SPATIOTEMPORAL POPULATION GENOMICS OF MARINE SPECIES: INVASION, EXPANSION, AND CONNECTIVITY

by

Eleanor Kathleen Bors

B.A., Oberlin College (2009)
B.Mus., Oberlin Conservatory of Music (2009)

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

and the

WOODS HOLE OCEANOGRAPHIC INSTITUTION

February 2017

Signature of Author

Joint Program in Oceanography/Applied Ocean Science and Engineering
Massachusetts Institute of Technology
and Woods Hole Oceanographic Institution
January 23, 2017

Certified by

Dr. Timothy M. Shank
Associate Scientist with Tenure
Woods Hole Oceanographic Institution, Biology Department
Thesis Supervisor

Accepted by

Dr. Ann M. Tarrant
Chair, Joint Committee for Biological Oceanography
Massachusetts Institute of Technology
Woods Hole Oceanographic Institution

SPATIOTEMPORAL POPULATION GENOMICS OF MARINE SPECIES: INVASION, EXPANSION, AND CONNECTIVITY

by

Eleanor Kathleen Bors

Submitted to the MIT-WHOI Joint Program in Oceanography and Applied Ocean Science and Engineering in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Biological Oceanography

ABSTRACT

Every genome tells a story. This dissertation contains four such stories, focused on shared themes of marine population dynamics and rapid change, with an emphasis on invasive marine species. Biological invasions are often characterized by a range expansion, during which strong genetic drift is hypothesized to result in decreased genetic diversity with increased distance from the center of the historic range, or the point of invasion. In this dissertation, population genetic and genomic tools are used to approach complex and previously intractable fundamental questions pertaining to the non-equilibrium dynamics of species invasions and rapid range expansions in two invasive marine species: the lionfish, *Pterois volitans*; and the shrimp, *Palaemon macrodactylus.* Using thousands of loci sequenced with restriction enzyme associated DNA sequencing in these two systems, this research tests theoretical predictions of the genomic signatures of range expansions. Additionally, the first chapter elucidates patterns of population genetic connectivity for deep-sea invertebrates in the New Zealand region demonstrating intimate relationships between genetics, oceanographic currents, and life history traits. Invasive shrimp results extend our understanding of marine population connectivity to suggest that human-mediated dispersal may be as important—if not more important—than oceanographic and life history considerations in determining genetic connectivity during specific phases of marine invasions. In invasive populations of lionfish, measures of genomic diversity, including a difference between observed and expected heterozygosity, were found to correlate with distance from the point of introduction, even in the absence of spatial metapopulation genetic structure. These results indicate a signal of rapid range expansion. The final study in this dissertation uses an innovative temporal approach to explore observed genomic patterns in the lionfish. In all, this dissertation provides a broad perspective through the study of multiple species undergoing superficially parallel processes that, under more intense scrutiny, are found to be mechanistically unique. It is only through comparative approaches that predictable patterns of population dynamics will emerge.

Thesis Supervisor: Dr. Timothy M. Shank
Title: Associate Scientist with Tenure, Woods Hole Oceanographic Institution

For my parents Kathy and Doug
and my brother Loren

ACKNOWLEDGEMENTS

I would like to thank the numerous funding sources that provided support for me and my research throughout graduate school. I was privileged to receive support through a National Science Foundation (NSF) Graduate Research Fellowship (Grant No. 1122374), a Woods Hole Oceanographic Institution (WHOI) Fellowship administered through the Academic Program Office, funding from NSF award OCE-1131620, and the National Oceanic and Atmospheric Administration (NOAA OER #NA08OAR4600757). Research funding for sample collection, laboratory materials, sequencing, and analysis was provided by the WHOI Coastal Ocean Institute, WHOI Ocean Ventures Fund, WHOI James Education Fund for Ocean Exploration, WHOI Biology Educational Fund, Woods Hole Sea Grant New Initiative Award, Massachusetts Office of Coastal Zone Management, Northeast Sea Grant Consortium, and the PADI Foundation (Grant No. 14904). Additional research support was provided by the NSF (OCE-1131620) to Timothy M. Shank. I was able to present this research at multiple national and international meetings through additional support from the WHOI Academic Programs Office, WHOI Coastal Ocean Institute, MIT Student Assistance Fund, and MIT Graduate Student Council Travel Fund.

First and foremost, I want to thank my advisor, Tim Shank. Through the difficult, victorious, and frequently absurd moments of graduate school, Tim was there with me, commiserating, celebrating, and laughing as each moment unfolded. I am profoundly thankful for the opportunities I was given in the Shank Lab including several fantastic cruises, extensive experience with telepresence, the opportunity to travel to enriching and important international workshops and conferences, and most of all, the opportunity to pursue research for which I have tremendous passion.

I benefited greatly from working with all the other members of the Shank Lab, past and present, who kept me smiling along the way. I would like to especially thank Cat Munro for her friendship and dedication to science, and Santiago Herrera who paved the way for me to do non-model genomic research. Santiago's friendship and mentorship made this work possible.

I am indebted to my thesis committee, Ann Tarrant, John Wakeley, and Eric Alm, who provided me with both intellectual and personal support as I took my nebulous interests and sculpted them into a compelling research portfolio. I deeply appreciate the advice that my committee provided through the years. I would also like to thank my academic advisory committee, Julie Kellner and Penny Chisholm, for helping me get through the pre-general exam phase of graduate school. All of our successes as graduate students at WHOI depend on the support of JCBO and the Academic Programs Office—thank you for keeping the wheels greased, and helping us through the process. A special thank you goes to Michael Neubert (chair of my dissertation defense) and Becky Gast who provided me with useful advice and conversation throughout the years. I would also like to thank Jim Carlton from the Williams-Mystic program for welcoming me to the world of invasive species research.

While this thesis is the result of tremendous scientific effort and focus, I have been lucky to maintain and cultivate other passions during graduate school.

My cello, Zoltan, has been a good companion through the years and I am thankful to the MIT Chamber Music Society and the Emerson Fellowship Program for providing me with musical opportunities. Through these programs and with the support of Marcus Thompson and David Deveau, I was able to nurture my musical interests. Thank you to my many musical collaborators. Music nourishes me and without it, I wither.

MIT has a bustling science policy community and I would like to thank Susan Solomon and Ken Oye for stimulating my thought processes and challenging me to think about the environment from a broader policy perspective.

I would like to thank the Fulbright Program for changing my life and making me a better person. Thank you to my beloved Aotearoa: your cultural, social, and natural aesthetics inspire me.

Thank you to the Woods Hole community, especially my friends at the Woods Hole Yacht Club. I can think of nothing that made me happier after a long day of work than a vigorous beat up the Nonamesset shoreline and a battle against the Woods Hole currents on the way home.

A rich life is nothing without good people to share it with. So now, I turn to my dear friends. I was lucky to have two outstanding cohorts of joint program students to claim as my colleagues, helping me to grow and learn. Graduate school allowed me to take my time growing up. What a privilege it has been to do so alongside so many brilliant and inspiring people.

I am surrounded by bright, inquisitive, determined, hilarious, and strong women who have helped shape me into the person I am and who will, inevitably, continue to challenge me to grow into a better human being with every passing day. Through breakups, breakthroughs, and breakdowns, they have been there, embracing my quirks and sense of humor. Thank you. To my Oberlin ladies in particular: I cannot imagine a more different assortment of X chromosomes. And yet here we all are, loving and supporting each other as we hurtle into adulthood.

And finally, I turn to my family, for whom the following words are insufficient.

The Bors Family Unit is a force to be reckoned with. We are a group of true characters and comics. From an early age, my parents instilled in me an unwavering appreciation for knowledge, art, and creativity. Our house was full of books, deep thoughts, and laughter. The ideal of *the intellectual* seemed to me an elusive dream. To dedicate one's life to knowledge and learning was the pinnacle of success. As every child yearns to please their parents, I too have been driven by a need to reach this ideal; to become a generator, keeper, and interpreter of knowledge. The good news is that it suits me. The other good news is that if it didn't, I know that I would still feel the endless love and support of my family wherever I go and whatever I do. All my love goes to them.

TABLE OF CONTENTS

CHAPTER 1

Introduction

I. PREAMBLE

This dissertation is a synthesis of concepts from population genetics, evolutionary biology, invasive species biology, oceanography, genomics, global change science, and conservation biology. While the technical substance of the research presented here is built on scaffolding of theoretical evolutionary biology and population genetics—specifically on models of gene flow, spatial dynamics, and evolutionary change—always in the background is the cognizance of a changing planet and of the life affected by that change. In the modern world, humans are part of every ecosystem. Now, at the outset of 2017 more than ever, scientific inquiry that probes the ways in which ecosystems respond to anthropogenic influence is essential to our understanding of the world around us. The necessity of this understanding does not lie solely in the perfect joy of discovery, but rather arises from an urgent need to manage resources better so we can feed a growing human population, to adapt our behaviors in order to ensure continued ecosystem integrity and function for human use, and to preserve critical marine and terrestrial habitats in order to sustain a clean, healthy environment for centuries to come.

The empirical results presented in the following four chapters lead to specific conclusions about specific systems, but they also contribute to a larger puzzle about the way marine species disperse in the world's oceans and how those processes affect evolution. In this introductory chapter, brief descriptions of the major topics of the dissertation are presented in "Background and Motivation," followed by a short summary of the contents of each chapter in the "Dissertation Overview."

II. BACKGROUND AND MOTIVATION

Genetic connectivity

Connectivity is a word that has numerous connotations across different sub-disciplines of biology. It means something slightly different on land than it does in the sea, and within the marine realm it can be used by biophysical modelers to imply different processes than those evoked by population geneticists. In these pages, the phrases *genetic connectivity*, and *population genetic connectivity* are defined as "the dispersal, survival, and reproduction of migrants, so that they contribute to the local gene pool" (Hedgecock & Barber 2007). Genetic connectivity encompasses temporal and spatial aspects of population genetics in order to infer the degree of genetic exchange among populations. It is, almost by definition, an averaged genetic signature of population processes through time and often describes patterns of population genetics within a specific region and the factors that could give rise to observed patterns.

Current anthropogenic pressures on the marine environment, including the deep sea, are unprecedented (Miles 2009; Barange *et al.* 2010; Ramirez-Llodra *et al.* 2011; Van Dover *et al.* 2012). To counteract anthropogenic impacts, many national and international regulations have been enacted to protect portions of the terrestrial and marine environment through limiting resource extraction and closing certain areas to human use (*e.g.,* New Zealand Biodiversity Strategy, 2000). However, the creation of such areas is often a political, social, and scientific challenge (Gaines *et al.* 2010a) because of competing priorities of different stakeholders. The goal of many area closures is to protect biodiversity at the genetic, species, and ecosystem levels. With such a broad goal, defining success and measuring the efficacy of closed area design is difficult. The development of genetic tools for evaluation of closed areas is an important step toward improving the design of closed areas as well as increasing their utility for biodiversity management and conservation. Genetic connectivity is considered important to marine protected area (MPA) design and evaluation (Palumbi 2003; Miller & Ayre 2008; Shank 2010; Gaines *et al.* 2010b) because it illuminates where the sources and sinks of marine populations are in order to effectively protect the diversity maintained by source-sink dynamics (Gaines *et al.* 2010b).

Physical oceanographic forces and life history traits work in concert to shape patterns of genetic connectivity. Oceanic currents are major drivers of larval dispersal in the oceans, sometimes determining population genetic structure patterns (White *et al.* 2010). Environments with high levels of oceanographic mixing and species with long distance larval dispersal—even if only stochastic—lead to lower levels of population structure and, notably, to genetic patterns that do not correlate with Euclidian distance (Cowen *et al.* 2007; White *et al.* 2010). However, local retention, high larval mortality, and other drivers of recruitment dynamics can result in unexpected or counterintuitive connectivity patterns. This phenomenon has been particularly well studied in reef systems where self-recruitment has been shown to be high in some cases (Jones *et al.* 2005). Indeed, reproductive strategies, dispersal capabilities, and other life history traits also play a crucial role in determining population genetic connectivity and geographic spread in many marine species (Selkoe *et al.* 2016).

Despite the inherent stochasticity of marine connectivity and the variability of anthropogenic disturbances, biologists often make equilibrium assumptions that impose stability on systems, behaviors, or dynamic processes, frequently resulting in the averaging of data across many generations. Population genetics and evolutionary biology of marine species, for example, often rely on inferring historical events from observed patterns, which entails describing the spatial distribution of genetic diversity at a fixed point in time and reconstructing a probable history of dynamic processes that could lead to the observed patterns. Doing so can result in the loss of temporal signals that could be highly variable. Sometimes this simplification is necessary for theoretical work; however, these assumptions (*e.g.*, those of Hardy Weinberg Equilibrium, including static population size, random mating, *etc.*) rarely apply in natural systems. As technologies and methods develop, biologists are able to abandon assumptions of equilibrium and push the theoretical envelope. One example of a process that violates equilibrium assumptions is rapid range expansion of species or populations.

