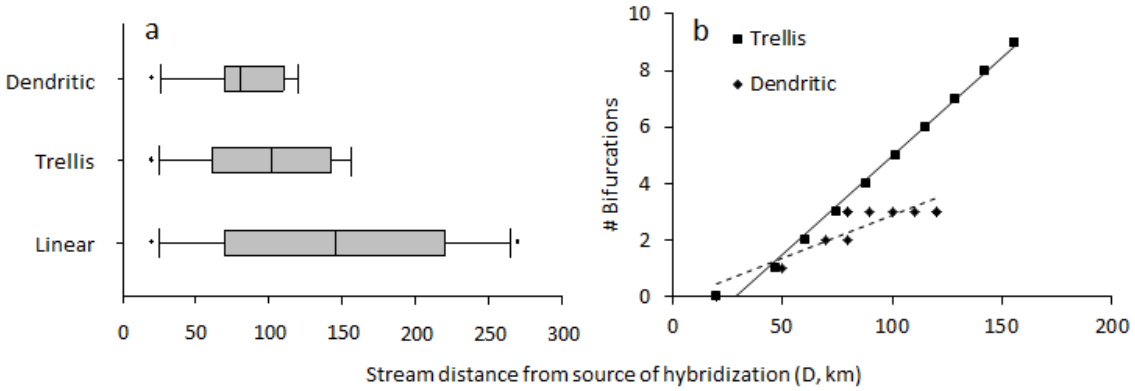

Figure 2.2: (a) Box-and-Whisker plots of the stream distance (D) to the source of hybridization (i.e., the stocking ground) for each spawning ground within each of the three simulated networks.The mid-line in the boxes is the median, the lower and the upper end of the boxes are the 25th and the 75th percentiles, the lower and upper whiskers are the 10th and the 90th percentiles. <10th and >90 percentiles are represented by the dots. (b) Number of bifurcations between stocking site and each spawning site versus the stream distance (D) between stocking site and spawning site for the networks containing bifurcations (dendritic and trellis).

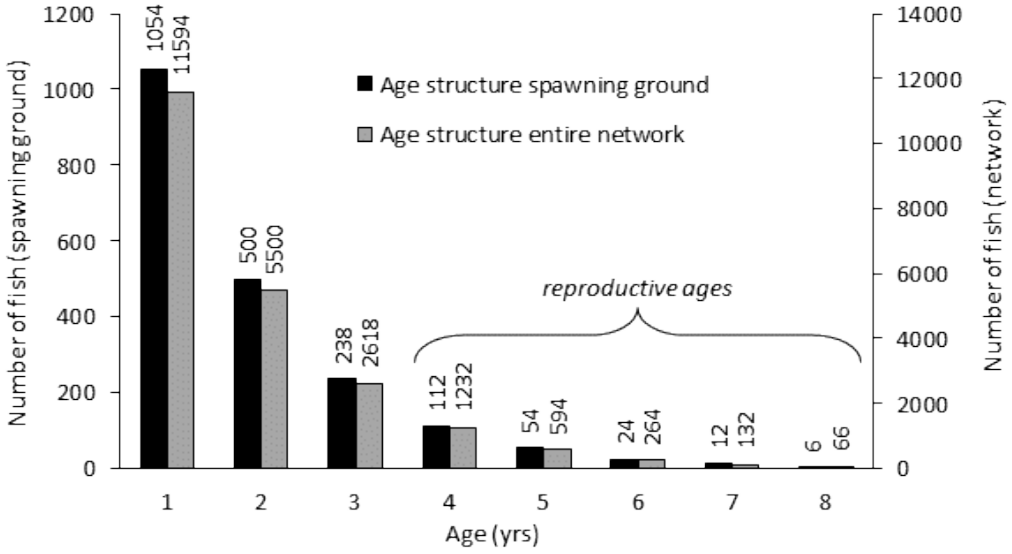

Figure 2.3: Initial population age structure (number of fish per year class) for a single spawning ground (black bars, variable over time) and population age structure for the entire network (grey bars, fixed over time).

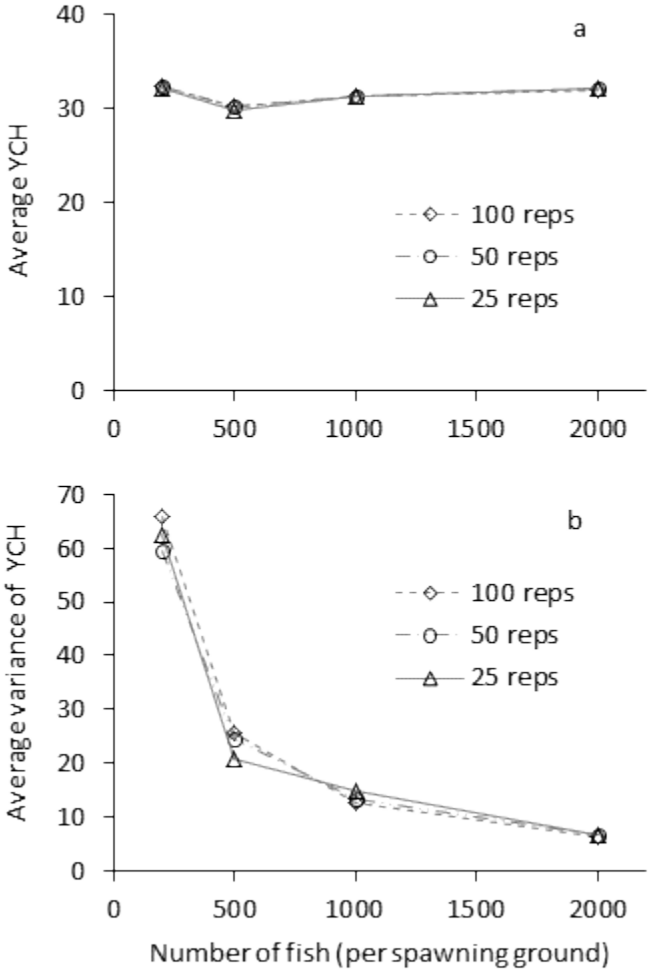

Figure 2.4: (a) Mean year of complete hybridization (YCH) and (b) mean variance of the YCH across spawning sites from replicated model runs based on simulations using 200, 500, 1000, and 2000 individuals per spawning ground. Data were generated using 25, 50, and 100 model repetitions. Variation in model results is due to the stochastic nature of the model.

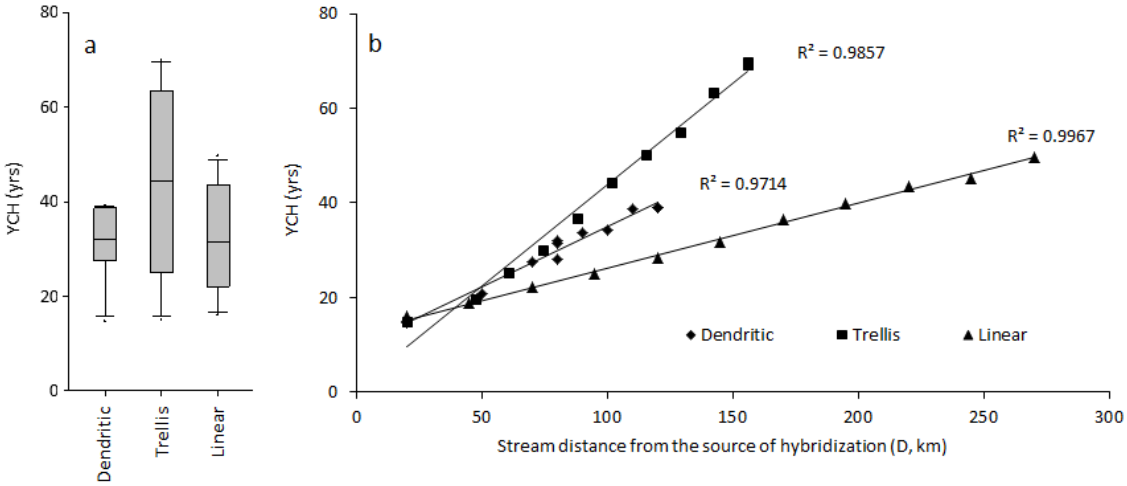

Figure 2.5: (a) Box-and-Whisker plots of the mean year of complete hybridization (YCH) across 25 replicate model runs for each of the three networks simulated. The mid-line in the boxes is the median, the lower and the upper end of the boxes are the 25th and the 75th percentiles, the lower and upper whiskers are the 10th and the 90th percentiles. <10th percentile and >90th percentile are represented by the dots. (b) Average YCH for each spawning ground versus the stream distance (D) separating the spawning ground from the source of hybridization (i.e., the stocking location) for the three simulated networks. r^2 values for the YCH vs. D relationship for the three networks are given next to each linear regression line (solid lines).

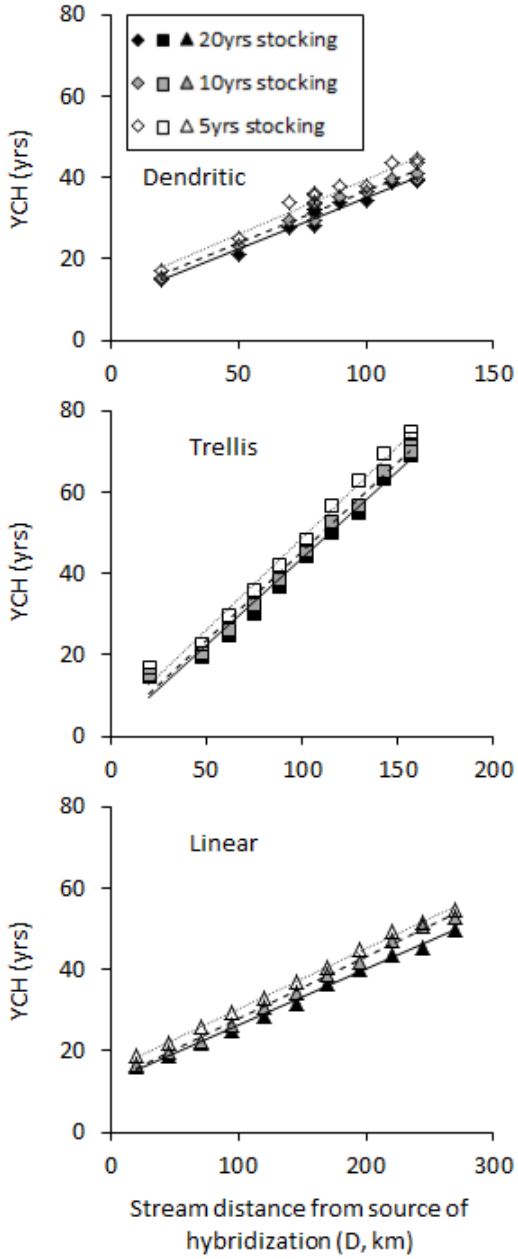

Figure 2.6: Mean year of complete hybridization (YCH) across 25 replicate model runs for each spawning ground versus the stream distance separating the spawning ground from the source of non-native genes (DTS) for the three simulated networks and the three stocking regimes simulated: 5 years (empty symbols, dotted regression line), 10 years (grey symbols, dashed regression line), and 20 years (black symbols, solid regression line).

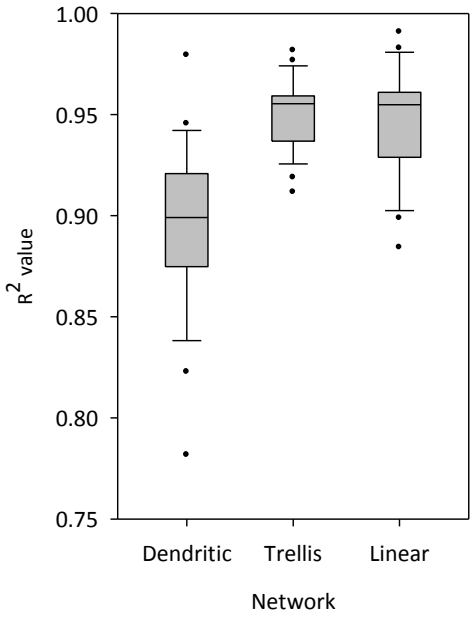

Figure 2.7: Box-and-Whisker plots summarizing the r^2 values for the YCH vs. DTS relationship for all the 25 repetitions conducted for the three simulated networks. The mid-line is the median, the lower and the upper ends of the box are the 25^{th} and the 75^{th} percentiles, the lower and upper whiskers are the 10^{th} and the 90^{th} percentiles. $<10^{th}$ and $>90^{th}$ percentiles are represented by the dots. Statistically significant differences ($p < 0.001$, from Tukey HSD multiple comparison test) among groups based on ANOVA are given by the letters above the networks' names.

CHAPTER 3

DETECTING AND QUANTIFYING INTROGRESSION IN HYBRIDIZED POPULATIONS: SIMPLIFYING ASSUMPTIONS YIELD OVERCONFIDENCE AND UNCERTAINTY

Contribution of Authors and Co-Authors

Manuscript in Chapter 3

Author: Patrick Della Croce

Contributions: Designed the study, performed simulations and analyzed data, and wrote the manuscript.

Co-author: Dr. Geoffrey C. Poole

Contributions: Assisted with study design, analyses, results discussion, co-wrote the manuscript.

Co-author: Dr. Gordon Luikart

Contributions: Assisted with study design and commented the manuscript.

Manuscript Information Page

Patrick Della Croce[1,2,#], Geoffrey C. Poole[1,3], Gordon Luikart[4]

Journal Name: Molecular Ecology Resources
Status of Manuscript:
___Prepared for submission to a peer-reviewed journal
_X__Officially submitted to a peer-reviewed journal
____Accepted by a peer-reviewed journal
____Published in a peer-reviewed journal

[1]Dept. of Land Resources & Environmental Sciences, Montana State University, P.O. Box 173120, Bozeman, MT 59717-3120
[2]Franklin University Switzerland, via Ponte Tresa 29, CH-6924 Sorengo-Lugano, Switzerland
[3]Institute on Ecosystems, Montana University System, Montana State University, 106 AJM Johnson Hall, Bozeman, MT 59717
[4]Flathead Lake Biological Station, Fish and Wildlife Genomics Group, Division of Biological Sciences, University of Montana, Missoula MT 59812

[#]Corresponding author: pdellacroce@fus.edu

Abstract

A major growing threat to the conservation of many native species worldwide is genetic introgression from non-native species. Although improved molecular genetic techniques are increasing the availability of species diagnostic markers for many species, efficient field sampling design and reliable data interpretation require accurate estimates of uncertainty associated with the detection of non-native alleles and the quantification of introgression. Using fish populations as examples, we developed a simulation model of an age-structured population that tracks the introduction and inheritance of non-native alleles across generations by simulating stochastic mating and survival of individual fish and the resulting transmission of diagnostic markers. To simulate detection and quantification of introgression, we sampled varying combinations of *n* fish and *m* diagnostic markers to detect and quantify introgression from thousands of virtual, independent fish populations for a wide range of hybridization scenarios. Using the result of simulated sampling, we quantified the extent to which common simplifying assumptions regarding population structure and inheritance mechanisms can lead to: 1) overconfidence in our ability to detect non-native alleles, and 2) unrealistically narrow confidence intervals for estimates of the proportion of non-native alleles present. Under some circumstances, commonly-used simplifying assumptions underestimate the probability of failing to detect ongoing introgression and the uncertainty in admixture estimations by orders of magnitude. Such overconfidence in our ability to detect and quantify introgression can affect critical conservation and management decisions regarding native species undergoing or at risk of introgression with non-native species.

Introduction

A large and growing number of native species worldwide are threatened by hybridization and introgression with introduced species (Rhymer & Simberloff 1996a; Kelly et al. 2010). Introgression can cause the loss of locally adapted native genotypes (Gozlan *et al.* 2010) and can eventually lead to the genomic extinction of a species (Muhlfeld *et al.* 2009a; 2014). For many native species, introgression with introduced genes has been so extensive that managers include slightly introgressed populations in protection and conservation efforts (e.g., Rhymer & Simberloff 1996a; Allendorf *et al.* 2001). For example, several states in the Western U.S. have granted protected status to introgressed populations of native cutthroat trout (*Oncorhynchus clarki* sp.) that have up to 10% admixture with introduced rainbow trout (*O. mykiss*) genes (Allendorf *et al.* 2004; Peacock & Kirchoff 2004). Consequently, for proper management and protection of native species undergoing or at risk of introgression, it is crucial for researchers and managers to be able to understand uncertainty associated with detection and quantification of admixture in populations.

Admixture in a population is ideally detected and quantified using *species-diagnostic markers* ("diagnostic markers") from a sample of individuals in the population. Diagnostic markers are specific sequences, or *loci*, in the genome of an individual that contain alleles known to be characteristic of (or fixed in) a particular species (Hohenlohe *et al.* 2011; Kalinowski *et al.* 2011). For any individual, the ratio of non-native alleles to the total number of alleles examined is assumed to represent the individual's admixture from non-native genes. The same ratio, using markers summed

across multiple individuals sampled from a single population, is assumed to represent the population proportion of admixture with non-native genes (e.g., Rasmussen *et al.* 2010; Amish *et al.* 2012). Currently, hundreds of species-diagnostic markers are becoming available to monitor introgression (e.g., Wang *et al.* 2009; Hohenlohe *et al.* 2013a; Campbell et al. 2014; Ali et al. 2015).

Sampling a population to assess introgression can fail to detect non-native alleles within the population, especially when non-native alleles are rare. The probability of failing to detect non-native alleles ($P_{n,m}$) in a population decreases as the number of individuals sampled (n) and the number of diagnostic markers assessed per individual (m) increase. Commonly (e.g., Rasmussen *et al.* 2010; Amish *et al.* 2012), $P_{n,m}$ is estimated by calculating the probability of failing to draw a non-native allele from a population, assuming a random draw of $2 \cdot m \cdot n$ samples (i.e., 2 alleles per marker) from a binomial (native vs. non-native) distribution:

$$P^E{}_{n,m} = (1 - A)^{(2 \cdot m \cdot n)} \qquad (1)$$

where $P^E{}_{n,m}$ denotes an estimate of $P_{n,m}$ derived from (E)quation 1 and A is population admixture.

Similarly, uncertainty surrounds any measure of admixture ($A_{n,m}$) – as an estimate of A – derived from a sample of n individuals and m diagnostic markers (Table 1). This uncertainty decreases with increasing n and m. Again, by assuming that $A_{n,m}$ is based on a random draw of markers from a binomial distribution, the 95% CI of A ($CI95_A$) can be estimated as:

$$CI95^E{}_A = A_{n,m} \pm 1.96 \cdot \sqrt{\frac{A_{n,m} \cdot (1 - A_{n,m})}{(2 \cdot m \cdot n)}} \qquad (2)$$

(e.g., Newcombe 1998; Kozfkay *et al.* 2007). Importantly, this estimate only incorporates sampling uncertainty. When $A_{n,m}$ is used as an estimate of A, Equation 2 ignores uncertainty due to the divergence of A and $A_{n,m}$ as marker frequency drifts over time as a result of Mendelian inheritance and stochastic mating and mortality.

Underlying Equations 1 and 2 and their applications, however, are simplifying assumptions, the effects of which on estimates of $P_{n,m}$ and $CI95_A$ are largely unquantified in the literature. Specifically, Equations 1 and 2 assume that a sample of n individuals and m diagnostic markers represents a random sample of alleles from a binomial distribution, which is true only if non-native alleles are randomly distributed among individuals and across diagnostic marker loci (i.e., when the population is a hybrid swarm). While hybrid swarms do occur in nature, populations to which Equations 1 and 2 are applied often violate such assumptions: non-native alleles are uniformly (not randomly) distributed within pure non-native individuals and F1 hybrids, and are clustered within subsets of individuals when pure non-natives and F1 hybrids are present.

Further, applications of Equation 1 (e.g., Bennett & Kershner 2009; Muhlfeld *et al.* 2009c) assume that population admixture measured using m diagnostic markers reflects the true admixture of the population (i.e., $A_m = A$ for any m). Yet, in real populations, A_m can depart from A over time because the frequencies of non-native alleles in diagnostic markers are subject to factors such as natural selection, drift, and chance events (e.g., stochastic mortality and mating). Finally, the form of Equation 2 yields symmetrical upper and lower confidence intervals, which is known to be incorrect, especially when

non-native alleles are rare or dominate the system and values of A_m approach 0 or 1 (Wilson 1927).

Using simulated fish populations as examples, here we quantify the error associated with uncertainty estimates derived from Equations 1 and 2 under a variety of circumstances that violated their underlying assumptions. Specifically, we analyze output from an individual-based simulation model of riverine fish hybridization (Della Croce *et al.* 2014) to estimate $P_{n,m}$ and $CI95_A$. By tracking the passages of individual alleles from generation to generation, based on a starting population with only pure native and pure non-native individuals, the model incorporates: (a) realistic representation of the change in distributions of non-native alleles over time, from invasion by non-natives to creation of a hybrid swarm; and (b) stochastic changes in the frequencies of non-native alleles within genetic markers over time – which lead to discrepancies between A_m and A.