Range expansions and distributional shifts

Populations of marine species are dynamic. They expand, contract, and fluctuate in density and distribution over many temporal and spatial scales; and they exist within a larger biological community and physical environment. Global change is affecting species distributions in both terrestrial (Parmesan & Yohe 2003; Thomas *et al.* 2004; Parmesan 2006; Sunday *et al.* 2012) and marine systems (Parmesan & Yohe 2003; Perry 2005; Sabatés *et al.* 2006; Sorte *et al.* 2010; Booth *et al.* 2011; Jones & Southward 2012; Sunday *et al.* 2012; Poloczanska *et al.* 2013). The resulting shifts in range boundaries are hypothesized to alter the population genomics of affected species in ways that will change adaptive potential and shape biodiversity (Excoffier *et al.* 2009). In fact, range expansion has been linked to a decreased response to selection in some systems (Pujol & Pannell 2008).

Characterizing how range expansions will affect the genetic diversity of populations will greatly improve our ability to predict the future resilience of species of ecological and economic importance. Theoretical literature addressing range expansion genetics has become more prominent in the last decade, but there remains little empirical research describing the genomic legacies of rapid range expansions in natural populations, especially within the marine realm. Without empirical studies like those in this dissertation, the applicability of hypotheses generated by theoretical models and simulations to natural systems cannot be evaluated. The potential for new tools, developed in the field of non-model population genomics and used throughout this dissertation, to address range expansion questions has been widely acknowledged (Kirk *et al.* 2013; Barrett 2014; Bock *et al.* 2014).

Range expansions are known to result in specific genetic consequences (Excoffier *et al.* 2009). The process that dominates much of the literature is known as “allele surfing” (alternatively called “gene surfing” or “mutation surfing”), in which a rare allele or new mutation rises to high frequency near a range margin because of repeated founder effects (*e.g.*, a random subsampling of the larger population that results in a decrease in genetic diversity) through space and through time (Edmonds *et al.* 2004; Klopfstein 2005; Hallatschek & Nelson 2008; Peischl *et al.* 2013). Depending on the demographic variables of the invasion—carrying capacity, growth rate, density, and dispersal—the strength of allele surfing may vary (Klopfstein 2005). In cases of strong allele surfing, the mutation or allele in question may become fixed at the range edge, even when the allele is disadvantageous (Travis *et al.* 2007; Peischl *et al.* 2013; Peischl & Excoffier 2015). Laboratory experiments have shown how allele surfing can result in sectored microbial growth patterns, which is a result of the fixation of an allele on any given expansion axis (Hallatschek *et al.* 2007). Furthermore, the shape and nature of the habitat (fragmented, containing obstacles) has been shown to change the way the process unfolds in more complex systems (Möbius *et al.* 2015).

A major limitation when trying to apply predictions from existing range expansion genetic theory and laboratory work to a natural system—especially to marine systems—is that models and simulations for expansion assume limited dispersal capabilities and adult migration (*e.g.*, assuming constant, small-distance dispersal (Edmonds *et al*. 2004)). Marine species, however, exhibit diverse life history strategies and dispersal capabilities that are shaped by their interaction with the dynamic ocean environment (see the genetic connectivity review above). Therefore, there is tremendous value in collecting and analyzing empirical data to test the predictions of range expansion theory. This dissertation generates some of the first data to achieve this goal. Chapters 3, 4, and 5 use invasive marine species to address the question: Do range-expanding metapopulations of marine species retain a signature of range expansion (*e.g.,* clines in genetic diversity)?

Species invasions from an ecological and evolutionary perspective

Biological invasions threaten biodiversity in terrestrial, aquatic, and marine ecosystems (Bax *et al.* 2003; Lowry *et al.* 2013; Thomaz *et al.* 2014). Invasive species frequently impede the conservation of global biodiversity by (1) driving other species locally or globally extinct through either competition or predation, (2) driving shifts in distributions of native species, or (3) degrading and altering habitats (Mainka & Howard 2010). For these ecological impacts, they have been called "homogenizers of biodiversity" (Cristescu 2015). Invasive species have featured prominently in four of the past seven horizon scans of global conservation issues published annually in the journal Trends in Ecology and Evolution, indicating that invasive species are presently a key element of conservation science (Sutherland *et al.* 2009; 2013; 2014; 2015; 2016). Just as they are ecologically damaging, species invasions are also economically harmful. In the US alone, invasive species have been estimated to cost $120 billion dollars annually in damages and control measures (Pimentel *et al.* 2005).

In addition to presenting a serious threat to biodiversity and economies, invasions represent dynamic ecological and evolutionary processes that break expectations of population equilibrium and can lend insight into the nature of spatial population processes (Geller *et al.* 2010; Bock *et al.* 2015; Brandvain & Wright 2016). Many invasions present an evolutionary paradox because the invading species is ecologically successful despite a high probability of experiencing reduced genetic diversity due to an initial founder event during introduction, often thought to decrease fitness and adaptive potential (*e.g.,* Tsutsui *et al.* 2000). The initial introduction of a species is one of four steps characterizing biological invasions: (1) transport, (2) introduction, (3) establishment, and (4) spread (Blackburn *et al.* 2011). Post-establishment spread is often characterized by range expansion, further highlighting the utility of using invasive species to study range expansion dynamics.

The lionfish, Pterois volitans

The invasion of the Indo-Pacific lionfish, *Pterois volitans,* into waters off the US Atlantic Coast, Gulf of Mexico, and Caribbean Sea is occurring at an unprecedented rate with unparalleled collateral ecological damage. The rate and extent of the invasion make it an ideal model for research focused on the genomic signatures of rapid range expansion and the drivers of marine invasions. In their native range, lionfish populations appear to be well controlled by predators and competitors (Kulbicki *et al.* 2012), but in their invaded range, lionfish are prolific breeders, insatiable predators, and habitat generalists (Morris & Akins 2009). First reported off Dania, Florida in 1985, the lionfish invasion is thought to have originated from a single introduction followed by a long incubation period and an immense post-establishment expansion (Betancur-R *et al.* 2011). In the late 1990s and early 2000s, lionfish began their northward expansion, and by 2004 sightings of juveniles were reported as far north as Cape Cod, although no known breeding populations have been established north of North Carolina to date. In 2004, lionfish spread to the Bahamas, and in the years since have been reported south through the Caribbean Sea to Brazil, southeast to the coast of South America, north through Panama, Belize, Mexico, and ultimately back into the Gulf of Mexico (Schofield 2010).

Traditionally viewed as a tropical reef fish, lionfish have been observed at surprising depths over 300 meters (Morris 2009). Lionfish have also been reported in estuarine river habitats (salinity 5.8 to 38.6 ‰) up to 5.5 km from the ocean in the Loxahatchee River in Florida (Jud *et al.* 2011), pointing to an immense capacity for either rapid evolutionary change or highly plastic physiological tolerance. Rapid evolutionary change on this time scale was until recently thought impossible; however, recent research using genomic methods like those used in this dissertation indicate a potential capacity for rapid evolutionary shifts over decadal time scales, seen specifically in freshwater evolution of the Stickleback fish on Middleton Island, Alaska (Lescak *et al.* 2015). Despite the remarkable nature of the lionfish invasion and the broad-reaching implications of rapid genetic change during a range expansion, there are notably few published population genetic studies on lionfish (reviewed in Chapter 4).

The Asian shrimp, Palaemon macrodactylus

Crustaceans are among the most common and successful marine invaders. The shrimp *Palaemon macrodactylus* Rathbun 1902, is native to Japan, Korea and China, and has invaded a wide range of biogeographic provinces worldwide (Ashelby *et al.* 2013). In the Northeastern United States, *P. macrodactylus* invaded the Bronx River in 2001, and as of 2014 has spread north to New Hampshire and likely south to the Chesapeake Bay (Fofonoff *et al.* 2003, accessed 2016). The exact expansion pathway was previously unknown, however, leaving open questions about the dynamics of the apparent post-establishment spread—questions that are addressed in Chapter 3.

From genetics to genomics and the development of new analytical tools

Since the advent of Next Generation Sequencing in the 1990s and now Third Generation Sequencing, fields reliant on DNA sequence data, like molecular evolution and population genetics, have undergone a revolution (for a review, see Heather & Chain (2016). The sequencing of millions of nucleotide bases in a single run is now possible (Heather & Chain 2016). Coincident with the advent of these sequencing technologies has been a computing revolution, including increased reliance on cloud computing (Stein 2010). Storing and processing terabytes of data is possible now at a significantly lower cost than it was even five years ago.

The existence and accessibility of sequencing and computing tools has spurred innovation in the methods used to generate genomic information and has led to new developments like restriction enzyme associated DNA sequencing (RAD-seq) which is used throughout this dissertation (Miller *et al.* 2007; Baird *et al.* 2008). The use of next generation sequencing and other emerging genomic tools to address long-standing questions in invasion biology is widely recognized as the frontier in invasion genetics research—promising a synergy between previously intractable questions and burgeoning technologies (Chown *et al.* 2014; Rius *et al.* 2015). While new analyses are now possible—as seen in Chapters 3, 4, and 5—that are facilitated by these new technologies, there is still great utility in Sanger sequencing, particularly of mitochondrial genes, as evidenced by the use of mitochondrial data for barcoding, broad-scale connectivity patterns, and tracking multiple introductions of invasive species in Chapters 2, 3, and 4.

III. Dissertation Overview

The overall goal of this dissertation was to use genetic and genomic techniques to better understand the processes of invasion, range expansion, and connectivity in marine populations. To that end, the dissertation includes four chapters focused on three distinct ecological systems in three regions of the world: the New Zealand deep sea, the coastal estuaries of the US Atlantic coast, and the coral reefs of the Caribbean Sea and western Atlantic. Focused on broad questions rather than specific species, the research uses polychaete worms, galatheid crabs, glass shrimp, and venomous predatory reef fish to address fundamental questions regarding the dynamics of marine populations and the evolution of marine species.

Chapter 2. Patterns of Deep-Sea Genetic Connectivity in the New Zealand Region: Implications for Management of Benthic Ecosystems

In Chapter 2, I describe the patterns of genetic connectivity among populations of benthic invertebrates found at three different deep-sea regions in New Zealand's exclusive economic zone—a prominent rise, an adjacent slope margin, and a nearby plateau—in order to evaluate the placement of benthic protection areas. The work focuses on two species—the squat lobster *Munida gracilis* (Henderson, 1885) and the onuphid, or "quill," worm *Hyalinoecia longibranchiata* (McIntosh, 1885). Both species are abundant and widely distributed in the New Zealand region, existing throughout the study area (Read & Clark 1999), but have markedly different life history characteristics, including reproduction and dispersal behaviors.

I address the following questions regarding population genetic connectivity in the New Zealand region with a broader perspective of benthic connectivity in general: (1) Is there regional genetic structure across the study area and if so, can this structure be explained by factors known to affect genetic connectivity (*e.g.*, currents, geographic distribution, topography, habitat availability)? (2) Is there significant genetic structure within the three regions? For example, populations that are located in different habitats but are geographically close together could be genetically different. Is there significant genetic structure between populations on the north and south flanks of the Chatham Rise, potentially influenced by the presence of the Subtropical Front? (3) Do the inferred life history strategies correlate with the observed patterns of genetic connectivity? (4) What implications do the patterns of genetic population connectivity between species and among sample sites, habitats, and regions have for MPA design and the efficacy of the current Benthic Protection Areas? These questions are addressed through population genetic analyses of mitochondrial sequence data. Chapter 2 represents a step towards understanding the spatial structure of benthic communities in New Zealand waters, and informing the future design of deep-water MPAs in the region.

Chapter 3. Multiple, Spatially Distinct Introductions and Not Range Expansion May Explain Colonization History in an Invasive Marine Species

Chapter 3 leaves the deep sea and turns to the coastal estuaries of New England. I present a combination of mitochondrial gene population genetics and new non-model species genomics for the invasive shrimp, *Palaemon macrodactylus*. I use sequence data from mitochondrial *cytochrome oxidase I* (*COI*) and data from 1,598 single nucleotide polymorphisms (SNPs) generated from restriction-enzyme-associated DNA sequencing (RAD-seq) to investigate population genetic patterns in the invaded region. Comparing mitochondrial DNA sequence data from recently collected samples to sequences generated from shrimp collected for previously published work lends unprecedented insight into this invasion.

Chapter 3 highlights the utility of population genetics for revealing invasion pathways and uncovering unexpected patterns of expansion. In the absence of systematic surveys of palaemonid shrimp along the U.S. Atlantic coast around the time of the first reported observation of the species in 2001, the precise location and timing of introduction in North America is not known. At least two colonization scenarios are possible. The first scenario is a progressive range expansion up the US Atlantic coast. The second scenario involves multiple introductions driving an increase in the species range. In Chapter 3, I examine which of these two scenarios best explains the nature of the establishment and distribution of *P. macrodactylus*. In testing these two possible scenarios, RAD-sequencing is used to describe the distribution of genetic diversity in the invaded area between New York and New Hampshire in the context of range expansion expectations.

Chapter 4. Non-Equilibrium Population Genomics of the Rapidly Invading Lionfish, Pterois volitans*, Reveals Expansion Signals Without Spatial Metapopulation Structure*

In Chapter 4, I again shift geographic focus, this time to the reefs of the Caribbean, where lionfish have invaded with tremendous speed and are causing unprecedented ecological and economic damage (Hixon *et al.* 2016). Chapter 4 contains the first genome-wide single nucleotide polymorphism (SNP) data for the invasive lionfish throughout the Caribbean Sea using 12,759 loci across nine populations. These SNP data are analyzed from a range expansion perspective, identifying changes in genetic diversity with distance from the point of invasion.