We compare estimates of $P_{n,m}$ and $CI95_A$ from Equations 1 and 2 with estimates derived from sampling simulated populations to describe conditions under which Equations 1 and 2 yield marked overconfidence in our ability to detect and quantify non-native alleles in populations (underestimate $P_{n,m}$ and $CI95_A$). Moreover, we verify our approach by showing that model-based estimates of $CI95_A$ more accurately reflect published field data than estimates from Equation 2. Finally, we demonstrate that, unlike Equation 2 or estimates derived from field data, our method incorporates the known asymmetry between upper and lower bounds for $CI95_A$, which is especially important for populations with low levels of admixture.

Methods

Variable Naming Convention

We use variable names (summarized in Table 1) to precisely and concisely describe the methods and results of our study. The variable names follow a naming convention that aids in interpretation. There are 2 base variables (*P* and *A*). *P* is the probability of failing to detect non-native alleles in a population. We define *A* as the true admixture proportion of a population While we acknowledge that the term "admixture" is often best applied to hybrid swarms, we use the term broadly to describe the fraction (proportion) of all alleles that are non-native alleles in a hybridizing population, including all genetically pure (native and non-native) and hybridized individuals (e.g., all F-1s and advanced generation hybrids).

Both *P* and the reliability of estimates of *A* are dependent upon sampling strategy, specifically the number of individuals sampled from a population (*n*) and the number of diagnostic markers examined per individual (*m*). We use the subscripts "*n*" and "*m*" to indicate estimates of *P* or *A* derived from a sample of size *n* and *m*. For instance, $A_{n,m}$ denotes an estimate of *A* derived from a sample of *n* individuals and *m* diagnostic markers. Further, we refer to the uncertainty associated with $A_{n,m}$ as the 95% confidence interval of *A* as $CI95_A$. Further, our results highlight a seldom considered source of uncertainty associated with estimates of $P_{n,m}$ (see Discussion). We calculate the 95th percentile of $P_{n,m}$ values ($Pct95_{Pn,m}$) to quantify this uncertainty.

Many of the variables described above can be derived from one of three sources: (E)quations 1 or 2, our (S)imulation model, or (F)ield data. We use a superscript letter

"*E*," "*S*," or "*F*" to indicate the source of the estimate. Thus, $A^{S}_{n,m}$ indicates an estimate of population admixture sampled from a simulated population.

Finally, the true admixture of an individual (rather than a population) is denoted with lower case "*a"*. The variable a_m represents an estimate of *a* derived from a sample of *m* diagnostic markers.

Overall Approach

To estimate $P_{n,m}$ and $CI95_A$ without employing the assumptions underlying Equations 1 and 2, we used a simplified version of a published computer simulation model (Della Croce *et al.* 2014) to create thousands of virtual fish populations with a differing proportion of admixture (i.e., by varying initial *A*-values). We sampled simulated populations using varying sample sizes of individuals (*n*) and diagnostic markers (*m*) to obtain estimates of admixture ($A_{n,m}$), which were used to calculate $P^{S}_{n,m}$ and $CI95^{S}_{A}$. By comparing $P^{S}_{n,m}$ to $P^{E}_{n,m}$ and $CI95^{S}_{A}$ to $CI95^{E}_{A}$, we assessed the influence of the assumptions underlying Equations 1 and 2 on estimates of $P_{n,m}$ and $CI95_A$.

Simulation Model

Our analysis is based on multiple runs of a simplified version of a model that mimics the life-cycle of individual fish in a population and simulates the transmission of alleles from one generation to the next (Della Croce et al. 2014). The original model, implemented in the Java programming language (Oracle, Redwood Shores, CA, USA), tracked the movement of fish among multiple spawning grounds in a stream network, while simulating introgression at each spawning ground. For computational efficiency,

we re-implemented the introgression portion of the Della Croce et al. (2014) model in "R" (R Core Team 2015), so that it tracked introgression dynamics within one population that lacked immigration and emigration.

During each model run, the model: (a) creates a population comprised of a user-defined number of genetically pure native and genetically pure non-native individuals; (b) stochastically selects mating pairs; (c) creates young-of-the-year, where the genetic composition of each individual is inherited from its parents; and (d) advances individuals to next year class while maintaining a user-defined age-class structure over time. The model repeats steps (b) through (d) for a user-defined number of simulation years. Throughout the simulations, both mortality and mate choice are stochastic: the model does not incorporate natural selection or assortative mating.

The model tracks the admixture "*a*" for each individual as an exact measure of its non-native ancestry: *a* for any individual is the average *a* from its parents. At the start of a model run initial fish are either pure native (a=0) or pure non-native (a=1). *A* for the population is the mean of *a* across all individuals. Because our simulations start with pure native (a=0) and pure non-native (a=1) fish, *A* is equal to the initial ratio of non-native to native fish in the simulation run. Thus, *A* varies among simulation runs. For each individual, the model also mimics a genome by tracking 100 neutral, independent (i.e., no linkage disequilibrium) biallelic loci (each locus has 2 alleles), which represent species-diagnostic markers. Each allele can be either native or non-native and for each diagnostic marker an offspring receives one allele from its mother and one from its father. For more details on the simulation model, see Della Croce et al. (2014).

Simulations versus Equations

We developed a "simplified mode" and "realistic mode" of operation for the simulation model. The model shares the same assumptions as Equations 1 and 2 when running in the simplified mode, and relaxes the assumptions of Equations 1 and 2 when running in the realistic mode (see Table 2). Because the simplified mode reproduced results from Equations 1 and 2 (see *Results*), any differences between realistic mode vs. Equations 1 and 2 are attributable to differences in the assumptions underlying Equations 1 and 2. Further, we drew samples from our simulated populations "early" in the simulation (when the presence of pure non-natives and F1 hybrids causes more uniform distributions of non-native alleles within individuals and a clustering of non-native alleles among individuals), and "late" in the simulation (after a hybrid swarm had formed and non-native alleles were randomly distributed within and among individuals).

Simulation Scenarios

We conducted all simulations using 1000 individuals, in which we tracked 100 species diagnostic markers. Della Croce et al. (2014) showed that the stochastic nature of the model allowed single individuals to have disproportionate effects on the genetic composition of future generations (akin to a founder effect) with populations smaller than 1000. Increasing the population size beyond 1000, however, did not have an appreciable effect on model results. Thus, populations of 1000 individuals minimize the computational requirements of the model without substantial effect on modeled outcomes.

The model requires some additional parameters to run realistic model simulations. Our realistic mode simulations included 8 year classes; individuals of age 4 and older were considered reproductively mature (Table 3). Each age class had an initial sex ratio of 1:1. Age structure, maturity age, and initial sex ratio are taken from Della Croce et al. (2014) and are based on data for salmonids from Behnke (2002) and Langeland and L'Abee-Lund (1996). The initial fraction of individuals that were non-native was based on a random number (between 0 and 1, inclusive – see below) and the resulting initial non-native individuals were randomly distributed across sex and age class.

All simulations were conducted until hybrid swarms were formed (simulated populations had no more pure native, non-native, or F-1 hybrid individuals but all individuals were hybrids with $0 < a < 1$ and a random distribution of non-native alleles across diagnostic markers). Formation of a hybrid swarm required about 12 generations, so we ran simplified mode scenarios for 15 years (15 generations) and realistic mode scenarios for 60 years (about 13 generations).

We estimated $P^S_{n,m}$ and $CI95^S_A$ at two time points during the simulations. The first sampling time point (referred to as "early sampling") occurred in the year where back-crossed ("F-2") hybrids first appeared in each population (after simulation year 3 for simplified mode populations, and after simulation year 6 for realistic mode populations). Early sampling represents samples drawn from a population where non-native alleles were poorly mixed within and among individuals, which violates a key assumption of Equations 1 and 2. The second sampling time point ("late sampling") occurred at the end of the simulations, when populations were hybrid swarms. Thus, late sampling represents

samples drawn from a population where non-native alleles were well mixed within and among individuals, in accordance with the assumptions of Equations 1 and 2.

Modeled Estimates of $P_{n,m}$

To determine how $P^S_{n,m}$ varies with population admixture (A), we conducted 4000 model runs (creating 4000 virtual populations) in both the simplified and realistic modes. Each population was assigned a random initial fraction of non-native individuals between 0 and 0.4 (inclusive), which established A for each population. At both the early and late sampling time points we drew 1000 random samples for n = 10, 25, 40, and 100 individuals, and from each sample, we analyzed m = 7, 40, and 100 diagnostic markers in each collected individual. For each population, the fraction of samples where no non-native alleles were found (i.e., where $A^S_{n,m} = 0$, see below) became our estimate of $P^S_{n,m}$ for the population. Using non-linear quantile regression on the 4000 $P^S_{n,m}$ vs. A pairs (1 pair from each simulated population for each combination of n and m), we characterized the 95th percentile of $P_{Sn,m}$ ($Pct95^S_{Pn,m}$) across the various values of A from the populations. Quantile regression was performed in R using the package *quantreg* (Koenker, 2015) by fitting a function of the same shape of Equation 1 through simulation-derived $P^S_{n,m}$ values (quantile q = 0.95).

Modeled Estimates of $A_{n,m}$

To determine the uncertainty in our ability to infer a population's A given an estimate of $A_{n,m}$ within our simulated populations, we used the model to create 10,000 populations under both the simplified and realistic modes. Each population had a

randomly assigned initial fraction of non-native individuals ranging between 0 and 1 (which, again, established *A* for the population). From each population, we drew one independent sample for each factorial combination of n=10, 25, 40, and 100 and m=7, 40, and 100. We calculated $A_{n,m}$ as the average a_m across each sample. We again used quantile regression (with quantiles = 0.025 and 0.975) to characterize $CI95^S{}_A$ across populations with differing *A*.

In the quantile regression, however, we also considered that Equation 2 is based on a commonly used but unrealistic equation for estimating the 95% CI of a binomial distribution. The equation yields: 1) symmetrical upper and lower confidence estimates; and 2) upper confidence estimates above 1 or lower confidence estimates below 0 when a binomial distribution is dominated by just one value (e.g., when non-native alleles are rare in an introgressed population), both of which are incorrect. Thus, instead of equation 2, we used the function describing Wilson score interval (Wilson 1927) as the template for our non-linear quantile regression. Specifically:

$$CI95^S{}_A = \frac{1}{1+\frac{1}{n\cdot m}s^2}\left[A_{n,m} + \frac{1}{2\cdot n\cdot m}s^2 \pm s \cdot \sqrt{\frac{1}{(2\cdot m\cdot n)}A_{n,m}\left(1-A_{n,m}\right) + \frac{1}{4(n\cdot m)^2}s^2}\right] \quad (3)$$

where *s* is the parameter to be fitted. The Wilson score interval (Equation 3) yields appropriately asymmetric confidence estimates that never extend beyond the range of 0 to 1.

For each factorial combination of *n* and *m*, we created plots following the templates shown in Figures 1 and 2, which compare simulation results with values from Equations 1 and 2. We include plots for *n* = 10 and 25 and *m* = 7 and 40, run under both simplified

and realistic modes and for early and late sampling, because these smaller samples sizes showed the largest deviations between model results and Equations 1 and 2. Remaining results (with other combinations of n and m) for $P^S_{n,m}$ vs. $P^E_{n,m}$ and for $CI95^S_A$ vs. $CI95^E_A$ are presented in the Appendix. For $P^S_{n,m}$ vs. $P^E_{n,m}$, we omitted results for $n>25$ or $m>40$ because $P^S_{n,m}$ was <0.001, regardless of the population admixture.

To characterize the interactive influence of n and m on $P^S_{n,m}$ and $CI95^S_A$ for the realistic mode simulations, we calculated the average $Pct95^S_{Pn,m}$ and $CI95^S_A$ for the range of admixtures simulated ($0 \leq A \leq 0.2$ for $Pct95^S_{Pn,m}$ and $0 \leq A \leq 0.2$ for $CI95^S_A$). To obtain these averages we used the R package *pracma* (Borchers, 2015) to calculate: (a) the area under the $Pct95^S_{Pn,m}$ curve (i.e., solid line, Figure 1a) across the range $0 \leq A \leq 0.2$ for each combination of n and m ($P^S_{n,m} \approx 0$ when $A \geq 0.2$, see Results), and (b) the area contained within $CI95^S_A$ curves (e.g., between solid black lines, Figure 1b) across the range $0 \leq A \leq 1$ for each combination of n and m. We then divided the areas by the respective ranges of A thereby obtaining average $Pct95^S_{Pn,m}$ and $CI95^S_A$for each combination of n and m across the examined range of A.

Model Output vs. Published Data

We compared $CI95^S_A$ from realistic mode simulations with estimates of $CI95^F_A$ derived from published field data. Boyer *et al*. (2008) present the mean and standard deviation of a^F_m values derived from applying the model STRUCTURE (Pritchard *et al.* 2000) to diagnostic marker data sampled from 25 sub-populations of *O.clarki*, with varying levels of admixture with *O. mykiss*, in the North Fork Flathead River, Montana, USA. The mean of a^F_m estimates is reported as "*q*" by Boyer *et al*. (2008) and can be

interpreted as a field-derived estimate of population admixture ($A^F_{n,m}$). We converted the published standard deviations of "q" to standard errors (dividing by $\sqrt{n}$), and then multiplied by 1.96 to estimate $CI95^F_A$.

Boyer *et al*. (2008) distinguished between populations with non-randomly distributed non-native alleles (akin to our "early" sampling time-points) and those with randomly distributed non-native alleles (akin to our "late" sampling time-point). Thus, when comparing $CI95^F_A$ to $CI95^S_A$, we used early and late sampling estimates accordingly. The data and field sample sizes from Boyer *et al*. (2008), along with our calculated values of $CI95^F_A$ associated with each sample, are provided in Appendix 2.

Results

Methodological Validation
(Late Sampling – Simplified Mode)

Results from our simulation model were almost indistinguishable from Equations 1 and 2 when (a) the model was run using the same assumptions underlying the equations (i.e., simplified model mode, Table 2), and (b) sampling of simulated populations occurred after the formation of a well-mixed hybrid swarm (i.e., 15 generations of admixture, "late" sampling). As visible in Figure 2 (1st column), simulation-derived probabilities of failing to detect non-native alleles ($P^S_{n,m}$) closely approximated the equation-derived probabilities ($P^E_{n,m}$) for any combination of number of individuals sampled (n) and number of diagnostic markers analyzed (m) when plotted against the true proportion of admixture of the population (A, Table 1). Finite sample sizes caused expected noise (variation along the y-axis in Figure 2, 1st column) in the relationship

between $P^S_{n,m}$ and A. Such noise caused the upper boundary of the 95% confidence interval of $P^S_{n,m}$ ($Pct95^S_{Pn,m}$) to be slightly higher than $P^E_{n,m}$. Similarly, Figure 3 (1st column) shows that model-derived 95% confidence interval of A ($CI95^S_A$) plotted atop of the 95% confidence interval of A obtained from Equation 2 ($CI95^E_A$). As expected, for all combinations of n and m, the $CI95^E_A$ from Equation 2 contained 95.0% of our simulated estimates of A ($A^S_{n,m}$) (Figure 3, pie graphs, 1st column).

Realistic Populations with Well Mixed Alleles (Late Sampling – Realistic Mode)

The assumptions of the realistic mode simulations allow allele frequencies to vary over time (Table 2). Simulation model results from the realistic mode did not match values derived from Equations 1 and 2, even if non-native alleles were allowed enough time to mix well among individuals (i.e., late sampling). Specifically, changes in alleles frequencies over time yielded substantial noise (though little bias) in the modeled relationship between $P^S_{n,m}$ and A, which caused $Pct95^S_{Pn,m}$ to be higher than $P^E_{n,m}$ (Figure 2, 2nd column). Also, modeled $CI95^S_A$ were wider than their corresponding $CI95^E_A$ (Figures 3, 2nd column), and 10% to 20% of simulated $A^S_{n,m}$-values fell outside the 95% confidence interval from Equation 2 (Figure 3, pie graphs, 2nd column). The additional noise in simulation results beyond that predicted by Equation 1, arises from the divergence of A and A_m (Table 1) over time in the realistic simulation mode (which does not occur in the simplified simulation mode, Figure 4). This divergence is caused by stochastic mortality and some males mating more the once each year, both of which occur in the realistic mode of the model (Table 2).

Non-Random Distribution of Non-Native
Alleles (Early Sampling – Simplified Mode)

When samples were drawn from simulated populations under simplified mode after only 3 generations (i.e., when non-native alleles were clustered among and uniformly distributed within genetically pure non-natives and F1 hybrids; "early" sampling), simulation-derived likelihood of failing to detect non-native alleles was consistently higher than the equation-derived counterparts (Figures 2, 3rd column). The consistent bias toward higher than expected $P^S_{n,m}$-values (rather than noise in the $P^S_{n,m}$ vs. A relationship), caused $Pct95^S_{Pn,m}$ to be higher than $P^E_{n,m}$. Likewise, $CI95^S_A$ were wider than $CI95^E_A$, and 20% to 55% of simulation-derived estimates of A fell outside the equation-derived $CI95^E_A$ (Figure 3, 3nd column). Thus, Equations 1 and 2 underestimate the likelihood of failing to detect non-native genes and the size of confidence intervals around estimates of admixture when sampling populations where non-native alleles were not randomly mixed within and among individuals.

Realistic Populations with Poorly Mixed
Alleles (Early Sampling – Realistic Mode)

As expected, sampling of realistic mode simulation results early in the simulations revealed a mixture of the effects of realistic population characteristics and non-random gene distribution described above. Specifically, both $P^S_{n,m} > P^E_{n,m}$ bias, and noise in the relationship between $P^S_{n,m}$ and A were evident (Figure 2, 4th column). The bias and noise combined to yield a greater deviance between $Pct95^S_{Pn,m}$ and $P^E_{n,m}$ than for other scenarios, for any given n and m. Alike, $CI95^S_A$ was considerably wider than $CI95^E_A$ (Figure 3, 4th column); fully 40% - 70% of simulated $A^S_{n,m}$-values fell outside $CI95^E_A$ (pie

graphs, Figure 3). Thus, the combined assumptions of (a) well mixed alleles and (b) unvarying alleles' frequencies over time, implicit in Equations 1 and 2, yielded substantial under-estimates of $P_{n,m}$ and $CI95_A$.

Difference Between Simulation- and Equation-derived 95% CI of A

Plots of the difference between $CI95^S_A$ and $CI95^E_A$ vs. $A_{n,m}$ (Figure 5) highlight an expected asymmetry (Wilson 1927) in upper vs. lower $CI95^S_A$ (Figure 3). Because Equation 2 does not accurately represent this asymmetry, the differences between upper $CI95^S_A$ and upper $CI95^E_A$ were large (relative to population admixture rates) when $A_{n,m}$ was approximately <0.2 (Figure 5), suggesting that Equation 2 systematically under-predicts the upper boundary of the 05% confidence interval of *A* when admixture with non-native genes is low.

Sensitivity to Variations in *n* and *m*

Increasing the number of diagnostic markers sampled (*m*) did not appreciably reduce the probability of failing to detect non-native alleles, nor did it narrow the 95% confidence interval associated with observed proportions of admixture for early sampling. Figure 6a shows that sampling a larger number of diagnostic markers (*m*) did not reduce the probability of failing to detect non-native alleles for early sampling, when non-native alleles were uniformly distributed among loci within pure-native and F1 hybrids. Similarly, Figure 6b shows that sampling more diagnostic markers did not yield a reduction in the cumulative uncertainty associated with $A_{n,m}$ for early sampling.

Model Output vs. Published Data

Our comparison of $CI95^{S}_{A}$ and $CI95^{E}_{A}$, with estimates of $CI95_{A}$ derived from field data ($CI95^{F}_{A}$) yielded two important conclusions. First, as visible in Figure 7, for populations with intermediate to high levels of admixture (roughly $0.2 < A_{n,m} < 0.8$), our simulated confidence intervals ($CI95^{S}_{A}$) are closer estimates of field observations ($CI95^{F}_{A}$) than the equation-derived $CI95^{E}_{A}$. Indeed, $CI95^{E}_{A}$ consistently underestimates the uncertainty associated with field observations. Second, for lower levels of admixture ($A_{n,m} < 0.2$), neither $CI95^{F}_{A}$ nor $CI95^{E}_{A}$ reflects the expected asymmetry between the upper and lower boundaries of $CI95_{A}$. Consequently, the upper boundaries of $CI95^{F}_{A}$ and $CI95^{E}_{A}$ are both substantially lower than those of $CI95^{S}_{A}$ (Figure 7).

Discussion

Simulation Results Match Equations 1 and 2 When Run Using the Same Assumptions

Results from simulated hybrid swarms in which allele frequencies were maintained constant over time (i.e., simplified mode (Table 2), late sampling) agree with predictions from Equations 1 and 2 (Figures 2 and 3, 1st column). The minor noise observed in the relationship between simulation-derived probability of failing to detect non-native alleles ($P^{S}_{n,m}$) and true population admixture (A) obtained from these simulations (Figure 2, 1st column) is expected because $P^{S}_{n,m}$ estimates are derived from finite sampling of finite populations. However, for any given A, this noise was small enough so that the upper boundary of the 95% confidence interval of $P^{S}_{n,m}$ ($Pct95^{S}_{Pn,m}$) was largely equal to the probability of failing to detect non-native alleles predicted by

Equation 1 ($P^{E}_{n,m}$). Similarly, the minute differences between simulated- and equation-derived 95% confidence intervals of *A* ($CI95^{S}_{A}$ and $CI95^{E}_{A}$ respectively, Figure 3, 1st column) are likely caused simply by finite sampling and by our use of Equation 3 rather than Equation 2 in the quantile regression. Based on the small magnitude of such differences, and following the approach of Epperson *et al.* (2010), we conclude that deviance between simulation predictions and Equations 1 and 2 for our other scenarios (realistic/late, simplified/early, and realistic/early) represents the effect of relaxing the underlying and often unrealistic assumptions of Equation 1 and 2 on values of $P_{n,m}$, $Pct95_{Pn,m}$, and $CI95_{A}$.

Assumptions of Random Gene
Mixture Impart Bias in $P^{E}_{n,m}$ and $CI95^{E}_{A}$

In our simulations, populations sampled at the early stages of introgression (i.e., sampling occurred only 3 generations after the introduction of non-native individuals) are dominated by pure native, pure non-native, and first-generation (F-1) hybrid individuals. In these populations, non-native alleles are clustered within few individuals (i.e., non-natives and F-1 hybrids) and uniformly distributed across the diagnostic markers of introgressed individuals (e.g., in F-1 hybrids each diagnostic marker carries a native and a non-native allele). This clustered (among individuals) and uniform (within individuals) distribution of non-native alleles violated the assumption of randomly distributed alleles underlying Equations 1 and 2. As a result, Equation 1 and 2 consistently underestimated $P_{n,m}$ and $CI95_{A}$ for early sampling under both simplified and realistic population assumptions (Figures 2 and 3, 3rd and 4th columns). Specifically, (a) the decreased chance

of sampling individuals carrying non-native alleles made the detection of introgression consistently more difficult than expected from Equation 1 (i.e., $P^S_{n,m}$ and $Pct95^S_{Pn,m}$ higher than $P^E_{n,m}$, Figure 2, 3rd and 4th columns) and (b) simulated samples had high variance in a_m which caused $CI95^S_A$ to be higher than $CI95^E_A$ (Figure 3, 3rd and 4th columns).

Assumption of Constant Alleles Frequencies Hide Uncertainty Caused by Divergence of A_m and A

In realistic simulation mode, the assumptions of Mendelian inheritance, stochastic survival, and the potential for multiple mating partners for males caused stochastic variation (loss or amplification) of non-native allele frequencies over the course of the 13-generations (60 year) simulations. As a result, estimates of population admixture using a finite number of diagnostic markers (A_m) drift away from the true population admixture (A), especially for small values of m (Figure 4). Such differences between A_m and A in a population create additional noise in the $P^S_{n,m}$ vs. A relationship that increases with time (Figure 2, 2nd vs. 4th columns). As a result, realistic population assumptions yield a distribution of $P^S_{n,m}$-values for populations with a given A. Therefore, for realistic mode, use of $Pct95^S_{Pn,m}$ rather than mean estimates of $P^S_{n,m}$ accounts for associated uncertainty in $P^S_{n,m}$ for any given A.

Assumptions of Symmetrical Upper and Lower CI95A Understate Uncertainty

Failure to consider the asymmetry in upper and lower limits of $CI95_A$ caused an especially large discrepancy between the simulation- and equation-based 95% confidence

intervals of A ($CI95^{S}_{A}$ and $CI95^{E}_{A}$) for low and high values of $A_{n,m}$ (Figure 3, 2nd and 4th columns). Visually, the difference between simulated- and equation derived uncertainty is small for late sampling (Figure 3, 2nd column). However, even for late sampling, the difference between the upper $CI95^{S}_{A}$ and $CI95^{E}_{A}$ can be a large proportion of $A^{S}_{n,m}$ at low levels of admixture ($A_{n,m} < 0.2$, Figure 5, left panel). For instance, at $A_{10,7} = 0.1$, the difference between the upper boundary of the simulation-derived $CI95_A$ and the upper boundary of the equation-derived $CI95_A$ (i.e., upper $CI95^{S}_{A}$ – upper $CI95^{E}_{A}$) is more than 20% of $A_{10,7}$ while at $A_{10,7} = 0.005$ the same difference is >600% of $A_{10,7}$.

The comparison of $CI95^{E}_{A}$ and $CI95^{S}_{A}$ against field data derived 95% confidence intervals ($CI95^{F}_{A}$, Figure 7) contains a paradox that can be resolved by considering the expected asymmetry (Wilson 1927) in upper and lower $CI95_A$ bounds. When admixture is low (roughly $A_{n,m} < 0.2$), Equation 2 predicts 95% confidence intervals that are very similar to field data-derived $CI95^{F}_{A}$, while our simulated $CI95^{S}_{A}$ values have higher upper boundaries. Yet our simulation results closely agree with $CI95^{F}_{A}$ at moderate to high levels of admixture (roughly $A_{n,m} > 0.2$), while Equation 2 underestimates $CI95^{F}_{A}$. We resolve this paradox by noting that: (a) Equation 2 and the model STRUCTURE (Pritchard *et al.* 2000) – from which $CI95^{F}_{A}$ values are derived (Boyer *et al.* 2008) – both yield symmetrical estimates of the upper and lower bounds for $CI95_A$; and (b) the expected asymmetry (Wilson 1927) will be most pronounced at low levels of admixture.

From this, we draw the following conclusions. First, the agreement between $CI95^{S}_{A}$ and $CI95^{F}_{A}$ at intermediate to high levels of admixture ($A_{n,m} > 0.2$) suggest that our modeling approach is providing reliable estimates of $CI95_A$. Further, for lower levels

of admixture ($A_{n,m} < 0.2$), we argue that the assumption of symmetrical confidence intervals substantially and artificially limits the upper confidence bound for both $CI95^{E}_{A}$ and $CI95^{F}_{A}$. Consider that, by definition, the lower confidence interval is constrained between $A_{n,m}$ and zero. If we assume that the upper confidence interval must be equal to the lower, then the upper confidence interval will be artificially constrained as $A_{n,m}$ approaches zero, even though much larger values of $A_{n,m}$ are randomly sampled from simulated populations with low levels of admixture. We conclude, then, that the lack of agreement between the upper limits of $CI95^{F}_{A}$ and $CI95^{S}_{A}$ at low levels of admixture (Figure 7) does not provide evidence that our simulation results are unreliable for low levels of admixture. Rather, we believe that $CI95^{F}_{A}$ does not provide a reliable benchmark for comparison at low levels of admixture because of the underlying assumption of symmetrical upper and lower $CI95_{A}$. Thus, although we have no benchmark for demonstrating the absolute accuracy of our results for low levels of admixture, our results call into question any estimate of $CI95^{F}_{A}$ that assumes symmetrical confidence intervals, especially for low levels of admixture.

Increasing m Might Not Improve Detection or Quantifications of Early Introgression

In our simulated populations, early stages of introgression are dominated by pure native, pure non-native, and F-1 hybrid individuals. For each of these genotypes, examination of just one diagnostic marker ($m = 1$) provides a precise and accurate measure of A for each individual (0.0 for natives, 1.0 for non-natives, and 0.5 for F-1 hybrids). Thus, increasing m had little effect on the probability of detecting introgression

or on the uncertainty associated with estimates of introgression level (Figure 6). This is in contrast with Equations 1 and 2, which assume that any sample of *n* individuals and *m* diagnostic markers is a random sample of the gene pool, and therefore that $P^{E}_{n,m}$ or $CI95^{E}_{A}$ react equally to variations in *n* or *m*. In Figure 6, late sampling shows the pattern expected from Equations 1 and 2 – the probability of failing to detect introgression and the uncertainty surrounding estimates of introgression is similar for *n*=100, *m*=40 and for *n*=40, *m*=100. Yet for early sampling, when pure native, pure non-native, and F-1 hybrid individuals dominate the population, *n*=100, *m*=40 yields a much lower probability of failing to detect introgression and a much lower uncertainty around estimates of introgression level than does *n*=40, *m*=100. Importantly, however, our results are less applicable to the situation where introgression of a native population is driven by immigration of F-2 hybrids (rather than pure non-native fish, as assumed in our simulations). Where F-2 hybrids are the predominant carriers of non-native genes early in the process of introgression, increases in *m* may still yield modest improvement in the detection and quantification of introgression because, if an F-2+ individuals is captured, increasing *m* will improved the likelihood of detecting non-native alleles in the individual and quantifying its admixture.

Implications and Recommendations for Management and Conservation

Comparison of our model results to values derived from Equations 1 and 2 and to results from field data reveals several mechanisms that cause deviations between the results of our simulation approach and results from other approaches. First, the non-

random distribution of non-native alleles associated with early sampling causes Equation 1 to exaggerate the ability of a sampling strategy to detect non-native alleles (Figure 2). This also causes Equation 2 to understate the uncertainty associated with $A_{n,m}$ as an estimate of A (Figure 3). Second, the realistic representation of populations (e.g., age structure, random mortality, and Mendelian inheritance rules) yields non-conservative diagnostic markers which cause A_m to diverge from A over time. Consequently, simulation results show noise in the relationships between $P^S_{n,m}$ vs. A (Figure 2) and $CI95^S_A$ vs. $A_{n,m}$ (Figure 3) which is not reflected in Equations 1 and 2. Third, failure to represent the expected asymmetry in upper and lower $CI95_A$ values likely causes both Equation 2 and field data to understate the upper $CI95_A$.

Non-random gene distribution also appears to render increases in m (the number of diagnostic markers analyzed) an ineffective approach to reducing the probability of failing to detect non-native alleles ($P_{n,m}$) or the uncertainty surrounding estimates of admixture ($CI95_A$) when populations are dominated by pure and F-1 individuals (Figure 6).

When considering strategies for sampling real populations, then, we expect that bias in $P^E_{n,m}$ and $CI95^E_A$ is likely to be largest early in the process of introgression, yet will decreases over time as alleles mix among and within individuals. In contrast, divergence of estimated admixture ($A_{n,m}$) from true A may be smallest early in the process of introgression, and increase over time as alleles are lost or magnified as a result of Mendelian inheritance rules and stochastic mating and survivorship. Such divergence has the effect of increasing $Pct95_{Pn,m}$ and $CI95_A$. Thus, our results show that the sources of

uncertainty may shift over time, driven at first by non-random gene distribution and later by changes in allele frequency. This change in the drivers of uncertainty over time has implications for sampling design and the interpretation of admixture data.

Accordingly, our first recommendation is that managers and researchers working to detect and quantify non-native alleles in a population allocate more resources toward sampling more individuals rather than toward the analysis of more diagnostic markers, especially in systems where introgression is likely to be in its early stages (especially when > 5-10 diagnostic loci are available). Such an initial sampling strategy will provide the best picture of the distribution of individuals among genotype classes (pure, F-1, or F-2+). Armed with this information, educated decisions can be made about the number of diagnostic markers to use as sampling of the population continues over time. The greater the number of F-2+ individuals, the more advantageous will be an increase in the number of diagnostic markers. This recommendation stands somewhat in contrast to what may be a somewhat dogmatic confidence in the benefits associated with the use of large *m*. Where it exists, any such dogma is especially worrisome with respect to our finding that detection and quantification of introgression may not benefit from increased *m* early in the introgression process, which is arguably the most important period for efforts to conserve native species. Specifically, studies and monitoring programs relying on Equation 1 to design sampling strategies to achieve a certain confidence of detecting non-native alleles in a population (e.g., Bennett & Kershner 2009; Muhlfeld *et al.* 2009c) could have an overly optimistic confidence in their sampling procedures for any populations where invading non-natives are pure or F-1 individuals.

Additionally, our results document how the fluctuation of allele frequencies can create heretofore undescribed variation in the probability of failing to detect non-native alleles ($P_{n,m}$) for a population of a given *A*. Thus we introduce the concept of $Pct95_{Pn,m}$ (as the upper boundary of the 95% confidence interval of $P^{S}_{n,m}$) as a means of incorporating this usually unaccounted source of uncertainty. Especially in populations where low levels of admixture might have been present for long periods of time, an extra measure of caution represented by the adoption of $Pct95_{Pn,m}$ rather than $P_{n,m}$ may be warranted in designing a sampling scheme to ensure a desired likelihood of detecting introgression in a population with a given *A*.

Also, in agreement with (Chikhi *et al.* 2001), our investigations highlight the importance of recognizing the uncertainty associated with observed values of admixture ($A_{n,m}$)and show that the same $A_{n,m}$-value can be derived from populations with a range of different true *A* (Figure 3). Therefore, observed $A_{n,m}$ values can only be interpreted correctly if the associated uncertainty is not only recognized, but also reliably estimated. Specifically, methods that ignore the complexity of real populations or that erroneously assume symmetric confidence intervals will not provide reliable estimates of $CI95_A$.

Unwarranted confidence in the precision or reliability of a given datum (e.g., $A_{n,m}$ or $P_{n,m}$) is apt to hinder the development of effective conservation and management plans for native species threatened by introgression with non-native genes. Therefore reliable estimates of uncertainty are important. Our modeling approach simulates relatively realistic populations, but still represents an analysis of a simplified system. For instance, our model ignores selection for or against non-native alleles and assortative mating,

which have been observed with introgressing species (e.g., De Rito 2004; Muhlfeld *et al.* 2009b; Hohenlohe *et al.* 2013b). Both selection and assortative mating are likely to increase the variance and uncertainty in $P_{n,m}$ and $A_{n,m}$. Similarly, potentially strong bias may result if non-diagnostic alleles are included as presumed diagnostic alleles (Hand et al. 2015). Thus, our estimates of $Pct95_{Pn,m}$ and $CI95_A$ may still understate the true uncertainty in determining the probability of failing to detect non-native alleles and the true uncertainty surrounding estimates of admixture. Yet, despite its limitations, results from our model improve upon past work and are free from many of the simplifying assumptions associated with Equations 1 and 2.

The research presented here highlights the problems associated with using Equations 1 and 2 to estimate $P_{n,m}$ and $CI95_A$, yet we could be accused of documenting a problem but offering no better alternative. We have derived equations to estimate $P_{n,m}$ that account for non-random distributions of non-native alleles among individuals and across diagnostic markers within individuals (Della Croce *et al.* submitted, *Chapter 4*). These new equations estimate $P_{n,m}$ reliably in the early stages of introgression (i.e., for non-hybrid swarms) when detection is most critical. The structure of Equation 3, however, combined with our observed trend of smaller $CI95_A$ over time, as genes mix within populations, makes derivation of equations for estimates of $CI95_A$ far less straight-forward. Figures 3 and 5 (along with Appendix 1) provide a sense of the error in Equation 2, but inherent in these results are assumptions of age class distributions, age of sexual maturity, mortality rates, etc. that are appropriate for trout populations, yet may

not be appropriate for other organisms. Thus, a general method for providing accurate estimates of $CI95_A$ remains open to additional research.

In conclusion, introgression between native and non-native species is cause of growing concern among conservation biologists In many cases the long-term conservation of native species can only be successful if both non-introgressed and introgressed populations are efficiently monitored and managed for admixture (Peacock & Kirchoff 2004; Schwartz *et al.* 2007). To do this, it is crucial to reliably detect and quantify introgression in populations. In agreement with Epperson *et al.* (2010), this study suggests that computational genetic methods that abide by unrealistic assumptions may mislead managers and researchers dealing with native species threatened by introgression from non-native genes. Specifically, use of Equations 1 and 2 likely yield (a) overly optimistic confidence in the ability to detect hybridization, and (b) will underestimate the uncertainty around estimates of admixture from non-native genes in a population. Under many circumstances, the uncertainty is underestimated by orders of magnitude. Such outcomes can mislead the design of sampling strategies and can even yield misclassified conservation status for populations. Improved methods, such as ours, for estimating uncertainty in our ability to detect and quantify non-native alleles in populations will facilitate conservation by ensuring that management decisions and study designs are based on more reliable information.

Acknowledgments

We are grateful to Dr. R. Payn, Dr. C. Baxter, Dr. W. Cross, and Dr. B. Gresswell for their comments and suggestions on this manuscript and to Dr. M. Boyer for helpful comments and support with this work. This work was supported by the U.S. NSF EPSCoR Track-2 EPS-0919557 and by the Swiss NSF grant PBSKP3 143944. (Disclaimer: "Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation"). Gordon Luihart was supported in part by U.S. National Science Foundation grants DEB-0742181 and DEB-1067613, NASA-Roses-2013 (Ecological forecasting), and Bonneville Power Administration grant #199101903 to Montana Fish Wildlife and Parks.

Tables and Figures

Table 3.1: Overview of the variables described in this study. The column "Level" indicates if the variable refers to an individual or to a population.

Variable	Description	Level
n	Number of individuals sampled from a population	Sample
m	Number of diagnostic markers examined from each individual in a sample of n individuals.	Sample
a	The true proportion of admixture with non-native alleles of an individual. Represented as a real number between 0 (pure native) and 1 (pure non-native). For each offspring produced in the model, calculated as the average a of the offspring's parents.	Individual
a_m	Observed admixture of an individual based on m diagnostic markers.	Individual
A	True admixture of the population. We calculate this value assuming Hardy-Weinberg equilibrium by averaging a across all the individuals of the population.	Population
A_m	Estimate of admixture (A) based on analysis of m diagnostic markers across all individuals in a population.	Population
$A_{n,m}$	Estimate of admixture (A) based on sampling n individuals and m diagnostic markers.	Population
$CI95_A$	95% confidence interval of A.	Population
$CI95^E_A$	Estimate of $CI95_A$ derived from Equation 2.	Population
$CI95^S_A$	Estimate of $CI95_A$ derived from simulation results.	Population
$CI95^F_A$	Estimate of $CI95_A$ derived from published field data.	Population
$P_{n,m}$	Probability that non-native alleles, when present, will not be detected in the population (false negative) given a sampling of n individuals and m diagnostic markers.	Population
$P^E_{n,m}$	Estimate of P derived from Equation 1 using a sample size of n individuals and m diagnostic markers.	Population
$P^S_{n,m}$	Estimate of P derived from resampling simulation results using a sample size of n individuals and m diagnostic markers.	Population
$Pct95^S_{Pn,m}$	Upper boundary of the 95% confidence interval of $P_{Sn,m}$ derived from simulation results.	Population

Table 3.2: Differences in model assumptions between simplified and realistic model modes. The assumptions of simplified mode are the same as those underlying Equations 1 and 2, and cause the frequency of non-native alleles in a population to remain constant over time. In realistic mode, Mendelian rules of inheritance, stochastic mortality, and stochastic mating allow for the stochastic loss and amplification of alleles during the course of a simulation run.

	Binomial Mode	**Realistic Mode**
Population Structure	Population has only one age class.Each simulation year, all parents are replaced with their offspring.	Population has multiple age classes of constant size. Mortality occurs by removing fish randomly as fish move up in age class.
	Individuals have no gender.	Gender ratio in age classes fluctuates due to random mortality.
Mating	All individuals are sexually mature and can mate.	Immature individuals (in age classes below a user-define age) do not mate.
	Number of mating pairs is equal to half the population size.	Number of mating pairs is equal to number of sexually mature females.
	Pairs are formed at random (though one cannot mate with itself) with no consideration of gender; individuals can mate more than once.	Pair formation is random between genders; as with binomial, and in salmonids populations individuals can mate more than once (e.g., Araki et al. 2007).
Reproduction and Mortality	Each mating pair produces two offspring; there is no stochastic effect of mortality.	Each mating pair produces a user-defined number of offspring to ensure an excess of young-of-the-year fish; gender of each offspring is assigned randomly (1:1 genders' ratio is not enforced in offspring population)
Inheritance	Each parent passes half of its alleles (randomly selected) to one offspring and the other half to the other offspring, thereby limiting random changes to native and non-native allele frequency.	Loci are assumed to be independent and inheritance of alleles is dictated by Mendelian rules of inheritance.

Table 3.3: Number of individuals per year class in realistic mode simulations.