Chapter 4 builds on the analyses from previous chapters of the dissertation and contains the most comprehensive analysis of RAD-seq data in this dissertation. The chapter includes genomic outlier analyses coupled with BLAST analysis to identify putative gene regions in the lionfish genome that may be experiencing selection or strong genetic drift during the invasion. Analyses in Chapter 4 are executed through the use of custom scripts to sort RAD loci in order to identify loci that could be exhibiting unique patterns relative to the rest of the genome that might indicate range-expansion-related signatures in the genome.

Chapter 5. Temporal Population Genomic Patterns Illuminate Ongoing Processes of Range Expansion in the Invasive Lionfish, Pterois volitans

In Chapter 5, I begin to illuminate a central component of range expansion biology: temporal changes to the genomic signatures of expansion. Building on the RAD-sequencing results presented in Chapter 5, I analyze data from 1,054 SNPs throughout the lionfish genome and describe differences in the genetic diversity patterns at two different time points in the invasion (2007-2009 and 2013-2014). The hypotheses tested stipulate differences in the invasion along the east coast of the United States and the invasion into the Caribbean as well as predict reductions in range expansion signals over time in the Caribbean.

LITERATURE CITED

Ashelby CW, Johnson ML, De Grave S (2013) The global invader *Palaemon macrodactylus* (Decapoda, Palaemonidae): an interrogation of records and a synthesis of data. *Crustaceana*, **86**, 594–624.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers (JC Fay, Ed,). *PLoS ONE*, **3**, e3376.

Barange M, Field JG, W S (2010) Marine Ecosystems and Global Change.

Bax N, Williamson A, Aguero M, Gonzalez E, Geeves W (2003) Marine invasive alien species: a threat to global biodiversity. *Marine Policy*, **27**, 313–323.

Betancur-R R, Hines A, Acero P A *et al.* (2011) Reconstructing the lionfish invasion: insights into Greater Caribbean biogeography. *Journal of Biogeography*, **38**, 1281–1293.

Blackburn TM, Pyšek P, Bacher S *et al.* (2011) A proposed unified framework for biological invasions. *Trends in Ecology & Evolution*, **26**, 333–339.

Bock DG, Caseys C, Cousens RD *et al.* (2015) What we still don't know about invasion genetics. *Molecular Ecology*, **24**, 2277–2297.

Booth DJ, Bond N, Macreadie P (2011) Detecting range shifts among Australian fishes in response to climate change. *Marine and Freshwater Research*, **62**, 1027–1042.

Brandvain Y, Wright SI (2016) The limits of natural selection in a nonequilibrium world. *Trends in Genetics*, **32**, 201–210.

Chown SL, Hodgins KA, Griffin PC *et al.* (2014) Biological invasions, climate change and genomics. *Evolutionary Applications*, **8**, 23–46.

Cowen RK, Gawarkiewic G, Pineda J, Thorrold SR, Werner FE (2007) Population connectivity in marine systems: an overview. *Oceanography*, **20**, 14–21.

Cristescu ME (2015) Genetic reconstructions of invasion history. *Molecular Ecology*, **24**, 2212–2225.

Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences*, **101**, 975–979.

Excoffier L, Foll M, Petit RJ (2009) Genetic Consequences of Range Expansions. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 481–501.

Fofonoff PW, Ruiz GM, Steves B, Carlton JT. (2016) *Palaemon macrodactylus,* in California Non-native Estuarine and Marine Organisms (Cal-NEMO) System. http://invasions.si.edu/nemesis/. Accessed December 13, 2016.

Gaines SD, Lester SE, Grorud-Colvert K, Costello C, Pollnac R (2010a) Evolving science of marine reserves: New developments and emerging research frontiers. *Proceedings of the National Academy of Sciences*, **107**, 18251–18255.

Gaines SD, White C, Carr MH, Palumbi SR (2010b) Designing marine reserve networks for both conservation and fisheries management. *Proceedings of the National Academy of Sciences*, **107**, 18286–18293.

Geller JB, Darling JA, Carlton JT (2010) Genetic perspectives on marine biological invasions. *Annual Review of Marine Science*, **2**, 367–393.

Hallatschek O, Nelson DR (2008) Gene surfing in expanding populations. *Theoretical Population Biology*, **73**, 158–170.

Hallatschek O, Hersen P, Ramanathan S, Nelson DR (2007) Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences*, **104**, 19926–19930.

Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics*, **107**, 1–8.

Hedgecock D, Barber PH (2007) Genetic approaches to measuring connectivity. *Oceanography*, **20**.

Hixon MA, Green SJ, Albins MA, Akins JL, Morris JA Jr (2016) Lionfish: a major marine invasion. *Marine Ecology Progress Series*, **558**, 161–165.

Jones GP, Planes S, Thorrold SR (2005) Coral reef fish larvae settle close to home. *Current Biology*, **15**, 1314–1318.

Jones SJ, Southward AJ (2012) Climate change and historical biogeography of the barnacle Semibalanus balanoides. *Global ecology and Biogeography*.

Jud ZR, Layman CA, Lee JA, Arrington DA (2011) Recent invasion of a Florida (USA) estuarine system by lionfish *Pterois volitans* / *P. miles*. *Aquatic Biology*, **13**, 21–26.

Klopfstein S (2005) The Fate of Mutations Surfing on the Wave of a Range Expansion. *Molecular Biology and Evolution*, **23**, 482–490.

Kulbicki M, Beets J, Chabanet P *et al.* (2012) Distributions of Indo-Pacific lionfishes Pterois spp. in their native ranges: implications for the Atlantic invasion. *Marine Ecology Progress Series*, **446**, 189–205.

Lescak EA, Bassham SL, Catchen J *et al.* (2015) Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proceedings of the National Academy of Sciences*, **112**, E7204–E7212.

Lowry E, Rollinson EJ, Laybourn AJ *et al.* (2013) Biological invasions: a field synopsis, systematic review, and database of the literature. *Ecology and Evolution*, **3**, 182–196.

Mainka SA, Howard GW (2010) Climate change and invasive species: double jeopardy. *Integrative Zoology*, **5**, 102–111.

Miles EL (2009) On the increasing vulnerability of the world ocean to multiple stresses. *Annual Review of Environment and Resources*, **34**, 17–41.

Miller KJ, Ayre DJ (2008) Protection of genetic diversity and maintenance of connectivity among reef corals within marine protected areas. *Conservation Biology*, **22**, 1245–1254.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Morris J (2009) The biology and ecology of the invasive Indo-Pacific lionfish. Doctoral Dissertation.

Morris JA Jr, Akins JL (2009) Feeding ecology of invasive lionfish (Pterois volitans) in the Bahamian archipelago. *Environmental Biology of Fishes*, **86**, 389–398.

Möbius W, Murray AW, Nelson DR (2015) How obstacles perturb population fronts and alter their genetic structure (L Excoffier, Ed,). *PLoS Computation Biology*, **11**, e1004615–30.

Palumbi SR (2003) Population genetics, demographic connectivity, and the design of marine reserves. *Ecological Applications*.

Parmesan C (2006) Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 637–669.

Parmesan C, Yohe G (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, **421**, 37–42.

Peischl S, Excoffier L (2015) Expansion load: recessive mutations and the role of standing genetic variation. *Molecular Ecology*, **24**, 2084–2094.

Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L (2013) On the accumulation of deleterious mutations during range expansions. *Molecular Ecology*, **22**, 5972–5982.

Perry AL (2005) Climate change and distribution shifts in marine fishes. *Science*, **308**, 1912–1915.

Pimentel D, Zuniga R, Morrison D (2005) Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics*.

Poloczanska ES, Brown CJ, Sydeman WJ *et al.* (2013) Global imprint of climate change on marine life. *Nature Climate Change*, **3**, 1–7.

Ramirez-Llodra E, Tyler PA, Baker MC *et al.* (2011) Man and the last great wilderness: human impact on the deep sea (P Roopnarine, Ed,). *PLoS ONE*, **6**, e22588–25.

Read GB, Clark H (1999) Ingestion of quill-worms by the astropectinid sea-star Proserpinaster neozelanicus (Mortensen). *New Zealand Journal of Zoology*, **26**, 49–54.

Rius M, Bourne S, Hornsby HG, Chapman MA (2015) Applications of next-generation sequencing to the study of biological invasions. *Current Zoology*, **61**, 488–504.

Sabatés A, Martín P, Llorer J, Raya V (2006) Sea warming and fish distribution: the case of the small pelagic fish, Sardinella aurita, in the western Mediterranean. *Global Change Biology*, **12**, 2209–2219.

Schofield P (2010) Update on geographic spread of invasive lionfishes (Pterois volitans [Linnaeus, 1758] and P. miles [Bennett, 1828]) in the Western North Atlantic Ocean, Caribbean Sea and Gulf of Mexico. *Aquatic Invasions*, **5**, S117–S122.

Selkoe KA, D'Aloia CC, Crandall ED *et al.* (2016) A decade of seascape genetics: contributions to basic and applied marine connectivity. *Marine Ecology Progress Series*, **554**, 1–19.

Shank TM (2010) Seamounts: deep-ocean laboratories of faunal connectivity, evolution, and endemism. *Oceanography*, **23**, 108–122.

Sorte CJ, Williams SL, Carlton JT (2010) Marine range shifts and species introductions: comparative spread rates and community impacts. *Global Ecology and Biogeography*, **19**, 303–316.

Stein LD (2010) The case for cloud computing in genome informatics. *Genome biology*, **11**, 207.

Sunday JM, Bates AE, Dulvy NK (2012) Thermal tolerance and the global redistribution of animals. **2**, 686–690.

Sutherland WJ, Aveling R, Brooks TM *et al.* (2014) A horizon scan of global conservation issues for 2014. *Trends in Ecology & Evolution*, **29**, 15–22.

Sutherland WJ, Bardsley S, Clout M *et al.* (2013) A horizon scan of global conservation issues for 2013. *Trends in Ecology & Evolution*, **28**, 16–22.

Sutherland WJ, Broad S, Caine J *et al.* (2016) A Horizon Scan of Global Conservation Issues for 2016. *Trends in Ecology & Evolution*, **31**, 44–53.

Sutherland WJ, Clout M, Côté IM *et al.* (2009) A horizon scan of global conservation issues for 2010. *Trends in Ecology & Evolution*, **24,** 1–7.

Sutherland WJ, Clout M, Depledge M *et al.* (2015) A horizon scan of global conservation issues for 2015. *Trends in Ecology & Evolution*, **30**, 17–24.

Thomas CD, Cameron A, Green RE *et al.* (2004) Extinction risk from climate change. *Nature*, **427**, 145–148.

Thomaz SM, Kovalenko KE, Havel JE, Kats LB (2014) Aquatic invasive species: general trends in the literature and introduction to the special issue. *Hydrobiologia*, **746**, 1–12.

Travis JM, Munkemuller T, Burton OJ *et al.* (2007) Deleterious mutations can surf to high densities on the wave front of an expanding Population. *Molecular Biology and Evolution*, **24**, 2334–2343.

Tsutsui ND, Suarez AV, Holway DA, Case TJ (2000) Reduced genetic variation and the success of an invasive species. *Proceedings of the National Academy of Sciences*, **97**, 5948–5953.

Van Dover CL, Smith CR, Ardron J *et al.* (2012) Designating networks of chemosynthetic ecosystem reserves in the deep sea. *Marine Policy*, **36**, 378–381.

White C, Selkoe KA, Watson J *et al.* (2010) Ocean currents help explain population genetic structure. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 1685–1694.

CHAPTER 2

Patterns of Deep-Sea Genetic Connectivity in the New Zealand Region: Implications for Management of Benthic Ecosystems

This research was originally published on November 21, 2012 in the journal PLOS ONE. It is available online: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049474

Bors EK, Rowden AA, Maas EW, Clark MR, Shank TM (2012) Patterns of Deep-Sea Genetic Connectivity in the New Zealand Region: Implications for Management of Benthic Ecosystems. PLoS ONE 7(11): e49474. doi:10.1371/journal.pone.0049474

ABSTRACT

Patterns of genetic connectivity are increasingly considered in the design of marine protected areas (MPAs) in both shallow and deep water. In the New Zealand Exclusive Economic Zone (EEZ), deep-sea communities at upper bathyal depths (<2000 m) are vulnerable to anthropogenic disturbance from fishing and potential mining operations. Currently, patterns of genetic connectivity among deep-sea populations throughout New Zealand's EEZ are not well understood. Using the mitochondrial *Cytochrome Oxidase I* and *16S rRNA* genes as genetic markers, this study aimed to elucidate patterns of genetic connectivity among populations of two common benthic invertebrates with contrasting life history strategies. Populations of the squat lobster *Munida gracilis* and the polychaete *Hyalinoecia longibranchiata* were sampled from continental slope, seamount, and offshore rise habitats on the Chatham Rise, Hikurangi Margin, and Challenger Plateau. For the polychaete, significant population structure was detected among distinct populations on the Chatham Rise, the Hikurangi Margin, and the Challenger Plateau. Significant genetic differences existed between slope and seamount populations on the Hikurangi Margin, as did evidence of population differentiation between the northeast and southwest parts of the Chatham Rise. In contrast, no significant population structure was detected across the study area for the squat lobster. Patterns of genetic connectivity in *Hyalinoecia longibranchiata* are likely influenced by a number of factors including current regimes that operate on varying spatial and temporal scales to produce potential barriers to dispersal. The striking difference in population structure between species can be attributed to differences in life history strategies. The results of this study are discussed in the context of existing conservation areas that are intended to manage anthropogenic threats to deep-sea benthic communities in the New Zealand region.