	Immature			**Mature**				
Age class	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
# of individuals	526	250	118	56	28	12	6	4

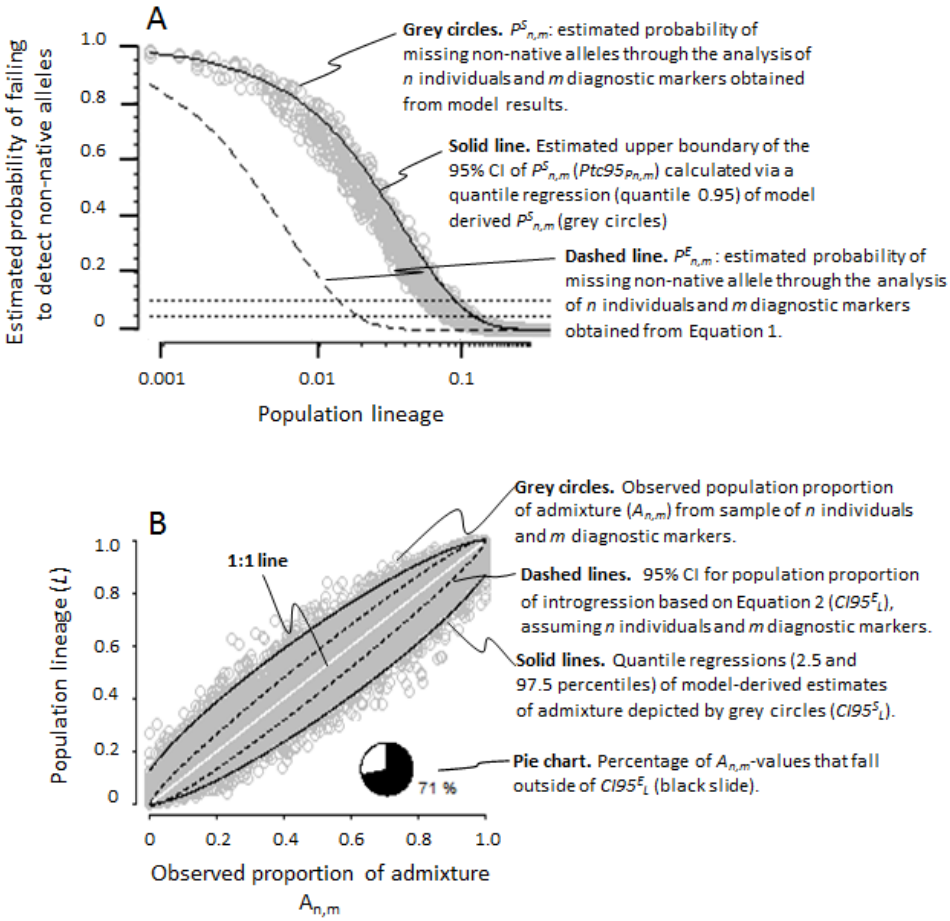

Figure 3.1: **(A)** Example plot comparing simulation- and equation-derived estimates of probability of missing non-native alleles (false negative) and the associated uncertainty for a given simulation scenario (i.e., combination of model mode and sampling time, see *Methods*) and a given sampling strategy (*n* individuals and *m* markers). When simulation results agree with Equation 1, model estimates (grey circles) and model-derived upper boundary of the 95% CI (solid line) will overlap with equation estimates (dashed line). Where equation estimates are too optimistic, model-derived estimates and upper boundary of the 95% CI will lay to the right of the equation estimates (as shown in this example).**(B)** Example plot comparing simulation- and equation-derived uncertainty associated with estimates of population admixture for a given simulation scenario and sampling strategy. When simulation results agree with Equation 2, quantile regressions of simulation results (solid lines) will track equation-derived 95% CI (dashed lines), and roughly 5% of simulation-derived estimates of population admixture (grey circles) will fall outside of the equation-derived 95% confidence interval (pie chart). Where Equation 2 is overly optimistic, quantile regression will be wider than equation-derived 95% CI and more than 5% of model derived estimates will fall outside of equation-derived 95% CI (as shown in this example).

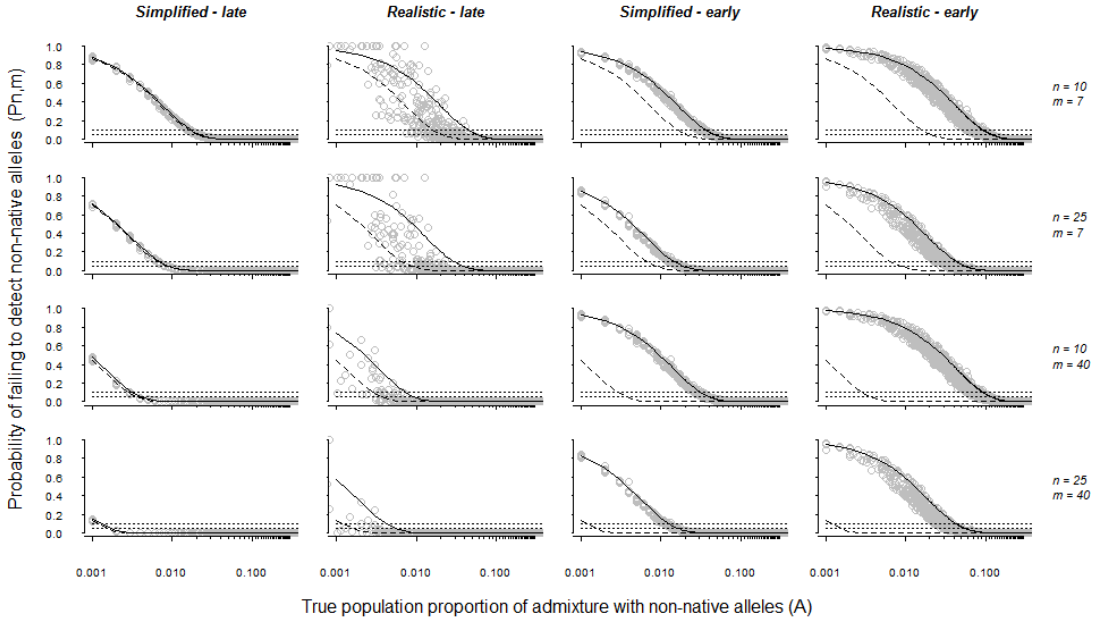

Figure 3.2: Simulation-derived probability of failing to detect non-native alleles (false negative) in a population through the analysis of n individuals and m diagnostic markers (PSn,m, grey circles) and upper boundary of the associated 95% confidence interval (Pct95SPn,m, solid line), along with PEn,m from Equation 1 (dashed line) as a function of the true proportion of admixture (A). Horizontal dotted lines represent Pn,m = 0.05 and 0.1. Shown in the plot are only the populations for which Pn,m>0 (approximately A<0.4).

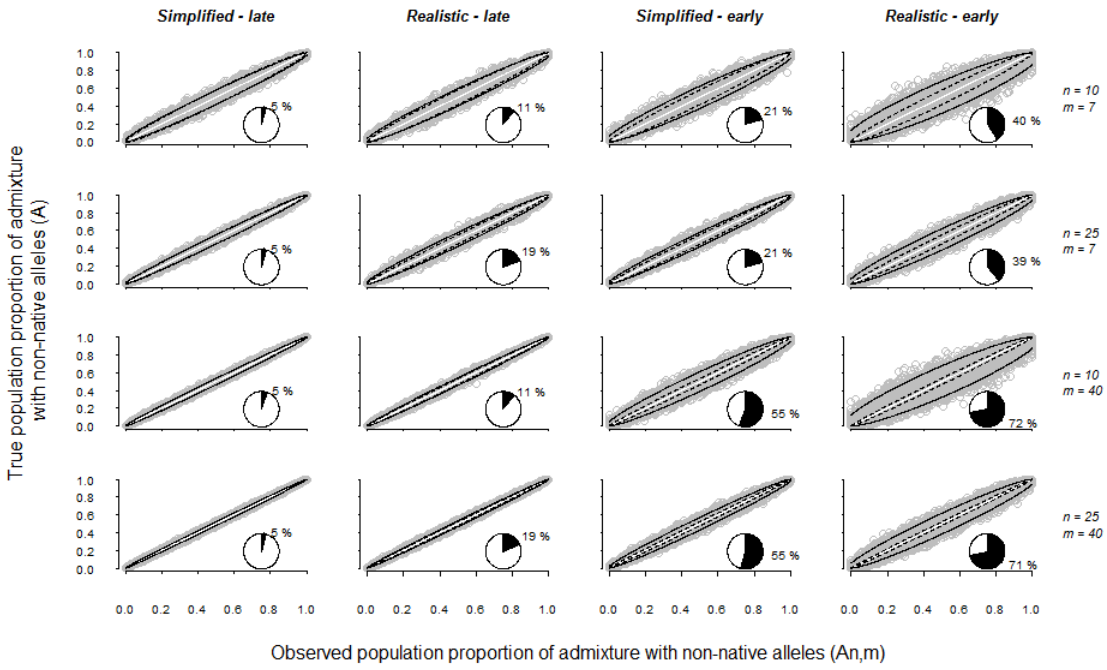

Figure 3.3: Simulation- (solid line) and equation-derived (dashed line) uncertainty (CI95A) associated with observed proportions of population admixture (An,m, grey circles) for selected combinations of individuals sampled (n) and markers analyzed (m). All other combinations of n and m are reported in the Appendix 1a-b.

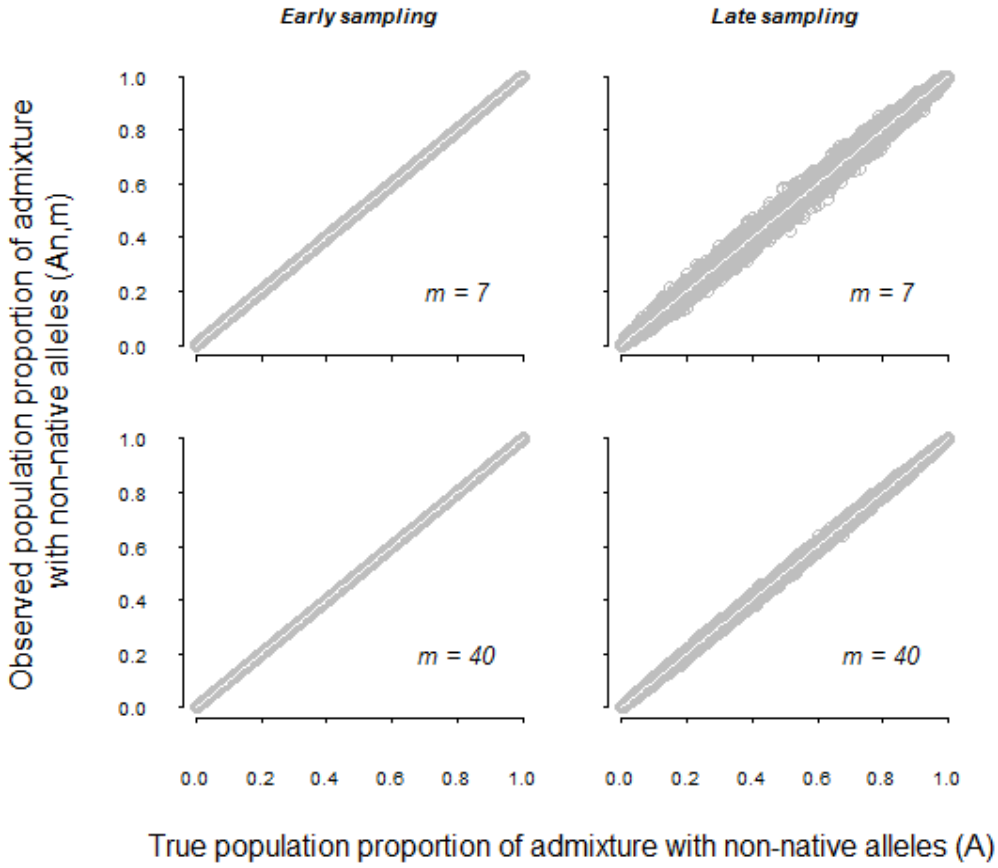

Figure 3.4: Population proportion of admixture (A_m, grey circles) of realistic mode populations versus true proportion of admixture (A) for m=7 and m=40 showing that under realistic mode assumptions A_m departs from A over time. Left column: early sampling (3 generations). Right column: late sampling (13 generations).

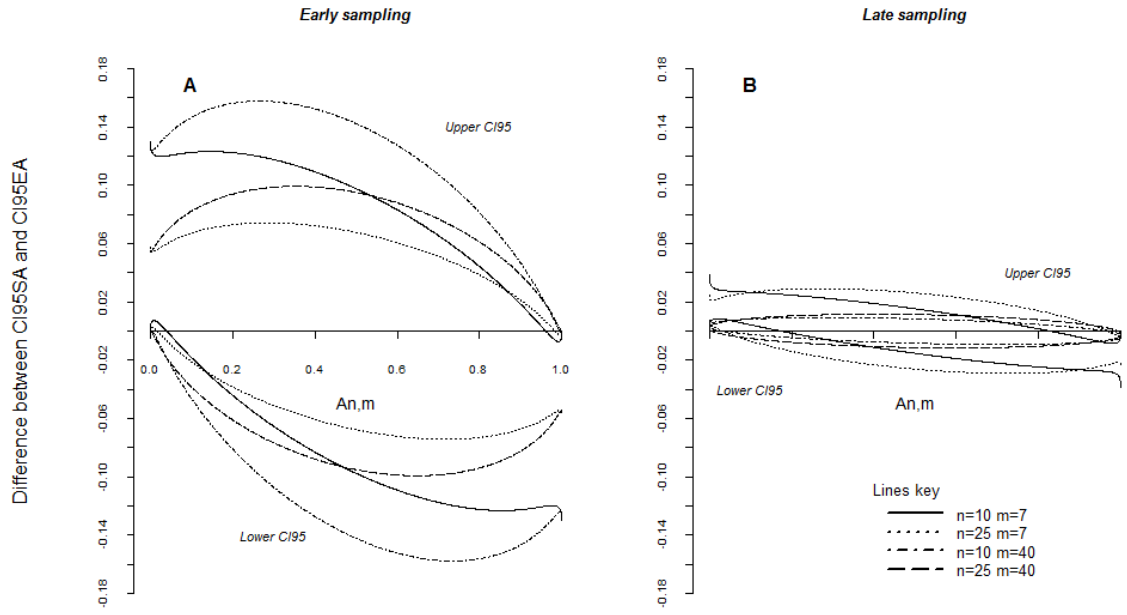

Figure 3.5: Difference between simulation- and equation-derived 95% confidence intervals of true proportion of admixture (i.e., CI95SA – CI95EA) from realistic mode populations. Left column: early sampling (3 generations). Right column: late sampling (13 generations).

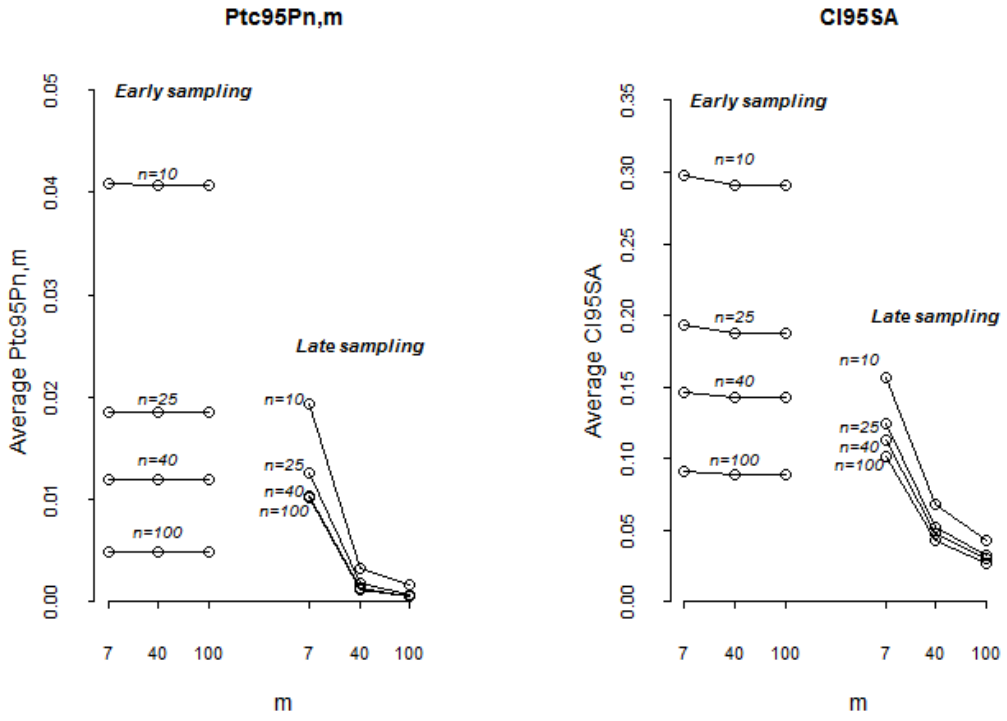

Figure 3.6: From populations simulated under the realistic model mode and all combinations of n and m used: (left) average Pct95SPn,m for $0 \leq A \leq 0.2$ and (right) average CI95SA for $0 \leq A \leq 1$ for early sampling and late sampling (3 and 13 generations respectively).

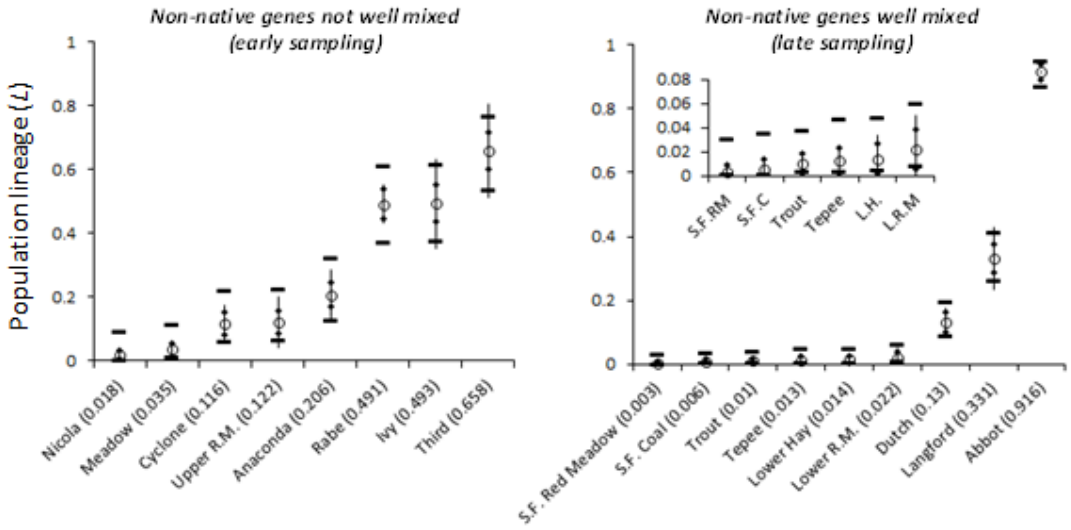

Figure 3.7: Proportions of population admixture from 7 microsatellite loci from Boyer et al. (2008) (empty circles) along with three different associated uncertainties (CI95A) derived from: (a) data reported in Boyer et al. (2008) shown with vertical lines (CI95FA); (b) Equation 2 shown as black diamonds (CI95EA), and; (c) realistic mode populations shown as horizontal bars (CI95SA). Data from introgressed populations that according to Boyer et al. (2008) showed a non-random distribution of non-native alleles were compared with CI95SA from early sampling (left panel). Data from introgressed populations that according to Boyer et al. (2008) showed a random distribution of non-native alleles (i.e., multiple generations of admixture) were compared with CI95SA from hybrid swarms (i.e., late sampling, right panel).

CHAPTER 4

EARLY DETECTION OF NON-NATIVE ALLELES IN FISH POPULATIONS: WHEN SAMPLE SIZE ACTUALLY MATTERS

Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: Patrick Della Croce

Contributions: Designed the study, performed simulations and analyzed data, and wrote the manuscript.

Co-author: Dr. Geoffrey C. Poole

Contributions: Assisted with study design, analyses, results discussion, edited and commented the manuscript.

Co-author: Dr. Robert A. Payn

Contributions: Assisted with study design and commented the manuscript.

Co-author: Dr. Robert E. Gresswell

Contributions: Commented the manuscript.

Manuscript Information Page

Patrick Della Croce[1,2,#], Geoffrey C. Poole[1,3], Robert A. Payn[1], Robert E. Gresswell[4]

Journal Name: Fisheries
Status of Manuscript:
_X__Prepared for submission to a peer-reviewed journal
____Officially submitted to a peer-reviewed journal
____Accepted by a peer-reviewed journal
____Published in a peer-reviewed journal

[1]Dept. of Land Resources & Environmental Sciences, Montana State University, P.O. Box 173120, Bozeman, MT 59717-3120
[2]Franklin University Switzerland, via Ponte Tresa 29, CH-6924 Sorengo-Lugano, Switzerland
[3]Institute on Ecosystems, Montana University System, Montana State University, 106 AJM Johnson Hall, Bozeman, MT 59717
[4] U.S. Geological Survey, Northern Rocky Mountain Science Center, Bozeman, Montana

[#]Corresponding author: pdellacroce@fus.edu

Abstract

Reliable detection of non-native alleles is a key element in the conservation of sensitive native fish populations at risk of introgression. Recent advances in genetic techniques have provided many diagnostic markers for a variety of fish species. These species-diagnostic markers can be used to detect the introgression of non-native species with native species. Commonly, the probability of detecting non-native alleles using species diagnostic markers is determined by assuming that a sample of n individuals and analysis of m diagnostic markers represents a random draw of 2nm alleles from a binomial (native and non-native) distribution. Here, we show that such an approach substantially overestimates of the likelihood of detecting non-native alleles. We derive alternative equations based on more realistic assumptions describing the genotypic population structure. Our new equations show that increasing the sample size of individual fish will improve the likelihood of detecting non-native alleles far more than increasing the number of genetic markers used in the analysis -- especially when attempting to detect non-native alleles during early stages of introgression. Therefore, we recommend: (1) a careful reconsideration of the number of individuals sampled when attempting to detect non-native alleles in a population; and (2) a new approach to determining the number of fish to sample and number of diagnostic markers to analyze when attempting to develop monitoring programs that will reliably detect the arrival of non-native alleles in native fish populations.

Introduction

Hybridization and introgression with non-native species have long been recognized as major threats to the conservation of many native fish species worldwide (e.g., Rhymer and Simberloff 1996, Perry et al. 2002). For example, hybridization and introgression have played an important role in 38% of fish extinctions in North America (Miller et al. 1989). This is especially relevant for cutthroat trout (*Oncorhynchus clarkii sp.*) throughout the interior western USA (Gresswell 1988, Behnke 2002) where efforts to protect native subspecies include isolating remaining native populations from potential sources of non-native alleles (Fausch et al. 2009, Muhlfeld et al. 2012).

Where introgression of non-native alleles in the gene pool of a native species is an issue, conservation efforts require reliable means of detecting the arrival of non-native alleles in native populations. Efforts to detect non-native alleles rely on the genetic analysis of *species-diagnostic markers* in a sample of individuals (e.g., Boyer et al. 2008, Bennett and Kershner 2009). As the name suggests, diagnostic markers are relatively short DNA allele sequences on specific genes that are characteristic of a given species and are used to detect non-native alleles in individuals, and thus the presence of hybrids in fish populations (Allendorf and Luikart 2007). Recent progress in genetic sequencing techniques has increased the number of diagnostic markers available for many fish species. For example, in the western USA, thousands of diagnostic markers are now available to detect introgression of alleles from introduced rainbow trout (*O. mykiss*) into populations of native westslope cutthroat trout (*O. clarkii lewisi*) (Hohenlohe et al. 2011).

Well-designed sampling campaigns for early detection of non-native alleles incorporate strategies where the number of individuals sampled from a population (*n*) and the number of diagnostic markers analyzed in each individual (*m*) are sufficient to minimize the chance of failing to detect non-native alleles in the sample. The probability that the analysis of *m* diagnostic markers in *n* individuals fails to detect any non-native alleles (probability of failing to detect non-native alleles, $P_{n,m}$) is often calculated as:

$$P_{n,m} = (1 - A)^{2nm} \quad [1]$$

(e.g., Bennett and Kershner 2009, Amish et al. 2012) where *A* is population admixture (the fraction of the population gene pool that is of non-native origin). For example, Equation 1 predicts that analysis of 8 diagnostic markers (*m* = 8) from 20 individuals (*n* = 20) sampled from a population where 2% of alleles are non-native (*A* = 0.02) would yield a 1% chance of failing to detect non-native alleles in the population ($P_{n,m}$ = 0.01) and thus a 99% chance of detecting non-native alleles(1 - $P_{n,m}$ = 0.99).

The basic assumption associated with this approach is that each allele sampled using *m* diagnostic markers in each of the *n* collected individuals represents a random draw from the entire population gene pool, which in turn is simplified to a binomial distribution of native and non-native alleles. Because each marker has two alleles, the size of the draw is two times the product of *n* and *m* (*2nm*).

Applied to a real population, the assumption of a random draw inherent in Equation 1 is met only if non-native alleles are randomly distributed among individuals in the populations and across diagnostic markers within individuals. In other words, the assumption is met only when the population is a hybrid swarm, in which all individuals

are backcrosses (F-2+ hybrids) that have similar levels of admixture with non-native alleles. The assumption underlying Equation 1 is therefore violated when non-native alleles are clustered in a subset of individuals (e.g., only few individuals carry non-native alleles) and/or non-native alleles are uniformly distributed across diagnostic markers within individuals (e.g., as with non-native and first-generation (F-1) hybrid individuals). Because clustered and/or uniformly distributed non-native alleles are especially likely during the early stages of introgression (e.g., Hitt et al. 2003, Peterson et al. 2004), the assumption of Equation 1 is violated most severely early in the process of introgression, when the opportunity for conservation intervention is greatest.

Della Croce et al. (2015, in review) show that Equation 1 can underestimate $P_{n,m}$ by orders of magnitude when applied to non-hybrid swarms, yielding unwarranted confidence in the probability to detect introgression in such populations. Addressing this flaw in Equation 1 requires that the genotypic structure of a sampled population (i.e., the distribution of individuals among genetically unaltered natives, genetically unaltered non-natives, F-1 hybrids, and F-2+ hybrids) be considered when estimating $P_{n,m}$. Here, we derive an alternative equation to calculate $P_{n,m}$ that incorporates the genotypic structure of the sampled population. We further present a sensitivity analysis of the new equation to illustrate how estimates of $P_{n,m}$ respond to variation in genotypic structure. Finally, we describe how the new equations can be applied to develop a sampling strategy (i.e., select a number of fish to sample and a number of diagnostic markers to analyze) that yields a desired level of $P_{n,m}$ for samples drawn from non-hybrid swarm populations.

Methods

The probability of failing to detect non-native alleles using *m* diagnostic markers in a sample of *n* individuals ($P_{n,m}$) can be thought of as the probability of detecting *only* native alleles in the sample. Because non-native alleles are *uniformly* (not randomly) distributed across diagnostic markers within non-native individuals and F-1 hybrids, analysis of any number of diagnostic markers drawn from non-native or F-1 hybrids will never yield only native alleles. Therefore, for any individual sampled from a population, only native alleles will be detected if:

(a) the individual is a pure native, OR;

(b) the individual is a backcross hybrid (F-2+) and the analyses of *m* diagnostic markers from the F-2+ individual yields, by chance, only native alleles.

Assuming a random distribution of non-native alleles across diagnostic markers within F-2+ hybrids, the probability of finding only native alleles when sampling *m* diagnostic markers from an F-2+ hybrid is:

$$P_{F2_m} = (1 - a)^{2m} \quad [2]$$

where *a* is the proportion of the *individual* admixture with non-native alleles (as opposed to *A*, which is the admixture of the population).

Assuming individuals are drawn at random from a population, the probability of sampling a pure native individual is equal to the fraction of the population that is native (F_N), and the probability of sampling an F-2+ hybrid is equal to the fraction of the population that is comprised of F-2+ hybrids (F_{F2}). Therefore, from (a) and (b) above, the

probability of sampling only native alleles (failing to detect non-native alleles) from a sample of a single individual using *m* diagnostic markers is:

$$P_m = (F_N + [F_{F2}(1-a)^{2m}]) \quad [3]$$

If we assume that the non-native alleles in F-2+ hybrids are randomly distributed among all F-2+ hybrids, then all F-2+ hybrids have a similar value of *a* (A_{F2}), and consequently the probability of finding only native alleles in a sample of *n* individuals is:

$$P_{n,m} = (P_m)^n = (F_N + [F_{F2}(1-A_{F2})^{2m}])^n \quad [4]$$

The assumption that all F-2+ hybrids have similar a proportion of admixture with non-native alleles may be violated in real populations, especially when pure non-natives and F-1 hybrids are present. Such a violation would likely cause Equation 4 to underestimate the probability of failing to detect (i.e., overestimate the probability of detecting) non-native alleles. However, incorporating an uneven distribution of non-native alleles across F-2+ hybrids would require estimates of *a* for every F-2+ hybrid in the population. The uncertainty associated with such an effort would likely negate any increased accuracy derived by avoiding the assumption that $a=A_{F2}$ for all F-2+ hybrids.

When considering Equation 4, the absence of terms describing the fraction of non-natives and F-1 hybrids (F_{NN} and F_{F1}, respectively) as well as the absence of population proportion of admixture with non-native alleles (*A*) is paradoxical. These values must indeed influence $P_{n,m}$. The paradox is resolved because *A*, F_{NN}, and F_{F1} are implicit in the value of A_{F2}. Specifically, *A* is the sum of the population fractions in native, non-native, F-1, and F-2+ individuals, weighted by the corresponding proportion

of admixture. Given that admixture for natives (A_N = 0.0), non-natives (A_{NN} = 1.0), and F-1 hybrids (A_{F-1} = 0.5) are constants, we find:

$$A = F_{NN} + 0.5F_{F1} + A_{F2}F_{F2} \quad [5]$$

Rearranged, Equation 5 can be used to calculate A_{F2} as:

$$A_{F2} = \frac{A - F_{NN} - 0.5F_{F1}}{F_{F2}} \quad [6]$$

Substituting Equation 6 into Equation 4 shows that estimates of $P_{n,m}$ are indeed a function of the genotypic structure of the population (i.e., F_N, F_{NN}, F_{F1}, and F_{F2}) and of the population proportion of admixture with non-native alleles (A), specifically:

$$P_{n,m} = \left(F_N + \left[F_{F2}\left(1 - \frac{A - F_{NN} - 0.5F_{F1}}{F_{F2}}\right)^{2m}\right]\right)^n \quad [7]$$

For a hybrid swarm (where non-native alleles are randomly distributed within and among individuals), F_N, F_{NN}, and F_{F1} are all equal to 0, and F_{F2} = 1. Thus, for hybrid swarms, Equation 7 simplifies to the conventional Equation 1. In contrast, if F-2+ hybrids are absent (i.e. all non-native alleles are in non-native or F-1hybrid individuals), Equation 7 simplifies to:

$$P_{n,m} = (F_N)^n \quad [8]$$

These simplifications demonstrate that equations 4 and 7 are consistent with conventional analyses, and yet provide a more general estimate of $P_{n,m}$ that can be applied to populations other than hybrid swarms.

Using the program R (R Core Team 2015), we created 5,000 unique populations with a 2% proportion of admixture (A = 0.02) but with different genotypic structures. Our code randomly generated the populations that satisfied Equation 5 and conformed to F_N + F_{NN} + F_{F1} + F_{F2} = 1.0. For each population, we than calculated $P_{n,m}$ using both Equation 1

and Equation 7 for all combinations of n = 10, 20, and 50 and m = 20, 40, and 100. Values of n and m were chosen to be similar to commonly used sampling strategies (e.g., Boyer et al. 2008, Amish et al. 2012). We noted that non-native alleles must be more clustered among individuals as F_N approaches 1.0. Thus, we investigated $P_{n,m}$ response to alleles clustering by plotting values of $P_{n,m}$ derived from Equation 1 and Equation 7 for each virtual population against the F_N of the population. We generated this plot for each examined combination of n and m. We further noted that non-native alleles must be more uniformly distributed within individuals as the fraction of non-native alleles found in pure non-natives or F-1 hybrids (i.e., $(F_{NN} + 0.5 \cdot F_{F1})/A$) approaches 1.0. Therefore, to illustrate the response of $P_{n,m}$ to both clustering of non-native alleles across individuals and uniform distribution of non-native alleles within individuals, we created a grey-scale/contour plot of $P_{n,m}$-values for each virtual population, derived from Equation 7, in the 2D space defined by F_N vs. $(F_{NN} + 0.5 \cdot F_{F1})/A$). We also created tables of $P_{n,m}$ values using key combinations of the variables in Equation 4 (n, m, F_N, and A_{F2}).

Using Equations 4 and 8, we developed a new method for designing sampling strategies (choosing combinations of n and m) that will provide a desired level of reliability for detecting non-native alleles when sampling populations where non-native alleles are expected to be clustered among individuals and uniformly distributed across diagnostic markers (e.g., early in the process of introgression between native and non-native species).

Results

Probabilities of failing to detect non-native alleles ($P_{n,m}$) derived from Equation 7 increase with increasing fraction of pure native individuals in the populations (F_N) (Figure 1). This stands in sharp contrast to values estimated using Equation 1, which are invariant across values of F_N and substantially underestimate the probability of failing to detect non-native alleles (i.e., $P_{n,m}$ from Equation 1 << $P_{n,m}$ from Equation 7). Equation 1 is especially unreliable for populations largely comprised of pure native individuals. The 90th and the 95th percentiles of $P_{n,m}$-values derived from Equation 7 can be between 2 (when m=20 and n=10) and roughly 80 (when m=100 and n=50), orders-of-magnitude higher than $P_{n,m}$-values calculated by Equation 1.

Sample size (n) and number of markers used (m) have different effects on $P_{n,m}$-values when the genetic structure of the population is considered (Figure 1), a behavior that cannot be explained by Equation 1. In fact, the 90th and the 95th percentiles of $P_{n,m}$ from Equation 7 show a stronger decrease if n is higher (top to bottom in any of the columns in Figure 1) than if m is higher (left to right in any row in Figure 1). Specifically, increasing m has little to no effect on $P_{n,m}$, when the fraction of pure native individuals in the population (F_N) approaches 1 (Figure 1).

As expected, F_N (as a measure of non-native alleles clustering among individuals) explains only part of the variability of $P_{n,m}$-values derived from Equation 7. Simultaneously considering the uniformity of non-native alleles distribution across diagnostic markers within individuals explains most of the remaining variation in $P_{n,m}$-values from Equation 7 (Figure 2). Most importantly, however, $P_{n,m}$ estimated from

Equation 7 is especially high not only when F_N is high (i.e., non-native alleles are clustered among individuals), but also when most of the non-native alleles in the populations are found in pure non-native individuals and/or in F-1 hybrids (non-native alleles are uniformly distributed across diagnostic markers). Indeed, in both cases, the assumption of randomly distributed non-native alleles within and among individuals is violated, leading Equation 1 to substantially underestimate $P_{n.m}$.

Finally, specific estimates of $P_{n,m}$ (Table 1) for various combinations of n, m, F_N, and A_{F2} (the variables in Equation 4) demonstrate that an increase in n yields substantial reductions in $P_{n,m}$, regardless of the fraction of native individuals in the population or the admixture levels of F-2+ hybrids. In contrast, an increase in m yields a reduction in $P_{n,m}$ only when *all* of the following conditions are true: 1) n is sufficiently large, given F_N of the population, to yield a high likelihood of capturing an individual with non-native alleles; 2) non-native alleles are carried predominantly by F-2+ hybrids rather than non-natives or F-1 hybrids; and 3) admixture levels for F-2+ hybrids are low (~0.05 or less). Where n is insufficient, where non-native alleles are concentrated in non-natives or F-1 hybrids, or where admixture levels for F-2+ hybrids are fairly high (>0.10), an increase in m has little to no effect on $P_{n,m}$.

Discussion

Equation 1 yields dramatic underestimates of $P_{n,m}$ for all conditions except when the population sampled is a hybrid swarm (i.e., all individuals are F-2+ hybrids with similar admixture levels). The underestimation of $P_{n,m}$ by Equation 1 stems from the

assumption that any draw of *2nm* alleles from a population is a random draw from a binomial distribution. For a given level of admixture (*A*), Equation 1 provides the same value of $P_{n,m}$ whether *n*=10 and *m*=50 or *n*=50 and *m*=10 because 1000 alleles are sampled in either case (*2nm*). Yet, in situations other than hybrid swarms, the structure of Equation 4 shows indisputably that *n* and *m* are not interchangeable because sampling markers of a fish from a non-hybrid swarm population is not a random sample of non-native alleles. In fact, because all diagnostic markers in non-native or F-1 hybrid individuals will have at least one non-native allele (i.e., non-native alleles are uniformly distributed across diagnostic markers), increasing *m* beyond *m* = 1 has *no* effect on the probability to detect non-native alleles in a population lacking F-2+ hybrids. In such a case, non-native alleles will be detected *any time an individual containing non-native alleles is captured.* Therefore, in populations where many of the non-native alleles are carried by pure non-native individuals and F-1 hybrids, detection of non-native alleles depends almost exclusively on the probability of capturing an individual carrying non-native alleles, which is determined by *n*.

This lack of interchangeability between *n* and *m* in non-hybrid swarm populations has important implications for the management of native populations. Indeed, populations with the highest conservation value are those in which large numbers of native individuals are still present and, therefore, non-native alleles are clustered in a small number of individuals. In such a case, assuring an adequate number of individuals (*n*) becomes paramount, while the number of diagnostic markers analyzed (*m*) becomes secondary. Further, even when non-native alleles are carried exclusively by F-2+ hybrids,

our results (Table 1, Figure 2) show that when native individuals dominate the population, the number of diagnostic markers analyzed has only a marginal effect on the ability to detect non-native alleles unless the number of individuals sampled is sufficient to provide a high probability of capturing individuals carrying the non-native alleles.

Indirectly, our results also suggest that the age class distribution of the individuals sampled may be important. Our results show that $P_{n,m}$ decreases as non-native alleles become more randomly distributed among individuals and across diagnostic markers. Assuming active crossing between native, non-native, and hybrid individuals in a population early in the process of introgression, non-native alleles will be more randomly distributed among individuals and across diagnostic markers in younger age classes than in older age classes. This implies that including young individuals in the sample will increase the chances of collecting individuals carrying non-native alleles and therefore increase the chances of detecting introgression. The common practice of preferably sampling large individuals may limit the probability of early detection of non-native alleles, which stresses the importance of in proportion to the age class distribution of the population.

In summary, our results suggest that sampling strategies derived from Equation 1 likely fail to ensure high probabilities of detecting non-native alleles in native populations, especially during the initial stages of introgression when detection is most critical. A sample of n individuals and m diagnostic markers does not represent a random draw of $2nm$ alleles from the population gene pool, and therefore the relative influence of

increasing n vs. m on $P_{n,m}$ is determined by the genotypic structure of the target population.

A Proposal for Sampling Design Strategy

Management actions to prevent introgression between native and non-native fish species are heavily dependent on detecting non-native alleles (Allendorf et al. 2001, Shepard et al. 2005) and include, for instance, the use of migration barriers to isolate native populations (Novinger and Rahel 2003). When making such management decisions, a reliable means of detecting non-native alleles is paramount. As our study suggests, sampling strategies designed using Equation 1 impart overconfidence in the ability to detect non-native alleles in populations dominated by pure native individuals and could thus undermine the efficacy of such expensive and long-term management decisions. For example, as a possible measure to prevent introgression between native westslope cutthroat trout and non-native rainbow trout, Muhlfeld et al. (2012) investigated the possibility of isolating the Akokala Creek watershed in Glacier National Park (Montana, U.S.). Previous analyses on 7 diagnostic markers in 32 fish sampled from Akokala Creek detected no rainbow trout alleles (Boyer et al. 2008). According to Equation 1 the probability of missing non-native alleles through such a sampling is <1%, even if we assume that only 1% of the gene pool of the population in Akokala Creek is comprised of non-native alleles (A=0.01). Yet, Equation 8 shows that there is a 72% probability of failing to detect non-native alleles using this sampling strategy, if non-native alleles are clustered in only few individuals (e.g., F_N=0.99). We therefore urge a careful reconsideration of how sample strategies are designed.

Equations 4 and 8 can be used to design more reliable sampling strategies. At a first glance application of Equations 4 and 8 may seem problematic; both equations require variables that describe the genotypic structure (F_N, F_{F2+} and A_{F2}) of the sampled population. If the genotypic structure of a population is known, the presence or absence of non-native alleles in that population is also known. Thus, the information necessary to accurately estimate $P_{n,m}$ for any given population *de facto* nullifies the need for the detection of non-native alleles, but the absence of such information frustrates any calculation of $P_{n,m}$. However, when combined with some conservative assumptions about the genotypic structure of the population to sample, Equations 4 and 8 become useful tools in designing sampling strategies that provide more reliable values for $P_{n,m}$.

The initial step to determine the sampling effort (combination of *n* and *m*) required to detect non-native alleles in a vulnerable native population is to determine the desired "invasion intensity" at which detection of non-native alleles is desired. For instance, if detection is desired when 1 individual out of 100 carries non-native alleles, the invasion intensity would be 0.01. From invasion intensity, F_N would be:

$$F_N = (1 - Invasion\ Intensity) \quad [9]$$

Second, the threshold for desired "reliability" of the sampling strategy is selected. For instance, if the chance of failure in detecting non-native alleles should not exceed 1%, the desired threshold for reliability is 0.99. From reliability, $P_{n,m}$ can be calculated as:

$$P_{n,m} = (1 - Reliability) \quad [10]$$

If the population is unlikely to include F-2+ hybrids, *m* will have little effect on $P_{n,m}$. Equation 8 can be used with F_N (derived from Equation 9) and $P_{n,m}$ (derived from Equation 10) to calculate *n*:

$$n = \left\lceil \frac{\log_{10}(P_{n,m})}{\log_{10}(F_N)} \right\rceil \quad [11]$$

(the single-headed brackets denote that *n* should be "rounded up" to the next integer). Theoretically, this estimate of *n* could be paired with an *m* as low as 1 because non-native alleles are uniformly distributed across markers in non-native and F-1 hybrid individuals. However, we recommend pairing any estimate of *n* derived from Equation 11 with at least a modest value of *m* (e.g., between 20 and 40), so that F-2+ hybrids have a reasonable chance of being detected should they be in the population unexpectedly.

In contrast, if the population is likely to include F-2+ hybrids, or if there is uncertainty about occurrence of F-2+ hybrids, Equation 4 should be used to estimate *n* rather than Equation 8. This requires appropriate values for F_{F2} and A_{F2} in addition to $P_{n,m}$ and F_N. A conservative estimate for A_{F2} is to assume all individuals carrying non-native alleles are F-2+. Therefore:

$$F_{F2} = Invasion\ Intensity \quad [12]$$

where *Invasion Intensity* is the same value used in Equation 9. The threshold for admixture level at which detection of F-2+ individuals is desired is then selected. For instance, to detect non-native alleles in a population where the F-2+ individuals have, on average, ≥5% non-native alleles, set A_{F2} to 0.05.