INTRODUCTION

Current anthropogenic pressures on the marine environment, including the deep sea, are unprecedented (Miles 2009; Barange *et al.* 2010; Ramirez-Llodra *et al.* 2011; Van Dover *et al.* 2012). As the human footprint in the oceans increases, international agreements like the UN Convention on Biodiversity have spurred the creation of several national biodiversity task forces that acknowledge the importance of marine protected areas (MPAs) — *i.e.*, any area of the marine environment that has been reserved by laws or regulations to provide lasting protection to part or all of the natural or cultural resources therein (*e.g.,* New Zealand Biodiversity Strategy, 2000; Department for Environment, Food, and Rural Affairs, 2011). However, the creation of such areas is often a political, social, and scientific challenge (Gaines *et al.* 2010a).

Genetic connectivity has recently come to the fore as a major scientific component of sound MPA design in both shallow and deep-sea environments (Palumbi 2003; Miller & Ayre 2008; Shank 2010; Gaines *et al.* 2010b). Genetic connectivity, or "the dispersal, survival, and reproduction of migrants, so that they contribute to the local gene pool" (Hedgecock & Barber 2007), examines temporal and spatial aspects of population genetics in order to infer the degree of genetic exchange among populations. The theoretical optimization of MPA design arises from understanding the sources and sinks of marine populations so that MPAs can protect sites that will export individuals to other areas, thus increasing the net benefit of the MPA (Gaines *et al.* 2010b). Genetic connectivity research generally focuses on patterns of population structure within a geographic area and the factors that could cause such population structure to arise.

Ranking as the sixth largest globally, the New Zealand Exclusive Economic Zone (EEZ) is one of the most topographically diverse seafloor environments in the world (Ramillien & Wright 2000). Benthic habitats are provided by a continental slope with canyons and cold seeps, while further off shore there are numerous plateau, rises, troughs, ridges, basins, seamounts (many with hydrothermal vents), as well as two ocean trenches (Thompson, 1991). The New Zealand EEZ supports rich biodiversity (Gordon *et al.* 2010), economically important and well-established fisheries (Gibbs, 2008), and provides for other extractive industries, including hydrocarbon and mineral mining (Crown Minerals, 2010; Glasby and Wright, 1990). Of the many species commercially targeted by New Zealand's fisheries, just ten deep-water species comprise 70% of the total catch volume (New Zealand Ministry of Fisheries, 2010), and bottom trawling occurs at depths down to 1500 m throughout the EEZ (*e.g.,* Baird *et al.*, 2011). The physical disturbance from trawl gear can have profound effects on deep-sea benthic communities, particularly on seamounts (Clark & Rowden 2009), where communities are thought to be more susceptible to disturbance from trawling because the fauna are less adapted to frequent natural disturbances and have life history traits that make them particularly vulnerable to fishing (Probert 1999; Clark *et al.* 2010). Bottom trawling on non-seamount habitat is extensive in the New Zealand EEZ, with areas of the seabed on the Chatham Rise having been subjected to tens of thousands of trawls between 1989-2005 (*e.g.,* fishing statistical area, Figure 17 from Baird *et al.,* 2011). While the impact of fishing on benthic communities at non-seamount habitats is generally unknown in the New Zealand EEZ, invertebrate by-catch studies have indicated a likely disturbance to soft sediment communities on the continental margin slope and certain areas of the Chatham Rise (Probert *et al.* 1997; Cryer *et al.* 2002). In addition to fishing operations, interest in mining has increased. Seafloor areas of the Chatham Rise contain significant deposits of phosphorite nodules (Glasby and Wright, 1990) and several companies have been granted exploratory permits (New Zealand Petroleum and Minerals, 2012).

Currently, there is no legislation that allows for the creation of marine reserves (defined by current New Zealand law as MPAs in which only scientific uses are allowed) in the New Zealand EEZ (*i.e.,* outside of the 12 nautical mile territorial seas), limiting the tools available for management of human activities in New Zealand's deep sea. There are areas closed to bottom trawling that include specific seamounts (Brodie and Clark, 2003) and fishing industry-created Benthic Protection Areas (BPAs) (Helson *et al.* 2010). But, activities such as mid-water trawling and mining are allowed at closed seamounts and in BPAs, a fact which has raised the concern that this specific type of closure does not fulfill biodiversity goals for New Zealand's EEZ. To date, only one published study has addressed the placement of the BPAs (Leathwick *et al.* 2008), despite their imminent 2013 review.

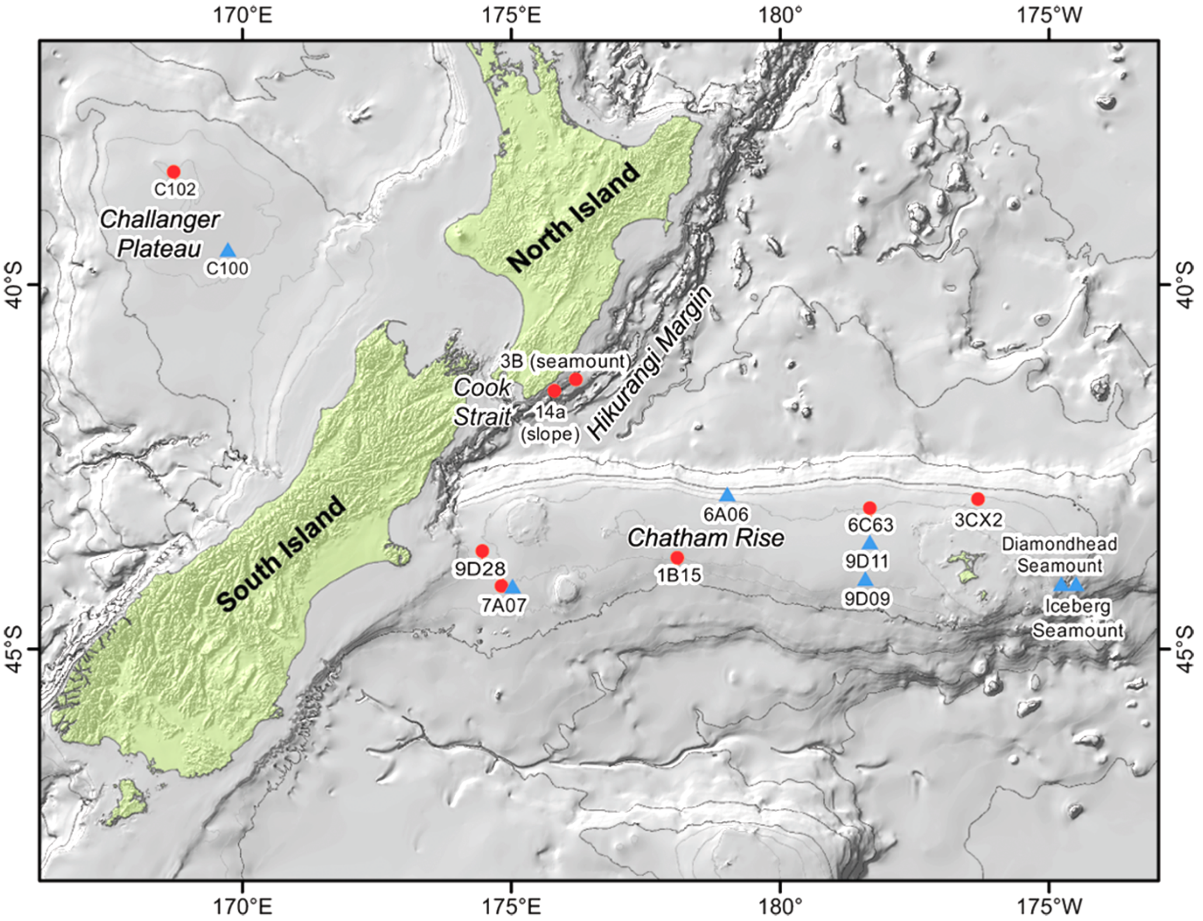
Most population genetic studies in the New Zealand region have been carried out in coastal waters (Miller 1997; Apte & Gardner 2002; Perrin *et al.* 2004; Ross *et al.* 2011), with relatively few studies of deep-water species. Smith *et al*. (2004) examined connectivity of hydrothermal vent mussels between two seamounts in the Kermadec Arc, north of New Zealand. Allozyme loci revealed unexpected levels of heterogeneity between the seamount populations despite only 50 km of separation. The authors attribute the finding to localized current regimes promoting isolation of these populations. Using mitochondrial *Cytochrome Oxidase I* (*COI*) data, Kojima *et al.* (2006) demonstrated that the population of *Lamellibrachia juni* tubeworms at Brothers seamount in the Kermadec Volcanic Arc contains two distinct genetic groups, one of which was phylogenetically related to samples from the TOTO caldera in the Mariana Volcanic Arc. These studies show how complex patterns can exist over various spatial scales. Similarly, genetic investigation of the *Internal Transcribed Spacer* regions 1 and 2, *COI* and *16S rRNA* genes in populations of the coral *Desmophyllum dianthus* from Chile, New Zealand, and Australia revealed greater variation between populations at different depths within a region than between populations at the same depth in two different regions (Miller *et al.* 2011). Corals in the New Zealand mid-depth stratum were more similar to corals in a mid-depth stratum in Australia than to corals in shallower water in New Zealand, and geographic genetic structure was not observed within the New Zealand region by this study (Miller *et al.* 2011). A study of Keratoisidinae bamboo corals in the Western Pacific using the INDEL#2 region of *16S rRNA* and a non-coding mitochondrial marker also found no genetic structure in the New Zealand region—which may present an accurate evolutionary pattern or may be the result of using evolutionarily conserved genetic markers that can be slow to change over time (Smith *et al.* 2004).

The present study aims to elucidate patterns of genetic connectivity among populations of benthic invertebrates found at three different deep-sea regions—a prominent rise, an adjacent slope margin, and a nearby plateau—and to consider the implications of the observed patterns for management decisions. The three study regions were the Chatham Rise, the Hikurangi Margin, and the Challenger Plateau (Figure 1). The Chatham Rise is a submerged feature that extends about 800 kilometers to the east of the South Island of New Zealand. There are numerous seamounts on the Rise, including the Graveyard Seamount cluster on the northern flank and the Andes Seamount cluster on the southeastern edge of the rise (Mackay *et al.,* 2005). The Subtropical Front (STF), a convergence zone between the subtropical and subantarctic water masses, extends west to east along the rise at the confluence of the East Cape Current and the Southland Current (Heath 1985). To the northwest of the Chatham Rise is Cook Strait, which separates the North and South Islands of New Zealand. The Hikurangi Margin is at the eastern opening of Cook Strait. Small seamounts are found across the slope of the Margin, which is also incised with numerous canyons. The Challenger Plateau extends off the continental margin to the west side of the Cook Strait.

We focus on two benthic invertebrates—the squat lobster *Munida gracilis* (Henderson, 1885) and the onuphid, or "quill," worm *Hyalinoecia longibranchiata* (McIntosh, 1885). Both species are abundant and widely distributed in the New Zealand region, existing throughout our study area (Read & Clark 1999). These species have strongly contrasting inferred modes of reproduction and dispersal. There is likely a long pelagic larval duration via planktonic dispersal in *M. gracilis*, as is typical for many *Munida* species (Baba *et al.,* 2011). Development is non-planktotrophic in onuphids (Paxton, 1986), with some *Hyalinoecia* species having incubated embryos (Carrasco, 1983; Orensanz, 1990). These species are used here to represent commonly occurring benthic organisms with contrasting life history strategies.

Our study of the population genetics of these two species aimed to address fundamental questions regarding connectivity of the deep benthos among some of the prominent geomorphic features in the New Zealand EEZ: (1) Is there regional genetic structure across the study area and if so, can this structure be explained by factors known to affect genetic connectivity (e.g., currents, geographic distribution, topography, habitat availability)?; (2) Is there significant genetic structure within the three regions? For example, is there a difference among populations that are found in different habitats but are geographically close together? Is there significant genetic structure between populations on the north and south flanks of the Chatham Rise, potentially influenced by the presence of the Subtropical Front?; (3) Do the inferred life history strategies correlate with the observed patterns of genetic connectivity?; and (4) What implications do the patterns of genetic population connectivity between species and among sample sites, habitats, and regions have for Marine Protected Area design and the efficacy of the current Benthic Protection Areas?

Figure 1. The location of the study area, including the North and South Islands of New Zealand (landmasses are in green), the Challenger Plateau, Hikurangi Margin, and Chatham Rise. Red circles mark sites from which *Munida gracilis* were collected; blue triangles mark sites from which *Hyalinoecia longibranchiata* were collected. Sites are labeled with their original site names. Samples were selected within a depth band of 400-800m with *Munida gracilis* between 421m and 634m, and *Hyalinoecia longibranchiata* between 478m to 746m. The depth of each site is listed in Table 1.