Finally, several reasonable values of *m* that are within the capability of genetic analysis labs (e.g. *m* = {10, 20, 40, 100}) should be selected to assess the sensitivity of

the reliability of the study to the number of markers used. F_N (derived from Equation 9), $P_{n,m}$ (derived from Equation 10), F_{F2} (derived from Equation 12), and A_{F2} (as described in the preceding paragraph) are used with a rearranged version of Equation 4 to calculate the sample size *n* required for each of the selected values of *m*:

$$n = \left\lceil \frac{\log_{10}(P_{n,m})}{\log_{10}(F_N + [F_{F2} \cdot (1 - A_{F2})^{2m}])} \right\rceil \quad [13]$$

Again, *n* should be "rounded up" to the next integer. Each value of *m* used in Equation 13 will yield a corresponding value of *n*. Because each *n,m* pair has the same likelihood of failing to detect non-native alleles (same $P_{n,m}$), we suggest selecting the *n,m* pair that is most appropriate or most cost effective, given the relative costs of collecting individuals vs. genetic analysis.

As an alternative, lookup tables for pairings of *n* and *m* derived from various combinations of $P_{n,m}$, F_N, F_{F2} and A_{F2} appear in the online supplement for this paper. Values in the table are derived from Equations 11 or 13, as described above.

Conclusion

Based on our findings, the use of Equation 1 to estimate necessary sample sizes of *n* and *m* for early detection of non-native alleles in native populations is questionable. Equation 1 provides inaccurate estimates of the probability of failing to detect non-native alleles associated with a sample of *n* individuals and *m* diagnostic markers, especially for populations in the early stages of introgression. The probability of failing to detect non-native alleles in predominantly native populations is orders of magnitude higher than that suggested by Equation 1, leading to marked overconfidence in any associated monitoring

results. We recommend use of Equation 11 or Equation 13, as described above, as a more appropriate strategy for estimating the sample size and number of genetic markers required to monitor for the arrival of non-native alleles in native populations.

Acknowledgments

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Tables and Figures

Table 4.1: Probability of failing to detect non-native alleles ($P_{n,m}$) derived from Equation 4 for various numbers of individuals collected (n), number of diagnostic markers examined (m), and fraction of native individuals in the sampled population (FN). In the last column, all invading individuals are assumed to be either pure non-native or F1 hybrids. In the remaining columns, all invading individuals are assumed to be F-2+ hybrids, with individual levels of admixture (a_{F2}) equal to the F-2+ population admixture (A_{F2}) listed in the column heading.

Fraction Native = 0.99

	F-2+ Invaders: F-2+ Admixture = 0.01				F-2+ Invaders: F-2+ Admixture = 0.05				F-2+ Invaders: F-2+ Admixture = 0.1				Non-native and/or F-1 invaders			
	m				m				m				m			
n	7	10	40	100	7	10	40	100	7	10	40	100	7	10	40	100
10	0.98695	0.98194	0.94611	0.91670	0.94993	0.93767	0.90589	0.90439	0.92550	0.91555	0.90440	0.90438	0.90438	0.90438	0.90438	0.90438
50	0.93644	0.91290	0.75805	0.64733	0.77351	0.72485	0.61007	0.60502	0.67902	0.64329	0.60507	0.60501	0.60501	0.60501	0.60501	0.60501
100	0.87692	0.83339	0.57464	0.41904	0.59831	0.52541	0.37219	0.36605	0.46106	0.41383	0.36611	0.36603	0.36603	0.36603	0.36603	0.36603
500	0.51856	0.40200	0.06266	0.01292	0.07667	0.04004	0.00714	0.00657	0.02084	0.01214	0.00658	0.00657	0.00657	0.00657	0.00657	0.00657

Fraction Native = 0.95

	m				m				m				m			
n	7	10	40	100	7	10	40	100	7	10	40	100	7	10	40	100
10	0.93628	0.91259	0.75569	0.64232	0.77144	0.72180	0.60396	0.59875	0.67486	0.63817	0.59881	0.59874	0.59874	0.59874	0.59874	0.59874
50	0.71949	0.63298	0.24644	0.10934	0.27321	0.19592	0.08036	0.07695	0.13998	0.10585	0.07699	0.07694	0.07694	0.07694	0.07694	0.07694
100	0.51766	0.40066	0.06073	0.01195	0.07465	0.03838	0.00646	0.00592	0.01960	0.01120	0.00593	0.00592	0.00592	0.00592	0.00592	0.00592
500	0.03717	0.01033	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Fraction Native = 0.9

	m				m				m				m			
n	7	10	40	100	7	10	40	100	7	10	40	100	7	10	40	100
10	0.87623	0.83213	0.56647	0.40420	0.59101	0.51529	0.35513	0.34869	0.44816	0.39875	0.34876	0.34868	0.34868	0.34868	0.34868	0.34868
50	0.51653	0.39897	0.05833	0.01079	0.07211	0.03633	0.00565	0.00515	0.01808	0.01008	0.00516	0.00515	0.00515	0.00515	0.00515	0.00515
100	0.26681	0.15918	0.00340	0.00012	0.00520	0.00132	0.00003	0.00003	0.00033	0.00010	0.00003	0.00003	0.00003	0.00003	0.00003	0.00003
500	0.00135	0.00010	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

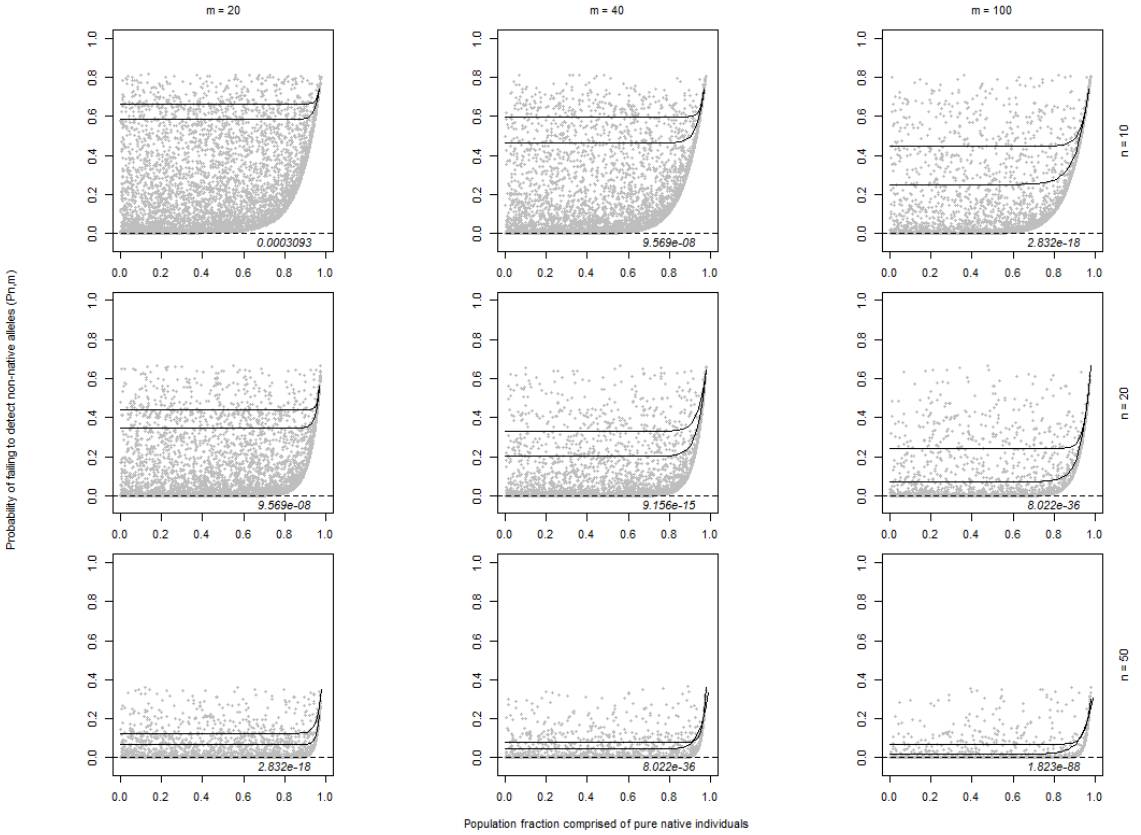

Figure 4.1: Probability of failing to detect non-native alleles in a population (Pn,m, grey diamonds) derived from Equation 7 compared to the fraction of the population that is comprised of pure native individuals (FN) for nine different combinations of number of individuals sampled (n, rows) and number of diagnostic markers analyzed (m, columns). Solid lines show the 90th and the 95th percentiles of Pn,m, while the dashed line shows Pn,m derived from Equation 1 (exact value in italics).

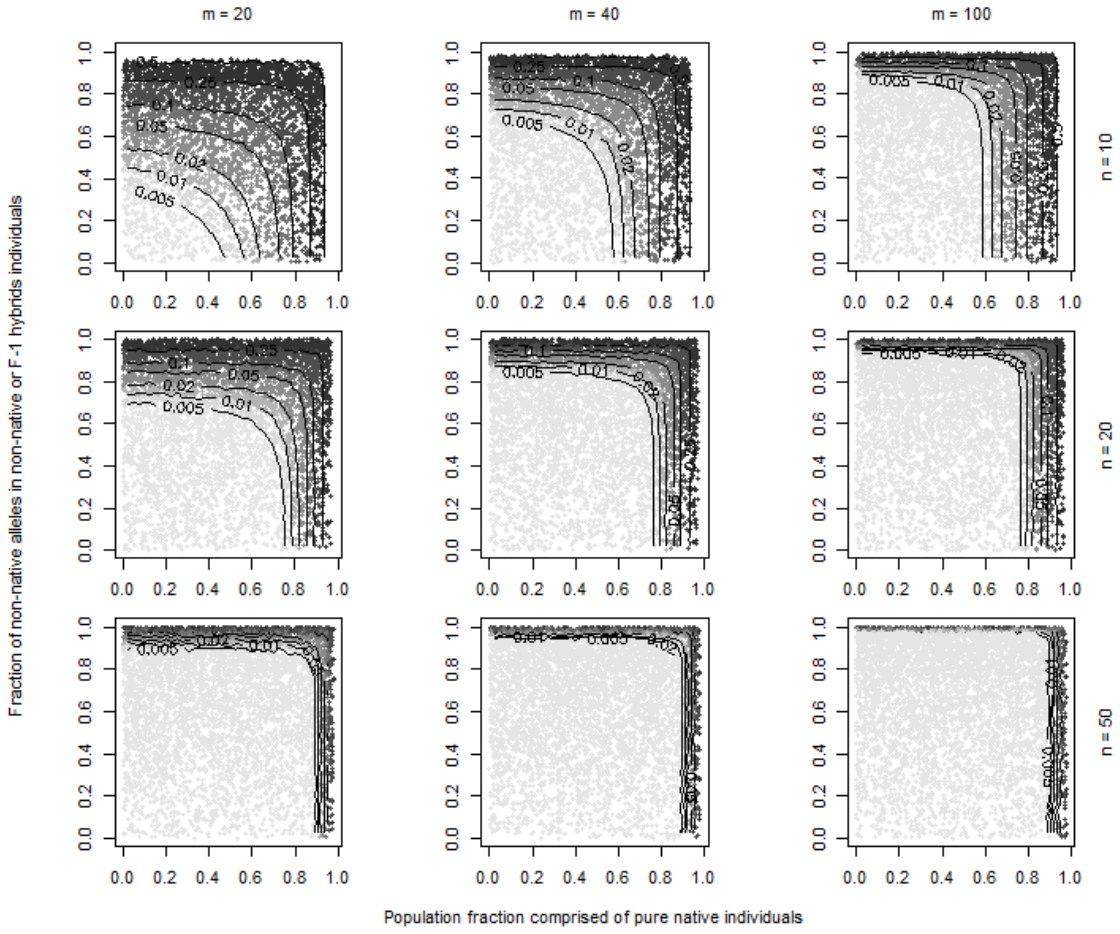

Figure 4.2: Contour plot showing the probability of failing to detect non-native alleles in a population (Pn,m, grey-black dots) derived from Equation 7 as a function of both the fraction of the population that is comprised of pure native individuals (FN, x-axis) and the fraction of non-native alleles found in non-native and F-1 hybrid individuals (y-axis) for nine different combinations of number of individuals sampled (n, rows) and number of diagnostic markers analyzed (m, columns). Contour lines are provided to show changes in Pn,m-values.

CHAPTER 5

CONCLUSION

Project Summary

As stated by Stone (2000), "it is an understatement to say that hybridization is a complex business". Indeed, hybridization with non-native species – especially with introgression – is not only a well-recognized threat to the conservation of native fish species (Miller *et al.* 1989b; Rhymer & Simberloff 1996b), but also poses a uniquely complicated set of tasks to managers and researchers alike (Allendorf *et al.* 2001).

First, inference about the occurrence and the magnitude of introgression in a population must often be drawn from genetic analyses on a sample of individuals from the population (see for example Boyer *et al.* 2008) – using so-called *species-diagnostic markers* (Amish *et al.* 2012). Genetic analyses are indeed necessary as hybrids are often morphologically undistinguishable from genetically pure parental individuals (Allendorf *et al.* 2004; Seiler *et al.* 2009). Moreover, it is usually unfeasible to collect all the individuals of a population. As a result, efforts to characterize the genotypic structure of a population depend on the strategy used to sample the population (i.e., the combination of number of fish sampled and the number of diagnostic markers analyzed in each individual). In particular, the sampling strategy affects both (a) the ability to detect non-native alleles in a population and (b) the uncertainty around field-estimates of the abundance of non-native alleles in the population's gene pool (typically expressed as the

proportion of admixture with non-native alleles of a population, see *Chapter 3* for examples).

Second, the spatiotemporal movement of non-native alleles (across generations within a population and among connected populations across a network) is interactively affected by: (a) landscape features such as connectivity patterns; (b) fitness and behavior characteristics of native, non-native, and hybrid individuals, and; (c) stochastic events associated with both the passage of genes across generations and the movement of individuals – and their genes – across the river network. Therefore, an effective interpretation of observed patterns of distribution and abundance of non-native alleles within single populations and across river networks must take all these factors into consideration.

To date, fisheries studies related to introgression between native and non-native species often simplify or ignore the effects of sampling strategies on the ability to detect and the precision to quantify non-native genes in a population (*Chapters 3* and *4*). Also, landscape features, genetically driven fitness and behavioral characteristics, and stochasticity are seldom taken into consideration during the interpretation of observed patterns of abundance and distribution of non-native alleles across river networks.

Following the ideas outlined by Epperson *et al.* (2010), this research project presents an individual-based simulation models (IBM) that mimics the spread of non-native alleles across a river network inhabited by trout populations (the complete description of the model is provided in *Chapter 2*). The main purpose of the model is to guide the design of field data collection and to help the interpretation of observed patterns

of introgression. Indeed, the possibility of repeatedly sampling simulated populations – and simulated diagnostic markers in each individual – allows for the quantification of the effects of sampling strategy on the metric of interest. This information can then be used to design the most effective sampling strategy. Furthermore, the model allows users to investigate how landscape features, individual fitness and behavior characteristics, and stochastic events influence – individually or interactively – spatiotemporal patterns of non-native alleles across a network. The results from such investigations can then provide fundamental information for a correct interpretation of observed patterns of introgression.

Model simulations conducted in *Chapter 2* show that by affecting the connectivity patterns among populations, network topology affects the spread of non-native alleles across a river network. Specifically, comparisons of the rate of spread of non-native alleles across three networks with different topologies (linear, trellis, and dendritic, Figure 2.1) show that network bifurcations act like semi-permeable barriers to the spread of non-native alleles thereby decreasing the rate at which introgression spreads from an initial source (Figure 2.5). Moreover, by forcing migrating individuals to randomly choose between two destinations, irregularly distributed bifurcations decrease the predictability of the spread of introgression (Figure 2.7). Results from Chapter 2 thus suggest that the topology of the network ought to be considered when spatial patterns of introgression are to be interpreted.

The work presented in *Chapter 3* and *Chapter 4* investigates the effects of sampling strategies on the probability to detect and the precision to quantify non-native alleles in a population. Results from these chapters call for a careful reconsideration of

the way native fish populations are sampled for non-native alleles. Specifically, *Chapters 3* and *4* show that simplistic binomial equations commonly used in fisheries studies (e.g., Newcombe 1998, Kozfkay *et al.* 2007, Amish *et al.* 2012) tend to overestimate the ability to detect non-native alleles in a population and to underestimate the uncertainty around estimates of population proportion admixture with non-native alleles associated with a given sampling strategy. This is especially true for non-hybrid swarm populations (populations in which not all individuals are hybrids).

In addition, simulations from *Chapter 3* and *Chapter 4* also show that in the early stages of invasion by non-native alleles (when the population is still not a hybrid swarm), the likelihood to detect non-native alleles and the precision of estimates of proportion of admixture are mainly limited by the number of individuals collected rather than by the number of diagnostic markers analyzed in each individual. This result is in contradiction with the current conviction that both number of individuals collected and number of diagnostic markers analyzed influence ability to detect and precision to quantify introgression.

In many cases, introgression with non-native alleles is so extensive and widespread that the long-term conservation of the native species is strongly dependent on the active and timely management of both the genetically unaltered and introgressed populations (Allendorf *et al.* 2001). Therefore, it is of paramount importance for researchers and managers to have realistic – and yet reliable – confidence in their sampling strategies. As mentioned above, the results from *Chapter 3* and *Chapter 4* suggest that the simplistic equations commonly used in fisheries studies should be

abandoned if effective conservation measures are to be guaranteed. As an alternative, *Chapter 4* proposes a new method to reliably estimate the power to detect non-native alleles in a population associated with a given sampling strategy. An alternative method to calculate more realistic uncertainties around estimates of proportion of admixture with non-native alleles derived from a sample of individuals and a sample of diagnostic markers, was beyond the scope of this research project and remains to be developed.

In conclusion, the model developed for this project is an attempt to capture and describe the complexity intrinsic to the spatiotemporal spread of introgression with non-native genes across river networks inhabited by sensitive native species. The model accounts for what are arguably the most critical drivers to the spread of introgression across river systems and helps understand how sampling strategies affect the ability to describe such spread. As we develop means to provide mechanistic explanations of observed patterns of introgression, the efficiency in managing native species undergoing introgression with non-native species will improve as will the likelihood of their long-term survival.

REFERENCES

Adams SB, Frissell CA, Rieman BE (2000) Movements of nonnative brook trout in relation to stream channel slope. *Transactions of the American Fisheries Society*, **129**, 623–638.

Adams SB, Frissell CA, Rieman BE (2001) Geography of invasion in mountain streams: Consequences of headwater lake fish introductions. *Ecosystems*, **4**, 296–307.

Allendorf FW, Leary RF, Hitt NP *et al.* (2004) Intercrosses and the US Endangered Species Act: Should hybridized populations be included as Westslope cutthroat trout? *Conservation Biology*, **18**, 1203–1213.

Allendorf FW, Leary RF, Spruell P, Wenburg JK (2001) The problems with hybrids: setting conservation guidelines. *Trends in Ecology & Evolution*, **16**, 613–622.

Allendorf FW, Luikart G (2007) *Conservation and the Genetics of Populations*. Blackwell Publishing.

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

Antolin M, Savage L, Eisen R (2006) Landscape features influence genetic structure of black-tailed prairie dogs (<i>Cynomys ludovicianus</i>). *Landscape Ecology*, **21**, 867–875.

Balkenhol N, Gugerli F, Cushman S *et al.* (2009) Identifying future research needs in landscape genetics: where to from here? *Landscape Ecology*, **24**, 455–463.

Baxter C V, Fausch KD, Murakami M, Chapman PL (2007) Invading rainbow trout usurp a terrestrial prey subsidy from native charr and reduce their growth and abundance. *Oecologia*, **153**, 461–470.

Behnke RJ (2002) Trout and salmon of North America. *Simon and Schuster Inc., New York.*

Benjamin JR, Dunham JB, Dare MR (2007) Invasion by nonnative brook trout in panther creek, Idaho: Roles of local habitat quality, biotic resistance, and connectivity to source habitats. *Transactions of the American Fisheries Society*, **136**, 875–888.

Bennett SN, Kershner JL (2009) Levels of Introgression in Westslope Cutthroat Trout Populations Nine Years after Changes to Rainbow Trout Stocking Programs in Southeastern British Columbia. *North American Journal of Fisheries Management*, **29**, 1271–1282.

Bennett SN, Olson JR, Kershner JL, Corbett P (2010) Propagule pressure and stream characteristics influence introgression: cutthroat and rainbow trout in British Columbia. *Ecological Applications*, **20**, 263–277.

Bímová BV, Macholán M, Baird SJE *et al.* (2011) Reinforcement selection acting on the European house mouse hybrid zone. *Molecular Ecology*, **20**, 2403–2424.

Blumler MA (2003) Introgression as a spatial phenomenon. *Physical Geography*, **24**, 414–432.

Borchers HW (2015) pracma: Practical Numerical Math Functions. R package version 1.8.6.

Boyer MC, Muhlfeld CC, Allendorf FW (2008) Rainbow trout (Oncorhynchus mykiss) invasion and the spread of hybridization with native westslope cutthroat trout (Oncorhynchus clarkii lewisi). *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 658–669.