Methods

Sample Collection and Study Sites

Populations of *Hyalinoecia longibranchiata* and *Munida gracilis* were collected during four research cruises onboard the R/V *Tangaroa*: TAN0705 (Chatham Rise, March 31st to April 29th 2007), TAN0707 (Challenger Plateau, May 28th to June 8th 2007), TAN0905 (Andes and Graveyard Seamounts, June 12th to June 30th 2009), and TAN1004 (Hikurangi Margin including the slope and seamounts near the eastern side of the Cook Strait, April 14th to April 29th 2010). All necessary permits were obtained for the described field studies. The specimens used in this study were taken from samples collected or obtained by New Zealand's National Institute of Water and Atmospheric Research (NIWA) under a "Special Permit (421)" issued by the New

Zealand Ministry of Fisheries for the taking of fish, aquatic life, and seaweed for the purposes of education and investigative research. Samples were collected using NIWA's epibenthic "seamount sled" (overall size 150 cm long, 50 cm high, and 100 cm wide; macro-invertebrates retained in a 30 mm stretched mesh size net that was covered in an anti-chaffing net of 100 mm stretched mesh size), a hyperbenthic "Brenke" sled, and a beam trawl. Upon collection, *M. gracilis* and *H. longibranchiata* specimens were preserved in ethanol, except for 29 *H. longibranchiata* individuals that were frozen upon collection. All specimens are stored in the NIWA Invertebrate Collection (NIC).

Due to limitations in the number of sampled individuals and resources available to this study, not all sample sites for which there were specimens in the NIC could be used in this study. In order to avoid confounding geographic site comparisons with depth variability, populations (the combined individual samples) from sites were selected from a restricted depth range (400 - 800 m). Sites did not straddle a previously identified depth disjunction in population structure between populations at <600 and >1000 m (Miller *et al.* 2011). Individuals of *H. longibranchiata* were sampled from 478m to 746m and *M. gracilis* from 421m to 634m depth. To explore the role of geomorphological habitat types as a factor for structuring connectivity, we identified sites that spanned habitat types in the three regions (*e.g.,* seamount and slope). When possible, sites with samples for both species were used.

The study sites (*i.e.,* sampled populations) are presented in Table 1 and Figure 1. One population for each species was identified on the Challenger Plateau: "C102" for *H. longibranchiata* and "C100" for *M. gracilis*. Two sites from the Hikurangi Margin were used for *H. longibranchiata*: "14a," a slope habitat site and "3B," a seamount site. On the Chatham Rise, five sites were used for *H. longibranchiata*: "7A07," "9D28," and "1B15" on the southwest part of the rise, and "6C63" and "3CX2" on the northeast part of the rise. Six sites on the Chatham Rise were identified for *M. gracilis*: "7A07" in the southwest, "6A06" centrally located on the northern flank of the rise, "9D11" and "9D09" located in the south-central region of the rise, and "Iceberg Seamount" and "Diamondhead Seamount" of the Andes Seamount cluster at the eastern end of the rise.

Table 1. Sites and collected samples included in this study.

Site Name	Cruise	Location	Latitude	Longitude	Date	Depth	Species
9D11	TAN0705	Chatham Rise	43.6287 S	178.3664 W	18-Apr-07	421	MG
9D09	TAN0705	Chatham Rise	44.0682 S	178.3295 W	18-Apr-07	450	MG
9D28	TAN0705	Chatham Rise	43.7257 S	174.458 E	27-Apr-07	550	HL
6A06	TAN0705	Chatham Rise	42.9935 S	178.9992 E	24-Apr-07	530	MG
7A07	TAN0705	Chatham Rise	44.1358 S	174.8438 E	4-Apr-07	518	HL, MG
3CX2	TAN0705	Chatham Rise	42.9988 S	176.3483 W	16-Apr-07	658	HL
1B15	TAN0705	Chatham Rise	43.8085 S	178.1173 E	7-Apr-07	497	HL
6C63	TAN0705	Chatham Rise	43.1575 S	178.3097 W	17-Apr-07	478	HL
C102	TAN0707	Challenger Plateau	38.3872 S	168.7397 E	29-May-07	482	HL
C100	TAN0707	Challenger Plateau	39.5437 S	169.7145 E	4-Jun-07	634	MG
Iceberg Seamount	TAN0905	Andes Seamounts	44.1582 S	174.555 W	28-Jun-09	551	MG
Diamondhead Seamount	TAN0905	Andes Seamounts	44.1473 S	174.6900 W	26-Jun-09	520	MG
14a (slope)	TAN1004	Hikurangi Margin	41.5195 S	175.8068 E	19-Apr-10	746	HL
3B (seamount)	TAN1004	Hikurangi Margin	41.3368 S	176.182 E	21-Apr-10	730	HL

DNA extraction, Polymerase Chain Reaction, and Sequencing

Mid-section muscular tissue from *H. longibranchiata* and leg tissue from *M. gracilis* were sub-sampled for genomic DNA (gDNA) extraction. To increase gDNA yield, many ethanol-preserved samples (n=61) were soaked for 24 hours in a buffer containing 500mM Tris-HCL (pH8), 20mM EDTA, and 10mM NaCl before extraction (Nielsen, 2005). Genomic DNA was extracted using the QIAGEN DNeasy Blood and Tissue extraction kit following the manufacturer's instructions (Qiagen GmbH, Germany) with a final elution into 25 to 200 µl of RNAase/DNAse free H_2O (Invitrogen Ltd, New Zealand), depending on the condition of the original tissue sample. For samples with poor tissue quality due to disintegration in ethanol, elution occurred in smaller volumes of water in order to achieve a higher concentration of gDNA. Genomic DNA was quantified using Quant-IT PicoGreen DNA quantification kit according to the manufacturer's instructions (Invitrogren Ltd, New Zealand), and working stocks of DNA (approximately 10 ng/µl) were stored at 4°C for up to six months prior to use.

For both target species, a fragment of the mitochondrial *COI* gene was amplified using universal primers (Folmer *et al.* 1994). The *COI* gene was amplified using iProof High-Fidelity DNA Polymerase Master Mix (Bio-Rad Ltd, Australia), using 1-5 µl of gDNA and primer concentrations of 0.2 mM each. A subset of reactions was trialed with HOT FIREPol Master Mix with 1.5mM $MgCl_2$ (Solis BioDyne) in an unsuccessful attempt to increase PCR yield. A "touch-up" PCR profile was used to eliminate non-specific binding. The profile used for *COI* consisted of denaturing at 98°C for 2 minutes followed by 10 cycles of denaturing at 98°C for 10 seconds, annealing at 49°C incrementally raising to 54°C for 30 seconds, and extension at 72°C for 30 seconds; followed by twenty cycles of denaturing at 98°C for 10 seconds, annealing at 54°C for 30 seconds, and extension at 72°C for 30 seconds; with a final extension step of 72°C for 7 minutes in a 2700 Applied Biosystems PCR machine. For some samples with low gDNA concentrations, an extra ten cycles (for a total of 30 cycles) were added in the final PCR profile.

Primers used for *H. longibranchiata 16S* gene amplification were from Zanol *et al.* (2010). The *16S* gene was amplified using a "touch-down" method as described in Zanol *et al.* (2010). iProof High-Fidelity DNA Polymerase Master Mix (Bio-Rad Ltd, Australia) was used with 1-5 µl of gDNA and primer concentrations of 0.2 mM each. A portion of *16S* was sequenced for a small subset of *M. gracilis* samples (n=7); however, the portion of the genetic marker that we were able to sequence exhibited no variation among the sequenced individuals. The same was true for a small set of *Internal Transcribed Spacer Region* sequences (n = 12) generated for *H. longibranchiata*.

PCR amplification was assessed using gel electrophoresis and Quant-iT PicoGreen DNA quantification kit according to the manufacturer's instructions (Invitrogren Ltd, New Zealand). PCR products of the correct size were purified using either a Zymogenetics PCR purification kit (Zymogentics D4013) or a Qiagen PCR purification kit (Qiagen GmbH, Germany) and eluted in DNA/RNAase-free water. Purified PCR reactions of approximately 10 ng were shipped to Macrogen, Inc. for sequencing.

Genetic Analysis

DNA sequences were edited and aligned (using CLUSTAL-W) in Geneious Pro 5.3.4 [50]. Bi-directional sequences were used with the exception of one *H. longibranchiata* individual from 7A07 for which only one direction was usable for the*16S* sequence. Final datasets consisted of 680 basepairs of *16S* for *H. longibranchiata*, 524 basepairs of *COI* for *H. longibranchiata*, and 526 basepairs of *COI* for *M. gracilis*. DNA sequences have been deposited in GenBank (JX219896 - JX219956, *H. longibranchiata 16S*; JX219786 - JX219843, *H. longibranchiata COI*; and JX219844 - JX219895, *M. gracilis COI*). Sequences (and subsequent species identity) were compared to the Genbank database (http://www.ncbi.nlm.nih.gov/) and alignments of *COI* were translated into amino acid sequences and checked for stop codons to assess whether or not the amplified fragments could be considered pseudo-genes. The genetic distances among individuals in each dataset were compared to genetic distances among species within the same families and/or genera to correlate genetic divergence and morphological species boundaries (*i.e.*, to assess the possibility of cryptic species). Genetic distances were calculated (using Kimura 2 Parameter model in PAUP) among the *COI* sequences from seven *Munida* species: *Munida spilota*, *Munida stia*, *Munida notata*, *Munida tyche*, *Munida zebra*, *Munida taenia*, and *Munida thoe* (Machordom & Macpherson 2004). Genetic distances were calculated (using Kimura 2 Parameter model in PAUP) among the *COI* sequences from seven species of onuphid worms: *Diopatra* cf. *ornate*, *Diopatra dentate*, *Diopatra dentate*, *Hyalinoecia sp.*, *Onuphis elegans*, *Onuphis* cf. *iridescens*, and *Paradiopatra quadricuspis* (Zanol *et al.* 2010).

Genetic diversity indices (the number of polymorphic sites in the sequence, the number of haplotypes represented at each site, the haplotype diversity, and the nucleotide diversity) for each population and gene were calculated in DnaSP, version 5.10.01 (Librado & Rozas 2009). In order to test for the significance of population genetic divergences, a measure of population pairwise divergence, or F_{ST}, was calculated with 110 replicates in Arlequin, version 3.11 (Excoffier & Laval 2005). In addition to geographically mapping the distribution of haplotypes, we constructed haplotype networks in TCS 1.21 using default program settings (Clement *et al.* 2000) to identify potential biogeographic patterns among populations and habitats.

In order to assess population genetic structure among and within populations, Analysis of Molecular Variance (AMOVA) was conducted in Arlequin version 3.11 by grouping sites into the three regions: Chatham Rise, Hikurangi Margin, and Challenger Plateau and running a standard AMOVA with default program settings. To address the question of spatial variation along the Chatham Rise, an AMOVA was conducted to separate the sites into two groups: a northeast subset that consisted of 6C63 and 3CX2 and a southwest subset that consisted of 7A07, 9D28 and 1B15. AMOVA tests were run only on *H. longibranchiata* given that initial assessments indicated there was no genetic structure for *M. gracilis*.

RESULTS

Genetic Diversity Indices

Prior to any population genetic analyses, the assessment and confirmation of phylogenetic species was conducted across all sequences for each of the target taxa. The genetic distances among *COI* sequences for various *Munida* species (listed in the Methods section) ranged from 8.2% to 15.4%, while the genetic distances of our *COI* dataset for *Munida gracilis* ranged from 0% to 1.9%. The genetic distances among *COI* sequences for various onuphid worms ranged from 4.6% to 28.5%. Genetic divergences in the *COI* dataset for *Hyalinoecia longibranchiata* ranged from 0% to 2.9%. These results indicate that the individuals we used from each respective morphospecies are within the genetic divergence diagnostic of their respective species, implying that there are no cryptic species in our samples.

A total of 58 *COI* and 61 *16S* partial sequences from *H. longibranchiata* were obtained. There were eight *H. longibranchiata* specimens for which we did not obtain *COI* sequences, but did obtain *16S* sequences. There were four *H. longibranchiata* specimens for which we did not obtain *16S* sequences but we did obtain *COI* sequences. A total of 52 *COI* sequences from *M. gracilis* were obtained. All sequence reads were unambiguous except for one gene sequence for *H. longibranchiata*, which had a single undetermined base. The lack of complete overlap between sequenced individuals for each gene prevented joint analysis of the two genetic datasets. The per-site number of sequences and genetic diversity indices for each gene in both species is presented in Table 2, Table 3, and Table 4.

Table 2. Intra-population mt*COI* diversity statistics for the squat lobster, *Munida gracilis*. Regions are designated as CP for Challenger Plateau and CR for Chatham Rise. *n* is the total number of individuals sampled for a site, S is the number of polymorphic nucleotide sites in the sequence, h is the number of haplotypes represented at the site, Hd is haplotype diversity, and pi is nucleotide diversity.

Site (region)	n	S	h	Hd	π
C100 (CP)	6	6	4	0.8	0.00418
7A07 (CR)	4	11	4	1.00000	0.01109
6A06 (CR)	8	12	7	0.96429	0.00808
9D11 (CR)	8	13	8	1.00000	0.00727
9D09 (CR)	8	12	8	1.00000	0.00693
Iceberg Seamount (CR)	10	19	9	0.97778	0.00837
Diamondhead Seamount (CR)	8	6	7	0.96429	0.00395
Total	52	47	36	0.96003	0.00691

Table 3. Intra-population *16S* diversity statistics for the quill worm, *Hyalinoecia longibranchiata*. Regions are designated as CP for Challenger Plateau and CR for Chatham Rise. *n* is the total number of individuals sampled for a site, S is the number of polymorphic nucleotide sites in the sequence, h is the number of haplotypes represented at the site, Hd is haplotype diversity, and pi is nucleotide diversity.

Site (region)	**n**	**S**	**h**	**Hd**	**π**
C102 (CP)	12	6	7	0.90909	0.00294
3B (HM)	6	3	2	0.53333	0.00235
14a (HM)	10	4	3	0.51111	0.00187
9D28 (CR)	8	3	4	0.64286	0.00110
7A07 (CR)	8	5	5	0.85714	0.00210
1B15 (CR)	6	1	2	0.33333	0.00049
6C63 (CR	5	2	3	0.80000	0.00147
3XC2(CR)	6	1	2	0.33333	0.00049
Total	61	19	19	0.86011	0.00433

Table 4. Intra-population mt*COI* diversity statistics for the quill worm, *Hyalinoecia longibranchiata*. The regions are designated as CP for Challenger Plateau and CR for Chatham Rise.*n* is the total number of individuals sampled for a site, S is the number of polymorphic nucleotide sites in the sequence, h is the number of haplotypes represented at the site, Hd is haplotype diversity, and pi is the is nucleotide diversity.

Site (region)	**n**	**S**	**h**	**Hd**	**π**
C102 (CP)	9	6	3	0.55556	0.00413
3B (HM)	7	5	2	0.47619	0.00454
14a (HM)	8	12	4	0.78571	0.00988
9D28 (CR)	7	5	5	0.85714	0.00309
7A07 (CR)	8	5	5	0.85714	0.00354
1B15 (CR)	7	5	3	0.66667	0.00309
6C63 (CR)	6	4	3	0.73333	0.00331
3XC2 (CR)	6	2	3	0.73333	0.00178
Total	58	30	20	0.92801	0.01381

Geographic Distribution of Haplotypes

The squat lobster, *M. gracilis* (n=52), had high haplotype diversity for the *COI* gene (Table 2). Of the 36 haplotypes, only three were shared and the remaining 33 were unique. Two of the three shared haplotypes were found across the Challenger Plateau and the Chatham Rise and the third was absent on the Challenger Plateau but present on the Chatham Rise (Figure 2).

The *COI* sequence dataset for *H. longibranchiata* (n=58) consisted of 10 haplotypes shared by more than one individual and 10 unique haplotypes for a total of 20 haplotypes (Figure

3). The *16S* sequence dataset for *H. longibranchiata* (n=61) consisted of nine haplotypes shared by more than one individual and 10 unique haplotypes, for a total of 19 haplotypes (Figure 4). All the haplotypes found at the Challenger Plateau site were unique to that population. There were two haplotypes only found at the Hikurangi Margin sites. Three shared haplotypes were only found on the Chatham Rise, spanning the length of the Rise.

Figure 2. Distribution of *COI* haplotypes across the study area for *Munida gracilis*. The map shows the location of the study sites with pie charts indicating the haplotype composition of the population from that site. Each color represents a haplotype with Red, Blue, and Yellow representing the three shared haplotypes that are found across the study area. Shades of grey and other muted colors represent unique haplotypes. Sample size for each site is indicated.

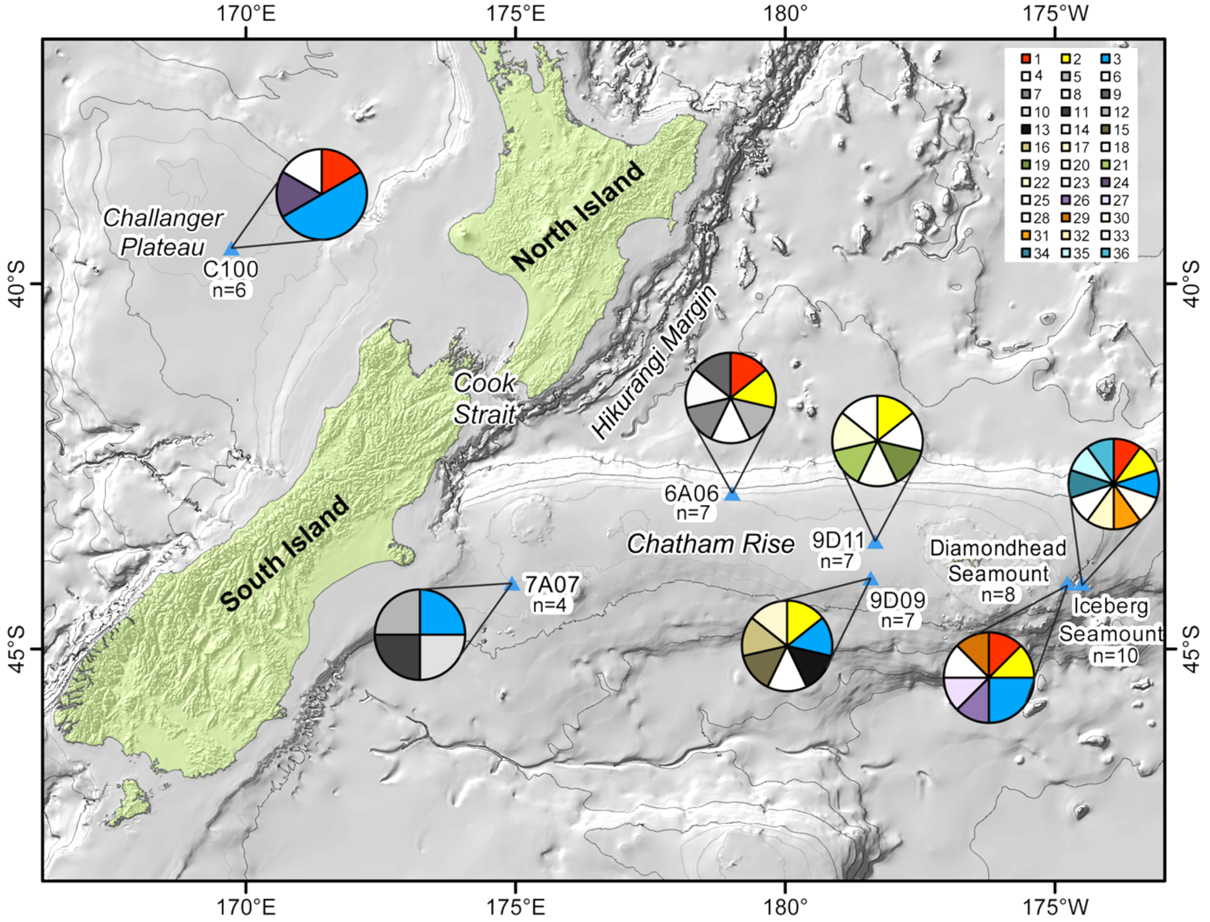


Figure 3. Distribution of *16S* haplotypes across the study area for *Hyalinoecia longibranchiata*. The map shows the location of the study sites with pie charts indicating the haplotype composition of the populations from that site. Each color represents a shared haplotype. White, black, and shades of grey represent unique haplotypes. Sample size for each site is indicated.

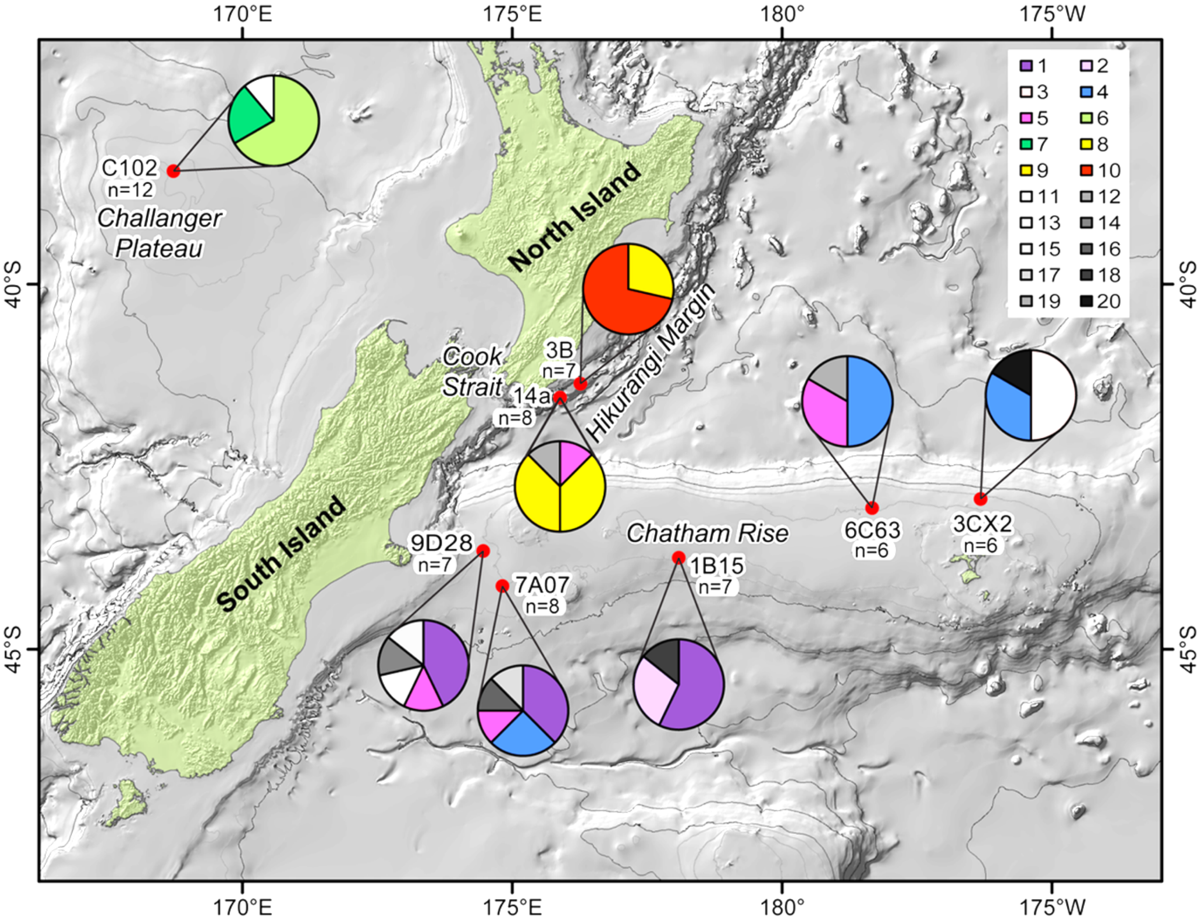
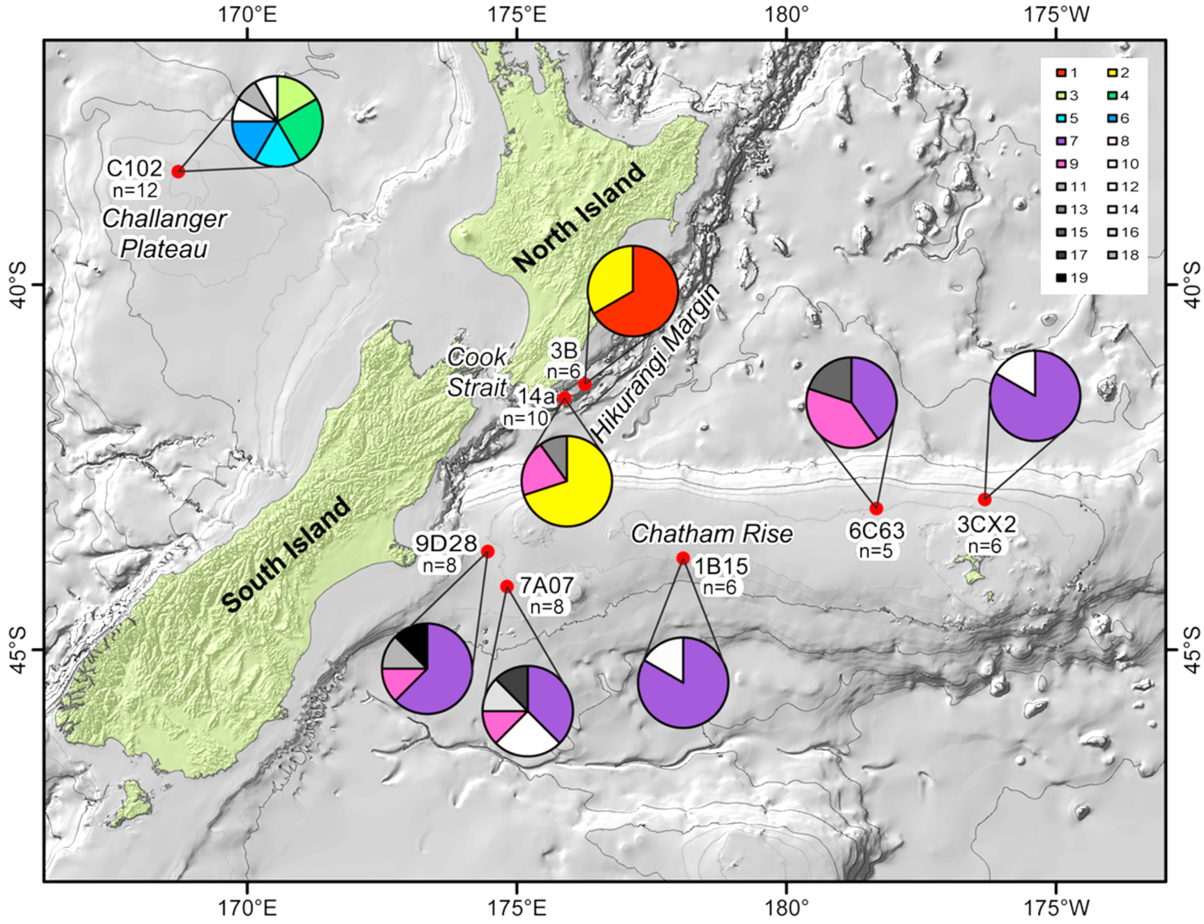


Figure 4. Distribution of *COI* haplotypes across the study area for *Hyalinoecia longibranchiata*. The location of the study sites with pie charts indicate the haplotype composition of the population from that site. Each color represents a shared haplotype. White, black, and shades of grey represent unique haplotypes. Sample sizes for each site are indicated.