Cassens I, Tiedemann R, Suchentrunk F, Hartl GB (2000) Mitochondrial DNA variation in the European otter (Lutra lutra) and the use of spatial autocorrelation analysis in conservation. *Journal of Heredity*, **91**, 31–35.

Chikhi L, Bruford MW, Beaumont MA (2001) Estimation of Admixture Proportions: A Likelihood-Based Approach Using Markov Chain Monte Carlo. *Genetics*, **158**, 1347–1362.

COSEWIC (2006) COSEWIC (Committee on the Status of Endangered Wildlife in Canada) assessment and update status report on the westslope cutthroat trout Oncorhynchus clarkii lewisii (British Columbia population and Alberta population) in Canada. *COSEWIC, Ottawa. Available: www.cosewic.gc.ca. (October 2009).*

Cote D, Kehler DG, Bourne C, Wiersma YF (2009) A new measure of longitudinal connectivity for stream networks. *Landscape Ecology*, **24**, 101–113.

Della Croce P, Poole GC, Luikart G (2015) Simplifying Assumptions Yield Overconfidence in Detecting and Quantifying Introgression in Hybridized Populations. *In review*.

Della Croce P, Poole GC, Payn RA, Izurieta C (2014) Simulating the effects of stream network topology on the spread of introgressive hybridization across fish populations. *Ecological Modelling*, **279**, 68–77.

Dionne Mél, Caron F, Dodson JJ, Bernatchez L (2008) Landscape genetics and hierarchical genetic structure in Atlantic salmon: the interaction of gene flow and local adaptation. *Molecular Ecology*, **17**, 2382–2396.

Dowling TE, Childs MR (1992) Impact of Hybridization on a Threatened Trout of the Southwestern United-States. *Conservation Biology*, **6**, 355–364.

Epperson BK, McRae BH, Scribner KIM *et al.* (2010) Utility of computer simulations in landscape genetics. *Molecular Ecology*, **19**, 3549–3564.

Epps CW, Wehausen JD, Bleich VC, Torres SG, Brashares JS (2007) Optimizing dispersal and corridor models using landscape genetics. *Journal of Applied Ecology*, **44**, 714–724.

Fausch KD (2007) Introduction, establishment and effects of non-native salmonids: considering the risk of rainbow trout invasion in the United Kingdom. *Journal of Fish Biology*, **71**, 1–32.

Fausch KD, Rieman BE, Dunham JB, Young MK, Peterson DP (2009) Invasion versus Isolation: Trade-Offs in Managing Native Salmonids with Barriers to Upstream Movement. *Conservation Biology*, ******.

Fausch KD, Taniguchi Y, Nakano S, Grossman GD, Townsend CR (2001) Flood disturbance regimes influence rainbow trout invasion success among five holarctic regions. *Ecological Applications*, **11**, 1438–1455.

Frankham R, Ballou JJD, Briscoe DDA (2010) *Introduction to Conservation Genetics*. Cambridge University Press.

Fritts TH, Rodda GH (1998) The role of introduced species in the degradation of island ecosystems: A case history of Guam. *Annual Review of Ecology and Systematics*, **29**, 113–140.

Gozlan RE, Britton JR, Cowx I, Copp GH (2010) Current knowledge on non-native freshwater fish introductions. *Journal of Fish Biology*, **76**, 751–786.

Gresswell RE (Ed.) (1988) Status and management of interior stocks of cutthroat trout. In: *American Fisheries Society, Symposium 4*. Bethesda, Maryland.

Gresswell RE, Hendricks SR (2007) Population-scale movement of coastal cutthroat trout in a naturally isolated stream network. *Transactions of the American Fisheries Society*, **136**, 238–253.

Grimm V, Berger U, Bastiansen F *et al.* (2006a) A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, **198**, 115–126.

Grimm V, Berger U, Bastiansen F *et al.* (2006b) A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, **198**, 115–126.

Gunnell K, Tada MK, Hawthorne FA, Keeley ER, Ptacek MB (2008) Geographic patterns of introgressive hybridization between native Yellowstone cutthroat trout (Oncorhynchus clarkii bouvieri) and introduced rainbow trout (O-mykiss) in the South Fork of the Snake River watershed, Idaho. *Conservation Genetics*, **9**, 49–64.

Guy TJ, Gresswell RE, Banks MA (2008) Landscape-scale evaluation of genetic structure among barrier-isolated populations of coastal cutthroat trout, Oncorhynchus clarkii clarkii. *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 1749–1762.

Halverson A (2010) *An entirely synthetic fish : how rainbow trout beguiled America and overran the world*. Yale University Press, New Haven.

Hendry SC, Stearns AP (2003) Evolution Illuminated: Salmon and Their Relatives. Oxford University Press.

Hilderbrand RH, Kershner JL (2000) Movement patterns of stream-resident cutthroat trout in Beaver Creek, Idaho-Utah. *Transactions of the American Fisheries Society*, **129**, 1160–1170.

Hitt NP, Frissell CA, Muhlfeld CC, Allendorf FW (2003) Spread of hybridization between native westslope cutthroat trout, Oncorhynchus clarki lewisi, and nonnative rainbow trout, Oncorhynchus mykiss. *Canadian Journal of Fisheries and Aquatic Sciences*, **60**, 1440–1451.

Hoban S, Bertorelle G, Gaggiotti OE (2012) Computer simulations: tools for population and evolutionary genetics. *Nature Reviews Genetics*, **13**, 110–122.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Hohenlohe PA, Day MD, Amish SJ *et al.* (2013a) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*.

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

Holderegger R, Wagner H (2006) A brief guide to Landscape Genetics. *Landscape Ecology*, **21**, 793–796.

Holway DA (1999) Competitive mechanisms underlying the displacement of native ants by the invasive Argentine ant. *Ecology*, **80**, 238–251.

Holzhauer S, Ekschmitt K, Sander A-C, Dauber J, Wolters V (2006) Effect of historic landscape change on the genetic structure of the bush-cricket <i>Metrioptera roeseli</i&gt. *Landscape Ecology*, **21**, 891–899.

Jager HI, Chandler JA, Lepla KB, Van Winkle W (2001) A theoretical study of river fragmentation by dams and its effects on white sturgeon populations. *Environmental Biology of Fishes*, **60**, 347–361.

Kalinowski ST, Novak BJ, Drinan DP, Jennings R deM, Vu N V (2011) Diagnostic single nucleotide polymorphisms for identifying westslope cutthroat trout (Oncorhynchus clarki lewisi), Yellowstone cutthroat trout (Oncorhynchus clarkii bouvieri) and rainbow trout (Oncorhynchus mykiss). *Molecular Ecology Resources*, **11**, 389–393.

Keeler-Foster C (2003) Rio Grande cutthroat trout (Oncorhynchus clarki virginalis) population genetics and hybridization with introduced rainbow trout (Oncorhynchus mykiss).

Kitano S, Hasegawa K, Maekawa K (2009) Evidence for interspecific hybridization between native white-spotted charr Salvelinus leucomaenis and non-native brown trout Salmo trutta on Hokkaido Island, Japan. *Journal of Fish Biology*, **74**, 467–473.

Koenker R (2015) quantreg: Quantile Regression. R package version 5.19.

Kozfkay CC, Campbell MR, Yundt SP, Peterson MP, Powell MS (2007) Incidence of Hybridization between Naturally Sympatric Westslope Cutthroat Trout and Rainbow Trout in the Middle Fork Salmon River Drainage, Idaho. *Transactions of the American Fisheries Society*, **136**, 624–638.

Krueger CC, May B (1991) ECOLOGICAL AND GENETIC-EFFECTS OF SALMONID INTRODUCTIONS IN NORTH-AMERICA. *Canadian Journal of Fisheries and Aquatic Sciences*, **48**, 66–77.

Kruse CG, Hubert WA, Rahel FJ (2000) Status of Yellowstone Cutthroat Trout in Wyoming Waters. *North American Journal of Fisheries Management*, **20**, 693–705.

Landguth EL, Muhlfeld CC, Luikart G (2012a) CDFISH: an individual-based, spatially-explicit, landscape genetics simulator for aquatic species in complex riverscapes. *Conservation Genetics Resources*, **4**, 133–136.

Landguth EL, Muhlfeld CC, Luikart G (2012b) CDFISH: an individual-based, spatially-explicit, landscape genetics simulator for aquatic species in complex riverscapes. *Conservation Genetics Resources*, **4**, 133–136.

Langeland A, L'Abee-Lund JH (1996) Habitat use, size and age structure in sympatric brown trout (Salmo trutta) and Arctic charr (Salvelinus alpinus) stocks: Resistance of populations to change following harvest. *Ecology of Freshwater Fish*, **5**, 49–58.

Leary RF, Allendorf FW, Sage GK (1995) Hybridization and introgression between introduced and native fish. *Uses and Effects of Cultured Fishes in Aquatic Ecosystems*, **15**, 91–101.

Mack RN, Simberloff D, Lonsdale WM *et al.* (2000) Biotic invasions: Causes, epidemiology, global consequences, and control. *Ecological Applications*, **10**, 689–710.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003a) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003b) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.

May BE, Albeke SE, Horton T (2007) *Range-Wide Status Assessment for Yellowstone Cutthroat Trout (Oncorhynchus clarkii bouvieri): 2006*.

McCleave JD (1967) Homing and orientation of cutthroat trout (Salmo clarki) in Yellowstone lake, with special reference to olfaction and vision. *Diss. Abstr.*, **28B**, 2190.

Meeuwig MH, Guy CS, Kalinowski ST, Fredenberg WA (2010) Landscape influences on genetic differentiation among bull trout populations in a stream-lake network. *Molecular Ecology*, **19**, 3620–3633.

Meyer KA, Schill DJ, Lamansky JA, Campbell MR, Kozfkay CC (2006) Status of Yellowstone Cutthroat Trout in Idaho. *Transactions of the American Fisheries Society*, **135**, 1329–1347.

Miller RR, Williams JD, Williams JE (1989a) Extinctions of North American fishes during the past century. *Fisheries*, **14**, 22–38.

Miller RR, Williams JD, Williams JE (1989b) Extinctions of North American fishes during the past century. *Fisheries*, **14**, 22–38.

Moritz C, Meynard CN, Devictor V *et al.* (2013) Disentangling the role of connectivity, environmental filtering, and spatial structure on metacommunity dynamics. *Oikos*, **122**, 1401–1410.

Muhlfeld CC, D'Angelo V, Kalinowski ST *et al.* (2012) A Fine-scale Assessment of Using Barriers to Conserve Native Stream Salmonids: A Case Study in Akokala Creek, Glacier National Park, USA. *Open Fish Science Journal*, **5**.

Muhlfeld CC, Kalinowski ST, McMahon TE *et al.* (2009a) Hybridization rapidly reduces fitness of a native trout in the wild. *Biology Letters*, **5**, 328–331.

Muhlfeld CC, McMahon TE, Belcer D, Kershner JL (2009b) Spatial and temporal spawning dynamics of native westslope cutthroat trout, Oncorhynchus clarkii lewisi, introduced rainbow trout, Oncorhynchus mykiss, and their hybrids. *Canadian Journal of Fisheries and Aquatic Sciences*, **66**, 1153–1168.

Muhlfeld CC, McMahon TE, Boyer MC, Gresswell RE (2009c) Local Habitat, Watershed, and Biotic Factors Influencing the Spread of Hybridization between Native Westslope Cutthroat Trout and Introduced Rainbow Trout. *Transactions of the American Fisheries Society*, **138**, 1036–1051.

Newcombe RG (1998) Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine*, **17**, 857–872.

Novinger DC, Rahel FJ (2003) Isolation management with artificial barriers as a conservation strategy for cutthroat trout in headwater streams. *Conservation Biology*, **17**, 772–781.

Olsen JB, Flannery BG, Beacham TD *et al.* (2008) The influence of hydrographic structure and seasonal run timing on genetic diversity and isolation-by-distance in chum salmon (Oncorhynchus keta). *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 2026–2042.

Ott L (1997) *An introduction to statistical methods and data analysis*. Duxbury Pacific Grove, CA.

Peacock MM, Kirchoff V (2004) Assessing the Conservation Value of Hybridized Cutthroat Trout Populations in the Quinn River Drainage, Nevada. *Transactions of the American Fisheries Society*, **133**, 309–325.

Perry WL, Lodge DM, Feder JL (2002) Importance of hybridization between indigenous and nonindigenous freshwater species: an overlooked threat to North American biodiversity. *Systematic Biology*, **51**, 255–275.

Peterson MP, Campbell MR, Cegelski CC, Powell MS (2004) Introgressive hybridization between westslope cutthroat trout and native rainbow trout in big creek, Idaho. *Propagated Fish in Resource Management*, **44**, 505–514.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Core Team (2015) R: A Language and Environment for Statistical Computing.

Rasmussen JB, Robinson MD, Heath DD (2010) Ecological consequences of hybridization between native westslope cutthroat (Oncorhynchus clarkii lewisi) and introduced rainbow (Oncorhynchus mykiss) trout: effects on life history and habitat use. *Canadian Journal of Fisheries and Aquatic Sciences*, **67**, 357–370.

Rhymer JM, Simberloff D (1996a) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, **27**, 83–109.

Rhymer JM, Simberloff D (1996b) Extinction By Hybridization and Introgression. *Annual Review of Ecology and Systematics*, **27**, 83–109.

De Rito JNJ (2004) Assessment of reproductive isolation between Yellowstone cutthroat trout and rainbow trout in the Yellowstone River, Montana. . *Master's thesis, Fish and Wildlife Management, Montana State University, Bozeman.*

Rozhnov V V (1993) Extinction of the European mink: ecological catastrophe or a natural process? *Lutreola (Moscow)*, **1**, 10–16.

Rubidge E, Corbett P, Taylor EB (2001) A molecular analysis of hybridization between native westslope cutthroat trout and introduced rainbow trout in southeastern British Columbia, Canada. *Journal of Fish Biology*, **59**, 42–54.

Rubidge EM, Taylor EB (2005) An analysis of spatial and environmental factors influencing hybridization between native westslope cutthroat trout (Oncorhynchus

clarki lewisi) and introduced rainbow trout (O-mykiss) in the upper Kootenay River drainage, British Columbia. *Conservation Genetics*, **6**, 369–384.

Schwartz MK, Luikart G, Waples RS (2007) Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology & Evolution*, **22**, 25–33.

Segelbacher G, Cushman SA, Epperson BK *et al.* (2010) Applications of landscape genetics in conservation biology: concepts and challenges. *Conservation Genetics*, **11**, 375–385.

Seiler SM, Gunnell K, Ptacek MB, Keeley ER (2009) Morphological Patterns of Hybridization between Yellowstone Cutthroat Trout and Introduced Rainbow Trout in the South Fork of the Snake River Watershed, Idaho and Wyoming. *North American Journal of Fisheries Management*, **29**, 1529–1539.

Seiler SM, Keeley ER (2007a) A comparison of aggressive and foraging behaviour between juvenile cutthroat trout, rainbow trout and F1 hybrids. *Animal Behaviour*, **74**, 1805–1812.

Seiler SM, Keeley ER (2007b) Morphological and swimming stamina differences between Yellowstone cutthroat trout (Oncorhynchus clarkii bouvieri), rainbow trout (Oncorhynchus mykiss), and their hybrids. *Canadian Journal of Fisheries and Aquatic Sciences*, **64**, 127–135.

Seiler SM, Keeley ER (2009) Competition between native and introduced salmonid fishes: cutthroat trout have lower growth rate in the presence of cutthroat-rainbow trout hybrids. *Canadian Journal of Fisheries and Aquatic Sciences*, **66**, 133–141.

Shepard BB (2004) Factors that may be influencing nonnative brook trout invasion and their displacement of native westslope cutthroat trout in three adjacent southwestern Montana streams. *North American Journal of Fisheries Management*, **24**, 1088–1100.

Shepard BB, May BE, Urie W (2005) Status and conservation of westslope cutthroat trout within the western United States. *North American Journal of Fisheries Management*, **25**, 1426–1440.

Storfer A, Murphy MA, Evans JS *et al.* (2006) Putting the /`landscape/' in landscape genetics. *Heredity*, **98**, 128–142.

Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP (2010) Landscape genetics: where are we now? *Molecular Ecology*, **19**, 3496–3514.

Wang IJ, Savage WK, Bradley Shaffer H (2009) Landscape genetics and least-cost path analysis reveal unexpected dispersal routes in the California tiger salamander (Ambystoma californiense). *Molecular Ecology*, **18**, 1365–1374.

Weigel DE, Peterson JT, Spruell P (2003) Introgressive hybridization between native cutthroat trout and introduced rainbow trout. *Ecological Applications*, **13**, 38–50.

Wilson EB (1927) Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209–212.

Wofford JEB, Gresswell RE, Banks MA (2005) Influence of barriers to movement on within-watershed genetic variation of coastal cutthroat trout. *Ecological Applications*, **15**, 628–637.

Yamamoto S, Kitano S, Maekawa K, Koizumi I, Morita K (2006) Introgressive hybridization between Dolly Varden Salvelinus malma and white-spotted charr Salvelinus leucomaenis on Hokkaido Island, Japan. *Journal of Fish Biology*, **68**, 68–85.

APPENDICES

APPENDIX A

SUPPORTING MATERIAL FOR CHAPTER 2

Model Functions

In the modeling experiments outlined in this paper, we parameterized the simulations such that all fish had equal fitness and behavior regardless of genotype. However, by altering appropriate model parameters, our model can account for a range of genetically driven differences among native, non-native, and hybridized individuals. We provide a more complete model description here.

The model uses published observations or hypotheses to account for: (a) selective mating (e.g., De Rito2004, Muhlfeld *et al.* 2009c); (b) changes in the fecundity of mature females due to non-native ancestry (Muhlfeld *et al.* 2009a); (c) differences in the straying propensity of offspring due to their lineage (e.g., Hitt *et al.* 2003, Boyer *et al.* 2008), and; (d) differences in survival ability among native, non-native, and hybridized individuals (e.g., Seiler and Keeley 2007a, 2009). These features were designed to allow theoretical exploration of how interactions between patterns of landscape connectivity and genetically driven differences among fish interact to determine the spread of introgressive hybridization across a system through time.

Mate Selection

Our model allows the user to weight the likelihood of two fish mating, based on genetic differences between the two fish. This approach is intended to simulate selective mating, where native fish may be more likely to mate with other native fish, and non-native fish may be more likely to mate with non-native fish.

Let P_{m_f}represent the probability of male *m* being selected as a mate by female *f*. If mating were random, P_{m_f} would simply be 1/*n*; where *n* is the number of males at the spawning site occupied by *f*. However, in our model, P_{m_f} can be adjusted by a weighting factor, ρ_{m_f}, which varies between 0 and 1. When $\rho_{m_f} = 0$, there is no chance for male *m* to mate with female *f*, whereas when $\rho_{m_f} = 1$, mating is random and the chance of male *m* to mate with female *f* is unweighted. Thus:

$$P_{m_f} = \rho_{m_f} / \sum_{i=1}^{n} \rho_{m_{f_i}}$$

In the model, ρ_{m_f} is determined from the difference in the lineage, *L*, between male *m* and female *f* (see *Fish Population and Life-Cycle Simulation* in the main text for a definition of lineage). As the lineage of the female, L_f, diverges from the lineage of the male, L_m, the model assumes a linear decrease in ρ_{m_f}, starting at 1 (when $L_m = L_f$), and decreasing toward some user defined minimum value, ρ_{min}, defined as the propensity to crossbreed for mating between a pure native and pure-nonnative fish (i.e., when $|L_f - L_m| = 1$):

$$\rho_{m_f} = (\rho_{min} - 1)(|L_f - L_m|) + 1$$

In all model runs described in the main body of this paper, ρ_{min} was set equal to 1, which yields random mating regardless of lineage.

Females' Fecundity

Our model accounts for the potential reduction in female fecundity as a function of increased hybridization. Specifically, the number of viable offspring produced by a female each year (Ω_f) is calculated as:

$$\Omega_f = \Omega_{max}\left[1 - L_f(1-\tau)\right]$$

where Ω_{max} represents the user-defined number of viable offspring produced by a genetically pure native female (i.e., with $L = 0$), L_f is the lineage of the female, and τ is a user-defined parameter ($0 \leq \tau < 1$) that quantifies the fraction of Ω_{max} produced by a genetically pure non-native female (i.e., with $L_f = 1$), and thus determines the slope of the linear relationship between Ω_f and L_f. For the purposes of this paper, τ was set to 1 such that all females produce the same number of viable offspring ($\Omega_{max} = 50$) regardless of lineage.

Also, our model also allows the user to conduct simulations accounting for an exponential decline in female fecundity with hybridization, as observed for CTT females introgressed with RBT in the North Fork of the Flathead River in Montana (Muhlfeld *et al.* 2009a). If this option is selected, then the number of viable offspring per female is calculated as

$$\Omega_f = \begin{cases} \Omega_{max} e^{-3.040 L_f} & L_f \neq 0.5 \\ \Omega_{max} & L_f = 0.5 \end{cases}$$

Muhlfeld *et al.* (2009a) observed that first generation hybrids (i.e., $L_f = 0.5$), had similar fecundity as genetically pure native females, and therefore the model sets $\Omega_f = \Omega_{max}$ for females with lineage indicative of first generation hybrids.