Haplotype Networks

The *COI* haplotype network for *M. gracilis* reflects the large number of haplotypes, many of which are separated by one or a few nucleotide changes (Figure 5). There is no clear ancestral haplotype, and no significant geographic pattern to the network. This is consistent with the high level of sequence diversity of the *COI* gene of *M. gracilis*.

The *COI* and *16S* haplotype networks (Figure 6) for *H. longibranchiata* are consistent with the geographic structure indicated by the haplotype maps. There appears to be a central, ancestral *16S* haplotype that is present across the Chatham Rise sites, radiating out to the Hikurangi Margin and on to the Challenger Plateau. The *COI* haplotype sequences were more diverse than the *16S* sequences, consistent with common rates of mutation in these genes (*e.g.,* Munasinghe *et al.,* 2003).

Figure 5. TCS haplotype network for *Munida gracilis*, *COI* sequences. Each circle represents an observed haplotype and the circles are proportional to the number of individuals sampled with that haplotype. Each color indicates a sampling site and when a haplotype was present at multiple sites, a pie chart indicates the proportions with absolute numbers appearing in text in the pie chart. Each line connecting colored circles represents a single nucleotide sequence change. Dotted lines indicate haplotypes that could also be connected to alternative nodes. Lines with small black circles indicate interior haplotypes not found in the dataset (multiple nucleotide changes between sampled haplotypes).

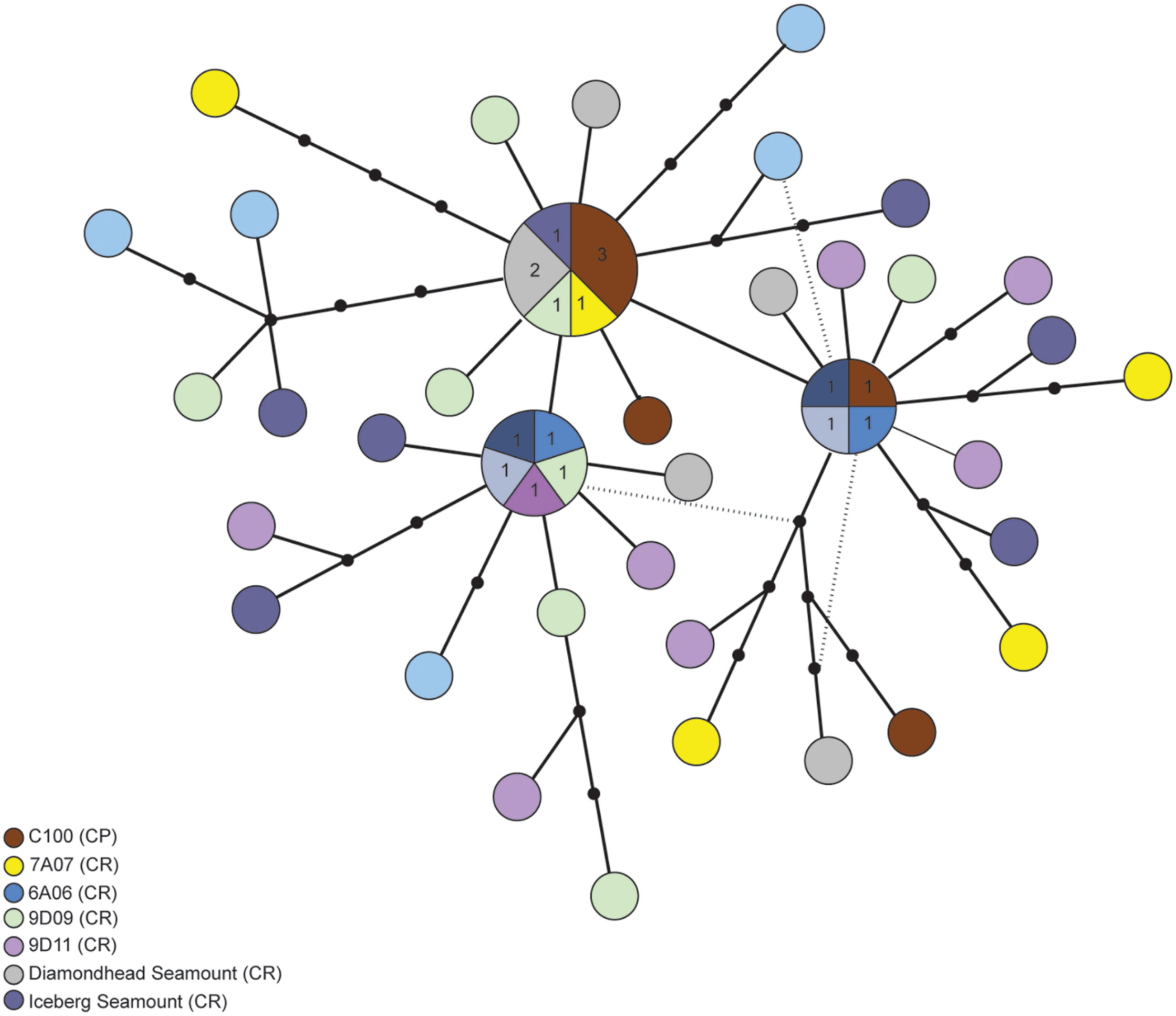
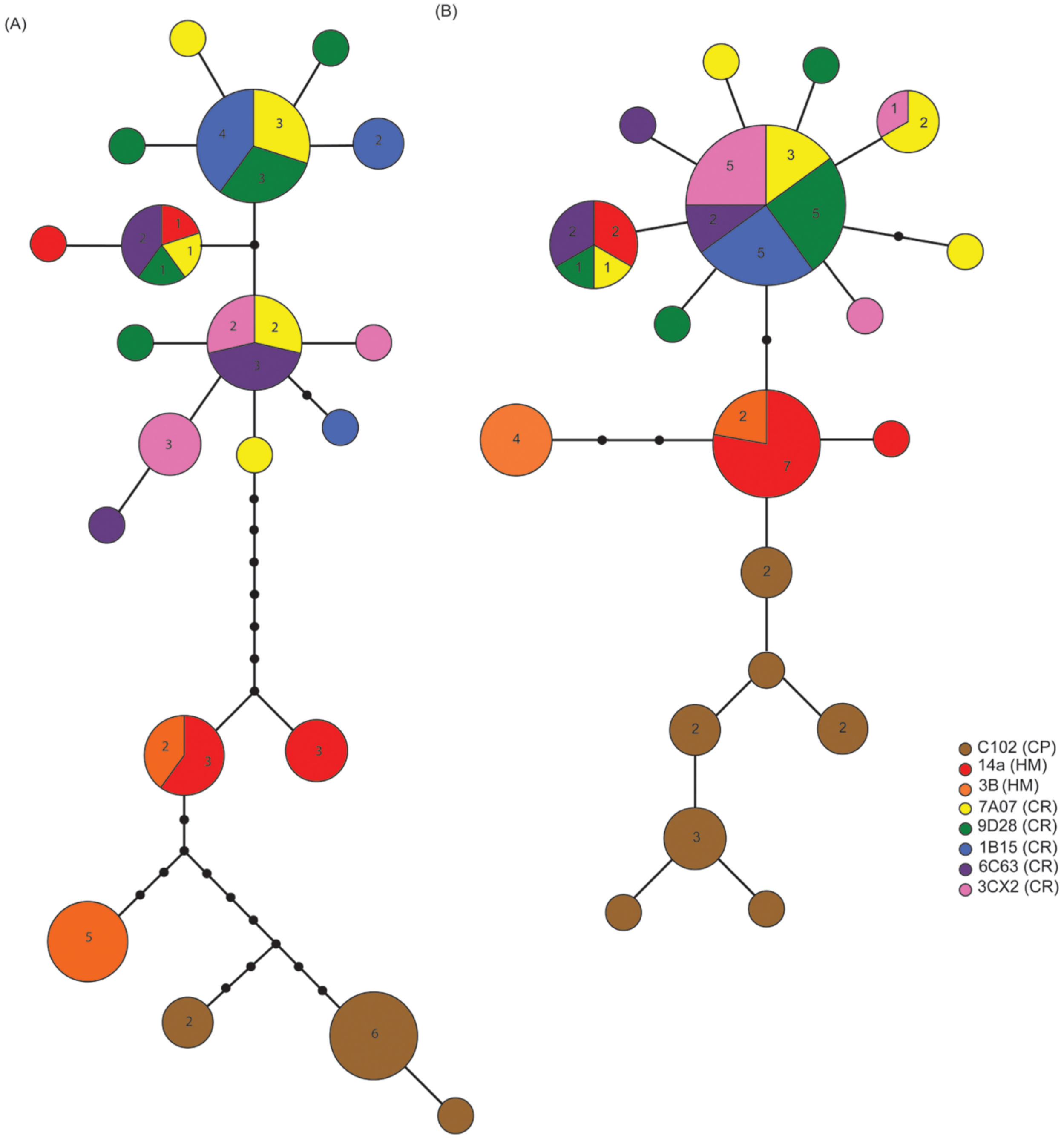


Figure 6. TCS haplotype networks for *Hyalinoecia longibranchiata.* Part (A) shows results for *COI* sequences and part (B) shows results for *16S* sequences. Each circle represents an observed haplotype and the circles are proportional to the number of individuals sampled with that haplotype. Each color indicates a sampling site and when a haplotype was present at multiple sites, a pie chart indicates the proportions with absolute numbers appearing in text in the pie chart. Each line connecting colored circles represents a single nucleotide sequence change. Lines with small black circles indicate interior haplotypes not found in the dataset (multiple nucleotide changes between sampled haplotypes).



Population Structure

F_{ST} data indicated no genetic structure in the sampled *M. gracilis* populations. Pairwise population F_{ST} values were all below 0.1 and none of the p values showed significant difference among the populations (Table 5). In the sampled *H. longibranchiata* populations, pairwise population F_{ST} values for *16S* indicated that the population at the Challenger Plateau site was significantly different ($p<0.05$) than all other populations ($F_{ST} >0.564$), as were both populations from the Hikurangi Margin sites ($F_{ST} >0.484$), which were also different from one another ($F_{ST} = 0.47989$). Differences between populations within the Chatham Rise were small (<0.17647) and only the difference between the two northeast populations at 6C63 and 3CX2 was statistically significant (Table 6).

Table 5. Pairwise F_{ST} values between populations of the squat lobster, *Munida gracilis*, using a fragment of the *COI* gene. Above the diagonal indicates ranges of p-values. The "-" denotes a p>0.05. The "*" denotes a p < 0.05. The "**" denotes a p < 0.01. The "***" denotes a p< 0.001.

	C100	**7A07**	**6A06**	**9D11**	**9D09**	**Diamondhead Seamount**	**Iceberg Seamount**
C100		-	-	-	-	-	-
7A07	0.04135		-	-	-	-	-
6A06	0.02383	0.03294		-	-	-	-
9D11	0.03655	0.04033	0.04692		-	-	-
9D09	-0.02666	0.02273	-0.0309	0.01299		-	-
Diamondhead Seamount	-0.04481	0.0536	-0.0064	-0.00304	-0.03896		-
Iceberg Seamount	-0.05198	-0.00004	-0.03325	-0.00142	-0.05346	-0.05184	

Table 6. Pairwise F_{ST} Values between populations of the quill worm, *Hyalinoecia longibranchiata,* using a fragment of the *16S* gene. Above the diagonal indicates ranges of p-values. The "-" denotes a p>0.05. The "*" denotes a p < 0.05. The "**" denotes a p < 0.01. The "***" denotes a p< 0.001.

	C102	**3B**	**14a**	**9D28**	**7A07**	**1B15**	**6C63**	**3CX2**
C102		***	***	***	***	***	***	***
3B	0.57829		**	***	***	***	***	***
14a	0.564	0.47989		**	***	***	*	***
9D28	0.73021	0.7446	0.53759		-	-	-	-
7A07	0.69463	0.68352	0.484	0.00408		-	-	-
1B15	0.73454	0.768	0.56923	-0.01659	0.01118		-	-
6C63	0.70138	0.71154	0.50162	0.01202	0.01408	0.15141		*
3CX2	0.74672	0.78462	0.58732	-0.0084	-0.05466	0.00201	0.17647	

The *COI* results for *H. longibranchiata* revealed similar trends to the *16S* data, however, there were significant differences between some Chatham Rise populations (F_{ST} >0.31089) (Table 7). Specifically, populations in the northeast were significantly different from those in the southwest and south central part of the Rise.