Straying

Our model was designed to account for increased straying propensity in juvenile individuals with non-native genes. Specifically, our model assumes that the probability of a juvenile to stray (P_i) increases linearly with its L_i to a maximum a_{max}. Thus, for each juvenile, the model calculates P_i as:

$$\mathrm{P}_i = a_n + L_i(a_{max} - a_n)$$

where $a_n$is the user-defined probability ($0 < a_n \leq 1$) that a pure native individual will stray. For the purposes of this paper, we assumed no genetically driven differences among fish relative to their straying propensity. We set $a_{max} = a_n$ which yields $\mathrm{P}_i = a_n$ for all fish, regardless of their lineage, and we set $a_n = 0.05$ yielding a 5% chance for all maturing fish to stray.

Survival

Similar to *Mate Selection*, our model allows the user to weight the chance of a given fish to be selected to grow to the next year class based on its lineage. As for P_{m_f}, the probability of individual *i* to survive to the next year class (S_i) is calculated in the model as

$$S_i = \sigma_i / \sum_{i=1}^{n} \sigma_i$$

where σ_i is the survival ability of individual *i*, and *n* is the total number of individuals of the same age as individual *i*. To account for increased or decreased survival ability due to non-native ancestry, σ_i is defined in the model as:

$$\sigma_i = 1 + s(1 - L_i)$$

where L_i is the lineage of individual i and s is a user-defined number that quantifies the difference in survival ability between pure native individuals and pure non-native individuals and can have any value between -1 (exclusive) and infinity. In its current form, σ_i assumes a linear relationship between the lineage of an individual and its ability to survive. When $s < 0$, a native individual has lower σ_i and consequently lower S_i than non-native individuals ($L_i > 0$). When $s = 0$, survival is random and all fish have equal σ_i and $S_i = 1/n$, regardless of their lineage (where n is equal to the total number of individual of the same year class as individual i). When $s > 0$, a native individual has higher σ_i and consequently higher S_i than non-native individuals ($L_i > 0$). The negative range of s is limited to -1 exclusive, since, due to the equation describing σ_i, when $s \leq -1$ native fish would have a negative survival ability, which of course is unrealistic. For the purposes of this paper, we set s = 0 such that survival is uniformly random, regardless of lineage.

Species Diagnostic Markers

For each simulated fish, the model can simulate a user-defined number of species diagnostic markers (Hohenlohe *et al.* 2011, Kalinowski *et al.* 2011). Each diagnostic marker is composed of two alleles that can be of type native or non-native. All diagnostic markers are assumed to be independent from each other (i.e., no linkage disequilibrium) and for each diagnostic marker offspring receive one allele from their father and one allele from their mother, according to Mendelian genetics.

Using the diagnostic markers of the simulated fish, the model can calculate individuals' admixture with non-native genes based on m diagnostic markers (a_m) as the

percentage of non-native alleles in the fish genome. Individuals' a_m can be used in all of the above reported equations instead of L_i to make the fitness and the behavior of that fish a function of diagnostic markers rather than lineage.

APPENDIX B

SUPPORTING FIGURES AND TABLES TO CHAPTER 3

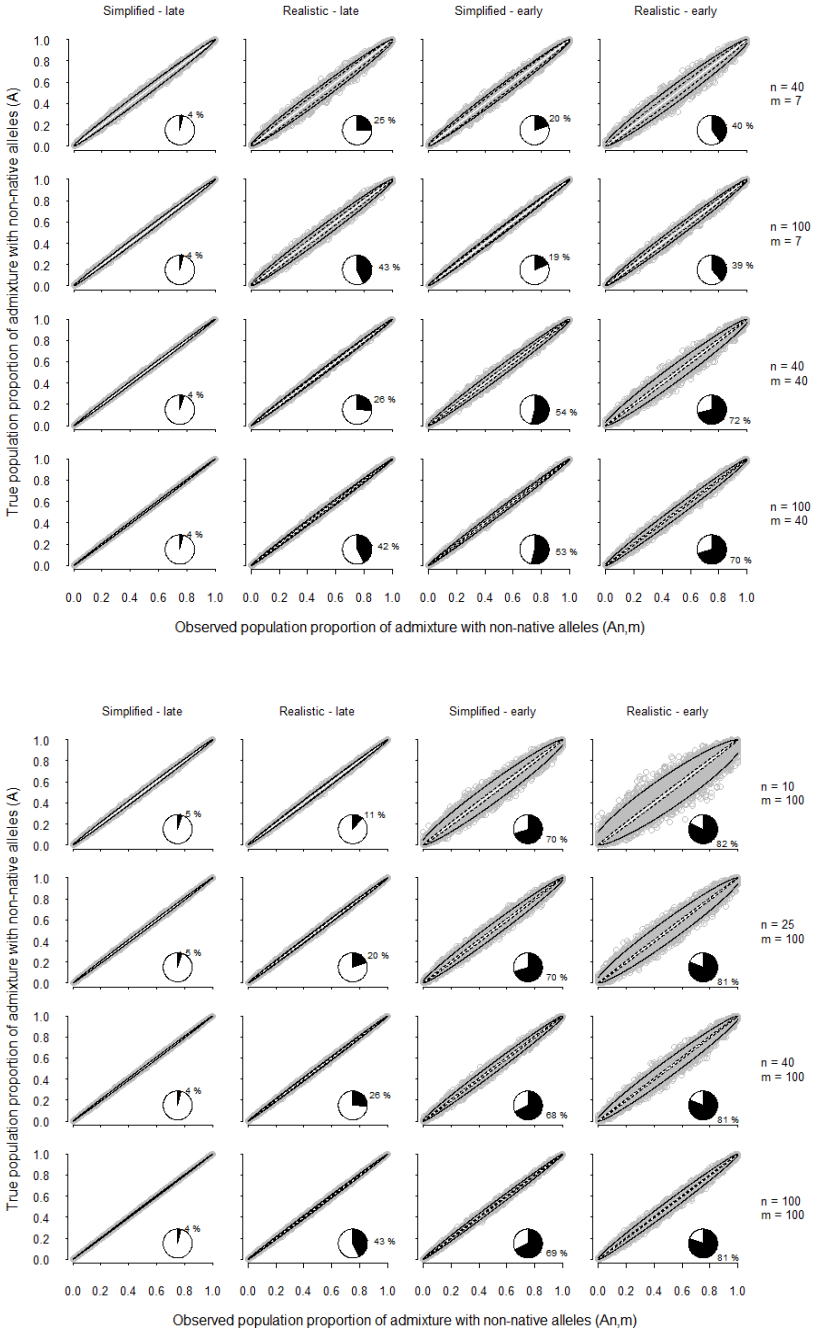

Figure B-1a-b: Model-derived (solid line) and equation-derived (dashed line) uncertainty (CI95A) associated with observed proportions of population admixture (An,m, grey circles) for the following n-m combinations: (Appendix 1a) 40-7 ; 40-40; 100-7; 100-40 and (Appendix 1b): 10-100; 25-100; 40-100; 100-100. Note that PSn,m vs. PEn,m results for these combinations of n and m were omitted because PEn,m was <0.001, regardless of the true proportion of admixture of the population(A).

Table A-3: Levels of admixture (An,m) with non-native rainbow trout genes and standard deviation (SD) from streams with An,m > 0 reported in Boyer et al. (2008) along with associated 95% confidence interval (CI95FA). Estimates of CI95EA were obtained using n and m as in Boyer et al. (2008), while n and m for CI95SA are reported in the table. Note that in Boyer et al. (2008) An,m is referred to as q. In Boyer et al. (2008), An,m and SD values were calculated using the program STRUCTURE (Pritchard et al. 2000).

			Boyer et al. (2008)						**Equation 2**		**Our model**			
							$95CI^F_A$		**$95CI^E_A$**				**$95CI^S_A$**	
	Code	**Site**	**n**	**m**	**$A_{n,m}$**	**SD**	**Upper**	**Lower**	**Upper**	**Lower**	**n**	**m**	**Upper**	**Lower**
Populations with non-random distribution of RBT genes	7	Nicola	32	7	0.018	0.04	0.032	0.004	0.030	0.006	25	7	0.139	0.002
	12	Meadow	25	7	0.035	0.06	0.059	0.011	0.054	0.016	25	7	0.165	0.007
	13	Cyclone	24	7	0.116	0.15	0.177	0.056	0.150	0.082	25	7	0.271	0.044
	21	Upper R.M.	24	7	0.122	0.2	0.202	0.042	0.157	0.087	25	7	0.279	0.048
	11	Anaconda	31	7	0.206	0.22	0.283	0.128	0.244	0.168	25	7	0.375	0.101
	3	Rabe	30	7	0.491	0.17	0.552	0.430	0.539	0.443	25	7	0.657	0.327
	2	Ivy	20	7	0.493	0.32	0.633	0.353	0.551	0.434	25	7	0.659	0.329
	4	Third	19	7	0.658	0.33	0.807	0.510	0.715	0.601	25	7	0.798	0.483
Populations with random distribution of RBT genes	19	S.F. Red Mead.	26	7	0.003	0.01	0.007	-0.001	0.009	-0.003	25	7	0.054	0.000
	15	S.F. Coal	26	7	0.006	0.02	0.014	-0.002	0.013	-0.002	25	7	0.059	0.001
	10	Trout	42	7	0.01	0.02	0.016	0.004	0.018	0.002	40	7	0.057	0.002
	25	Tepee	32	7	0.013	0.03	0.023	0.002	0.023	0.002	25	7	0.071	0.002
	17	Lower Hay	25	7	0.014	0.05	0.033	-0.006	0.026	0.002	25	7	0.072	0.003
	20	Lower R.M.	23	7	0.022	0.07	0.051	-0.007	0.038	0.006	25	7	0.084	0.005
	9	Dutch	32	7	0.13	0.13	0.175	0.085	0.161	0.099	25	7	0.224	0.072
	5	Langford	30	7	0.331	0.27	0.428	0.234	0.376	0.286	25	7	0.443	0.236
	1	Abbot	35	7	0.916	0.01	0.919	0.913	0.940	0.891	40	7	0.957	0.842

APPENDIX C

SUPPORTING TABLES TO CHAPTER 4

Table A-4: Number of individuals (n) to sample and diagnostic markers (m)to analyze necessary to detect non-native alleles in a population for different Invasion intensities (proportion of alleles in the gene pool that are of non-native origins) and levels of confidence (Reliability).

Invasion intensity = 0.01, Reliability = 0.99

	n						
	no F-2+	F-2+ present					
		Admixture of F-2+					
m		0.005	0.01	0.025	0.05	0.1
7		6794	3507	1541	897	595
10		4826	2527	1157	716	522
20		2533	1389	721	527	466
40		1392	832	529	466	459
60	459	1017	655	482	460	459
80		833	574	467	459	459
100		726	530	462	459	459
200		530	467	459	459	459
500		462	459	459	459	459

Invasion intensity = 0.02, Reliability = 0.99

	n						
	no F-2+	F-2+ present					
		Admixture of F-2+					
m		0.005	0.01	0.025	0.05	0.1
7		3396	1752	770	448	297
10		2412	1263	578	357	260
20		1266	694	360	262	232
40		695	415	263	232	228
60	228	508	327	240	229	228
80		416	286	233	229	228
100		362	264	230	228	228
200		264	233	228	228	228
500		230	228	228	228	228

Invasion intensity = 0.01, Reliability = 0.95

	n						
	no F-2+	F-2+ present					
		Admixture of F-2+					
m		0.005	0.01	0.025	0.05	0.1
7		4419	2281	1003	584	387
10		3140	1644	753	466	340
20		1648	904	469	343	303
40		906	541	344	304	299
60	299	662	427	314	299	299
80		542	374	304	299	299
100		472	345	300	299	299
200		345	304	299	299	299
500		301	299	299	299	299

Invasion intensity = 0.02, Reliability = 0.95

	n						
	no F-2+	F-2+ present					
		Admixture of F-2+					
m		0.005	0.01	0.025	0.05	0.1
7		2209	1140	501	291	193
10		1569	822	376	232	170
20		823	451	234	171	151
40		452	270	172	151	149
60	149	330	213	156	149	149
80		271	186	151	149	149
100		236	172	150	149	149
200		172	152	149	149	149
500		150	149	149	149	149

Invasion intensity = 0.01, Reliability = 0.90

	n						
	no F-2+	F-2+ present					
		Admixture of F-2+					
m		0.005	0.01	0.025	0.05	0.1
7		3397	1754	771	449	298
10		2413	1264	579	358	261
20		1267	695	361	264	233
40		696	416	265	233	230
60	230	509	328	241	230	230
80		417	287	234	230	230
100		363	265	231	230	230
200		265	234	230	230	230
500		231	230	230	230	230

Invasion intensity = 0.02, Reliability = 0.90

	n						
	no F-2+	F-2+ present					
		Admixture of F-2+					
m		0.005	0.01	0.025	0.05	0.1
7		1698	876	385	224	149
10		1206	632	289	179	130
20		633	347	180	131	116
40		348	208	132	116	114
60	114	254	164	120	115	114
80		208	143	117	115	114
100		181	132	115	114	114
200		132	117	114	114	114
500		115	114	114	114	114

Invasion intensity = 0.01, Reliability = 0.75

	n						
	no F-2+	F-2+ present					
		Admixture of F-2+					
m		0.005	0.01	0.025	0.05	0.1
7		2045	1056	464	270	180
10		1453	761	349	216	158
20		763	419	218	159	141
40		419	251	160	141	138
60	138	306	198	145	139	138
80		251	173	141	138	138
100		219	160	139	138	138
200		160	141	138	138	138
500		139	138	138	138	138

Invasion intensity = 0.02, Reliability = 0.75

	n						
	no F-2+	F-2+ present					
		Admixture of F-2+					
m		0.005	0.01	0.025	0.05	0.1
7		1023	528	232	135	90
10		726	380	174	108	79
20		381	209	109	79	70
40		210	125	80	70	69
60	69	153	99	73	69	69
80		125	86	70	69	69
100		109	80	70	69	69
200		80	70	69	69	69
500		70	69	69	69	69

Invasion intensity = 0.05
Reliability = 0.99

m	n (no F-2+)	n (F-2+ present) Admixture of F-2+ 0.005	0.01	0.025	0.05	0.1
7		1357	700	307	178	118
10		964	504	230	142	103
20		505	276	143	104	92
40		277	165	104	92	90
60	90	202	130	95	90	90
80		165	113	92	90	90
100		144	105	91	90	90
200		105	92	90	90	90
500		91	90	90	90	90

Invasion intensity = 0.10
Reliability = 0.99

m	n (no F-2+)	n (F-2+ present) Admixture of F-2+ 0.005	0.01	0.025	0.05	0.1
7		678	349	152	88	58
10		481	251	114	70	51
20		252	137	70	51	45
40		138	82	51	45	44
60	44	100	64	47	44	44
80		82	56	45	44	44
100		71	51	45	44	44
200		51	45	44	44	44
500		45	44	44	44	44

Invasion intensity = 0.05
Reliability = 0.95

m	n (no F-2+)	n (F-2+ present) Admixture of F-2+ 0.005	0.01	0.025	0.05	0.1
7		883	455	200	116	77
10		627	328	150	92	67
20		329	180	93	68	60
40		180	107	68	60	59
60	59	132	85	62	59	59
80		108	74	60	59	59
100		94	68	59	59	59
200		68	60	59	59	59
500		59	59	59	59	59

Invasion intensity = 0.10
Reliability = 0.95

m	n (no F-2+)	n (F-2+ present) Admixture of F-2+ 0.005	0.01	0.025	0.05	0.1
7		441	227	99	57	38
10		313	164	74	46	33
20		164	89	46	33	29
40		90	53	33	29	29
60	29	65	42	30	29	29
80		53	36	29	29	29
100		46	34	29	29	29
200		34	29	29	29	29
500		29	29	29	29	29

Invasion intensity = 0.05
Reliability = 0.90

m	n (no F-2+)	n (F-2+ present) Admixture of F-2+ 0.005	0.01	0.025	0.05	0.1
7		679	350	154	89	59
10		482	252	115	71	52
20		253	138	72	52	46
40		139	83	52	46	45
60	45	101	65	48	45	45
80		83	57	46	45	45
100		72	53	46	45	45
200		53	46	45	45	45
500		46	45	45	45	45

Invasion intensity = 0.10
Reliability = 0.90

m	n (no F-2+)	n (F-2+ present) Admixture of F-2+ 0.005	0.01	0.025	0.05	0.1
7		339	175	76	44	29
10		241	126	57	35	26
20		126	69	35	26	23
40		69	41	26	23	22
60	22	50	32	24	22	22
80		41	28	23	22	22
100		36	26	23	22	22
200		26	23	22	22	22
500		23	22	22	22	22

Invasion intensity = 0.05
Reliability = 0.75

m	n (no F-2+)	n (F-2+ present) Admixture of F-2+ 0.005	0.01	0.025	0.05	0.1
7		409	211	93	54	36
10		290	152	70	43	31
20		152	84	43	32	28
40		84	50	32	28	28
60	28	61	39	29	28	28
80		50	34	28	28	28
100		44	32	28	28	28
200		32	28	28	28	28
500		28	28	28	28	28

Invasion intensity = 0.10
Reliability = 0.75

m	n (no F-2+)	n (F-2+ present) Admixture of F-2+ 0.005	0.01	0.025	0.05	0.1
7		204	105	46	27	18
10		145	76	35	21	16
20		76	42	22	16	14
40		42	25	16	14	14
60	14	30	20	14	14	14
80		25	17	14	14	14
100		22	16	14	14	14
200		16	14	14	14	14
500		14	14	14	14	14

Invasion intensity = 0.20 Reliability = 0.99							Invasion intensity = 0.25 Reliability = 0.99						
		n							n				
	no F-2+	F-2+ present						no F-2+	F-2+ present				
		Admixture of F-2+							Admixture of F-2+				
m		0.005	0.01	0.025	0.05	0.1	m		0.005	0.01	0.025	0.05	0.1
7		338	174	75	43	28	7		270	139	60	34	22
10		240	125	56	34	24	10		191	99	45	27	19
20		125	68	34	25	21	20		100	54	27	19	17
40		68	40	25	22	21	40		54	31	19	17	17
60	21	49	31	22	21	21	60	17	39	24	17	17	17
80		40	27	22	21	21	80		32	21	17	17	17
100		35	25	21	21	21	100		27	19	17	17	17
200		25	22	21	21	21	200		19	17	17	17	17
500		21	21	21	21	21	500		17	17	17	17	17

Invasion intensity = 0.20 Reliability = 0.95							Invasion intensity = 0.25 Reliability = 0.95						
		n							n				
	no F-2+	F-2+ present						no F-2+	F-2+ present				
		Admixture of F-2+							Admixture of F-2+				
m		0.005	0.01	0.025	0.05	0.1	m		0.005	0.01	0.025	0.05	0.1
7		220	113	49	28	18	7		176	90	39	22	14
10		156	81	37	22	16	10		125	65	29	18	13
20		81	44	22	16	14	20		65	35	18	13	11
40		44	26	16	14	14	40		35	21	13	11	11
60	14	32	20	15	14	14	60	11	25	16	12	11	11
80		26	18	14	14	14	80		21	14	11	11	11
100		23	16	14	14	14	100		18	13	11	11	11
200		16	14	14	14	14	200		13	11	11	11	11
500		14	14	14	14	14	500		11	11	11	11	11

Invasion intensity = 0.20 Reliability = 0.90							Invasion intensity = 0.25 Reliability = 0.90						
		n							n				
	no F-2+	F-2+ present						no F-2+	F-2+ present				
		Admixture of F-2+							Admixture of F-2+				
m		0.005	0.01	0.025	0.05	0.1	m		0.005	0.01	0.025	0.05	0.1
7		169	87	38	22	14	7		135	70	30	17	11
10		120	63	28	17	12	10		96	50	23	14	10
20		63	34	17	13	11	20		50	27	14	10	9
40		34	20	13	11	11	40		27	16	10	9	9
60	11	25	16	11	11	11	60	9	20	12	9	9	9
80		20	14	11	11	11	80		16	11	9	9	9
100		18	13	11	11	11	100		14	10	9	9	9
200		13	11	11	11	11	200		10	9	9	9	9
500		11	11	11	11	11	500		9	9	9	9	9

Invasion intensity = 0.20 Reliability = 0.75							Invasion intensity = 0.25 Reliability = 0.75						
		n							n				
	no F-2+	F-2+ present						no F-2+	F-2+ present				
		Admixture of F-2+							Admixture of F-2+				
m		0.005	0.01	0.025	0.05	0.1	m		0.005	0.01	0.025	0.05	0.1
7		102	53	23	13	9	7		82	42	18	11	7
10		72	38	17	11	8	10		58	30	14	8	6
20		38	21	11	8	7	20		30	17	8	6	5
40		21	12	8	7	7	40		17	10	6	5	5
60	7	15	10	7	7	7	60	5	12	8	6	5	5
80		12	8	7	7	7	80		10	7	5	5	5
100		11	8	7	7	7	100		9	6	5	5	5
200		8	7	7	7	7	200		6	5	5	5	5
500		7	7	7	7	7	500		5	5	5	5	5