Table 7. Pairwise F_{ST} values between populations of the quill worm, *Hyalinoecia longibranchiata,* using a fragment of the *COI* gene. Above the diagonal indicates ranges of p-values. The “-“ denotes a p>0.05. The “*” denotes a p < 0.05. The “**” denotes a p < 0.01. The “***” denotes a p< 0.001.

	C102	**3B**	**14a**	**9D28**	**7A07**	**1B15**	**6C63**	**3XC2**
C102		***	***	***	***	***	***	***
3B	0.76723		**	***	***	***	***	***
14a	0.66914	0.38421		***	**	***	***	***
9D28	0.84878	0.81818	0.58117		-	-	**	***
7A07	0.83886	0.80383	0.56703	-0.00474		-	-	**
1B15	0.85298	0.82278	0.59817	-0.0303	0.06264		***	**
6C63	0.84071	0.80985	0.54635	0.31089	0.10061	0.38353		-
3XC2	0.86515	0.84923	0.62322	0.53833	0.33767	0.5605	0.11111	

Analysis of Molecular Variance

Hyalinoecia longibranchiata AMOVA results (Table 8 and Table 9) were consistent with our other analyses. The AMOVA for the three regions revealed that there was higher variation between regions than within regions, indicating that there is significant genetic structure across the study area. The Hikurangi Margin sites were shown to be statistically different from the Chatham Rise sites. The AMOVA tests for differences between groups of sites on the southwest and northeast of the Chatham Rise revealed a greater diversity in population structure within groups than between groups, indicating that this test does not show significant structure based on the northeast-southwest divide.

Table 8. *16S* AMOVA results for *Hyalinoecia longibranchiata*.

Test	**Source of variation**	**df**	**SS**	**Var. comp.**	**% V**	**P value**
Three Regions	among regional groups	2	49.641	1.24887 Va	63.54	0.00782 ± 0.00280
	among populations within regional groups	5	7.231	0.12367 Vb	6.29	0.00000 ± 0.00000
	within populations	54	32.015	0.59288 Vc	30.17	0.00000 ± 0.00000
Margin v. Rise	among regional groups	1	18.894	0.79467 Va	55.88	0.04399 ± 0.00714
	among populations within regional groups	5	7.231	0.13876 Vb	9.76	0.00293 ± 0.00164
	within populations	43	21.015	0.48873 Vc	34.37	0.00000 ± 0.00000
NE v. SW CR	among regional groups	1	0.195	-0.02046 Va	-5.27	1.00000 ± 0.00000
	among populations within regional groups	3	1.548	0.01853 Vb	4.77	0.10459 ± 0.00793
	within populations	29	11.315	0.39019 Vc	100.5	0.24829 ± 0.01653

Table 9. *COI* AMOVA results for *Hyalinoecia longibranchiata*.

Test	**Source of variation**	**df**	**SS**	**Var. comp.**	**% V**	**P value**
Three regions	among regional groups	2	127.433	3.58675 Va	69.07	0.00978 ± 0.00294
	among populations within regional groups	5	21.941	0.47015 Vb	9.05	0.00000 ± 0.00000
	within populations	50	56.815	1.13631 Vc	21.88	0.00000 ± 0.00000
Margin v. Rise	among regional groups	1	64.726	2.88919 Va	64.14	0.04106 ± 0.00536
	among populations within regional groups	5	21.941	0.46869 Vb	10.41	0.00000 ± 0.00000
	within populations	42	48.149	1.14640 Vc	25.45	0.00000 ± 0.00000
NE v. SW CR	among regional groups	1	7.899	0.44418 Va	35.09	0.07722 ± 0.01012
	among populations within regional groups	3	3.043	0.03277 Vb	2.59	0.21799 ± 0.01260
	within populations	29	22.881	0.78900 Vc	62.32	0.00000 ± 0.00000

DISCUSSION

Our study is one of few that have examined genetic connectivity of deep-sea invertebrate populations in the New Zealand EEZ. Using mitochondrial *COI* and *16S* genes as genetic markers, we tested for genetic structure among populations of the squat lobster *Munida gracilis* and the quill worm *Hyalinoecia longibranchiata* at sites across three deep-sea regions near New Zealand: the Chatham Rise, Hikurangi Margin, and Challenger Plateau. The study aimed to address a number of questions related to the factors that determine genetic connectivity in the deep sea as well as inform the design and evaluation of MPAs in the deep sea. Our results are discussed below in relation to each of these study questions (paraphrased below).

Is there regional genetic structure across the study area?

The population structure on a regional scale for *H. longibranchiata* provides evidence of little to no historic gene flow between the Challenger Plateau, the Hikurangi Margin, and the Chatham Rise. In contrast, the sampled *M. gracilis* population data demonstrated high haplotype diversity for *COI* and no population structure at the geographic scale examined in this study. A number of factors are known to affect genetic connectivity including large and small-scale current regimes, topography, settlement habitat, depth, dispersal strategies, adult mobility and reproductive success, etc. (Cowen & Sponaugle 2009). While it was not our goal to isolate a single factor as the cause of an observed population genetic structure, we can examine the consistency and potential interplay of each factor in relation to the results.

The observed large-scale genetic differences in *H. longibranchiata* populations between the three regions can be explained partly by geographic distribution and partly by currents. Large geographic distance between sites can limit connectivity, especially for species with low dispersal capability. The Hikurangi Eddy, located to the East of Cook Strait could create an isolated water mass around the study sites at the southern end of the Hikurangi Margin (Chiswell & Booth 1999), and limit dispersal. There is one shared haplotype between the Rise and one of the sites on the margin, suggesting that historically there has been some ability of individuals to disperse between the two regions. Variation in the spatial extent of the Hikurangi Eddy could transport larvae or adults between the Margin and the Rise. The lack of any shared haplotypes between the Chatham Rise/Hikurangi Margin and the Challenger Plateau is consistent with the Cook Strait functioning as a barrier to dispersal, rather than a conduit for transporting larvae or adults between the western and eastern side of New Zealand. Barnes (1985) found that, despite large tidal flow, a front exists in Cook Strait with up to a 2°C gradient that causes negligible net flow—at least near the sea surface—through the Strait.

In contrast to the quill worm, we found unstructured yet genetically diverse populations of the squat lobster *M. gracilis* throughout the study area. Based on the large proportion of unique haplotypes, we conclude that the genetic diversity of *COI* in the *M. gracilis* population has not been fully ascertained. Considering the high levels of diversity in the mitochondrial genes we sampled, it is difficult to draw conclusions about the *M. gracilis* population except to say that the presence of certain haplotypes across the study area indicates that there is likely a single population with high levels of mixing not impeded by geographic distance or current patterns.

As with all population genetic studies, the numbers of individuals, loci, and sites have to be considered in the interpretation of the data. The available sample sizes at various sites in the study area (which we have termed populations) are not considered large, nor are they consistent across sites. However, this is not atypical for deep-sea population genetic studies in which collecting large sample sizes yielding highly robust estimates of genetic diversity is considerably difficult given the inaccessibility and expense of obtaining these populations. Comparing haplotype diversity between sites with highly variable sample sizes could lead to inappropriate assumptions about spatial patterns of haplotype diversity, including the performance of unbiased estimators and rarefaction methods (Pruett & Winker 2008), with lower sample sizes likely underestimating levels of diversity. We calculated F_{ST} from haplotype-frequencies and pairwise DNA sequence diversity. Haplotype frequency-based statistics are more sensitive for small sample sizes, while the sequence-based statistic can be considered a more sensitive method for detecting population structure in highly polymorphic loci. Our goal was not to examine the effect of small or variable sample sizes on genetic estimates in genetically diverse datasets (Pruett & Winker 2008), and we caution the over-interpretation of our results. Given the high COI haplotype diversity of *in M. gracilis*, it is likely that more individuals would provide more informative results. Despite the limited *H. longibranchiata* data set, the data provide insight into the population structure of this species, and the results are supported by both genetic markers (*16S* and *COI*). From our initial genetic survey and for future studies of these species, the *COI* gene can be considered a useful marker for resolving genetic structure in *H. longibranchiata*, and somewhat less in *M. gracilis*.

Is there genetic structure within the three regions?

In addition to the larger scale patterns discussed above, population structure for *H. longibranchiata* was observed between the seamount and slope sites on the Hikurangi Margin and potential differentiation was detected between populations on the northeast and southwest sites on the Chatham Rise. No genetic structure between sites of varying habitat types—specifically between seamount and slope—was observed for *M. gracilis* in any region.

Addressing questions of genetic connectivity is especially complex in the deep sea given that suitable habitats can be patchy over large spatial scales (hundreds to thousands of km). For example, several thousand kilometers may separate hydrothermal vent fields or seamounts and yet gene flow may occur between the geographically distant sites of the same habitat type (Craddock *et al.* 1995; Plouviez *et al.* 2010; Vrijenhoek 2010; Cho & Shank 2010). The opposite can also be true where small distances between patches of the same habitat do not necessarily translate into genetic connectivity among populations if there are physical or biological barriers to dispersal (Cho & Shank 2010).

An added layer of complexity to deep-sea connectivity is the potential for inter-habitat connectivity when different habitats may be found in a small geographic area. For example, in the Norfolk Ridge seamount system, populations on the seamounts have been shown to be genetically connected to populations on the island slope (Samadi *et al.* 2006). Conversely, populations at different habitats may not be well connected when the physical or biological attributes of one habitat serve to isolate it from other suitable habitats (*e.g.,* the presence of local isolating hydrographic features as seen in some seamount systems as in Lavelle and Mohn, 2010).

Because of the constrained sample availability, the sites from which we were able to obtain samples and the sample sizes did not provide a robust enough dataset to fully understand the extent of inter-habitat genetic connectivity. Nevertheless, our data provide some interesting indications about inter-habitat connectivity. While populations of *H. longibrachiata* at the Hikurangi Margin sites were significantly different from the Challenger Plateau sites and the Chatham Rise sites, they were also significantly different from each other. These sites—one a seamount, the other a slope—are only ~38 km apart, but are separated by a small canyon. It is possible that there is a local current regime on this margin that limits the connectivity of these two populations or perhaps some habitat preference that results in the observed difference in genetic structure. The seamount in question may likely be considered too small for localized isolating hydrographic features. So, alternatively, it could be the predominance of down slope currents associated with canyons rather than along slope currents on this margin that are responsible for the limited connectivity between the populations of quill worms at the seamount and slope sites. For *M. gracilis*, the two seamount sites in the Andes Seamount cluster—Diamondhead Seamount and Iceberg Seamount—shared haplotypes with populations at sites elsewhere on the Chatham Rise, indicating that for this species in the New Zealand region, habitat type may not play a strong role in genetic connectivity.

Results suggest that there is variation in the level of connectivity across the Rise. Other studies have found marked differences in benthic community structure between the northern and southern flanks of the rise, which have been attributed to the different environmental and biological conditions imposed by the location of the Subtropical Front (Probert & McKnight 1993; McKnight & Probert 1997; Nodder *et al.* 2003; Berkenbusch *et al.* 2011). It is possible that the currents that maintain the STF present a significant barrier to dispersal of individuals among populations of the same species between northern and southern sites. While the location of our study sites for *H. longibranchiata* did not allow us to separate strict north-south effects from possible east-west effects, we were able to test for northeast to southwest variation in the genetic make-up of populations. *COI*, but not *16S*, data provide some support for this hypothesis for *H. longibrachiata* because populations at sites on the northeast of the rise were sufficiently different from those on the southwest of the rise.

Depth has been shown to play a major role in connectivity within and between deep-sea ocean basins and slopes (McClain *et al.* 2010) and seamounts (Cho & Shank 2010). It is worth remembering here that our samples were explicitly chosen to fall within a small range of depths, a method of sample selection that could have resulted in a reduced ability to detect whether there was an effect of the STF on population structure on the rise. Nodder *et al.* (2003) found that the most notable differences in benthic communities between northern and southern sites were evident at greater depths (*e.g.,* sites at 2300 meters). It is possible that populations located deeper than our study sites (below 746 m) could be less well connected across the rise than our results suggest. It is also possible that shallower populations on the crest of the rise (approximately 200 m) may be well connected with one another within the core of the STF, yet be poorly connected to populations at greater depths.

Do the inferred life history strategies correlate with the observed patterns of genetic connectivity?

The results of our study indicate that the genetic connectivity patterns of the two study species are different. Differences in life history strategy and inferred pelagic larval duration likely explain the difference between observed patterns in the two species. Larval dispersal and adult mobility contribute to making the squat lobster a better potential disperser than the quill worm and correlate directly with the differences in inferred patterns of connectivity. The relationship between life history and population structure across seamounts is well documented in Samadi *et al.* (2006), in which the authors found that species with broad dispersal potential had limited to no population structure while the other species with limited dispersal potential had clear population structure.

What implications do the patterns of genetic population connectivity have for MPA design and the efficacy of the current BPAs?

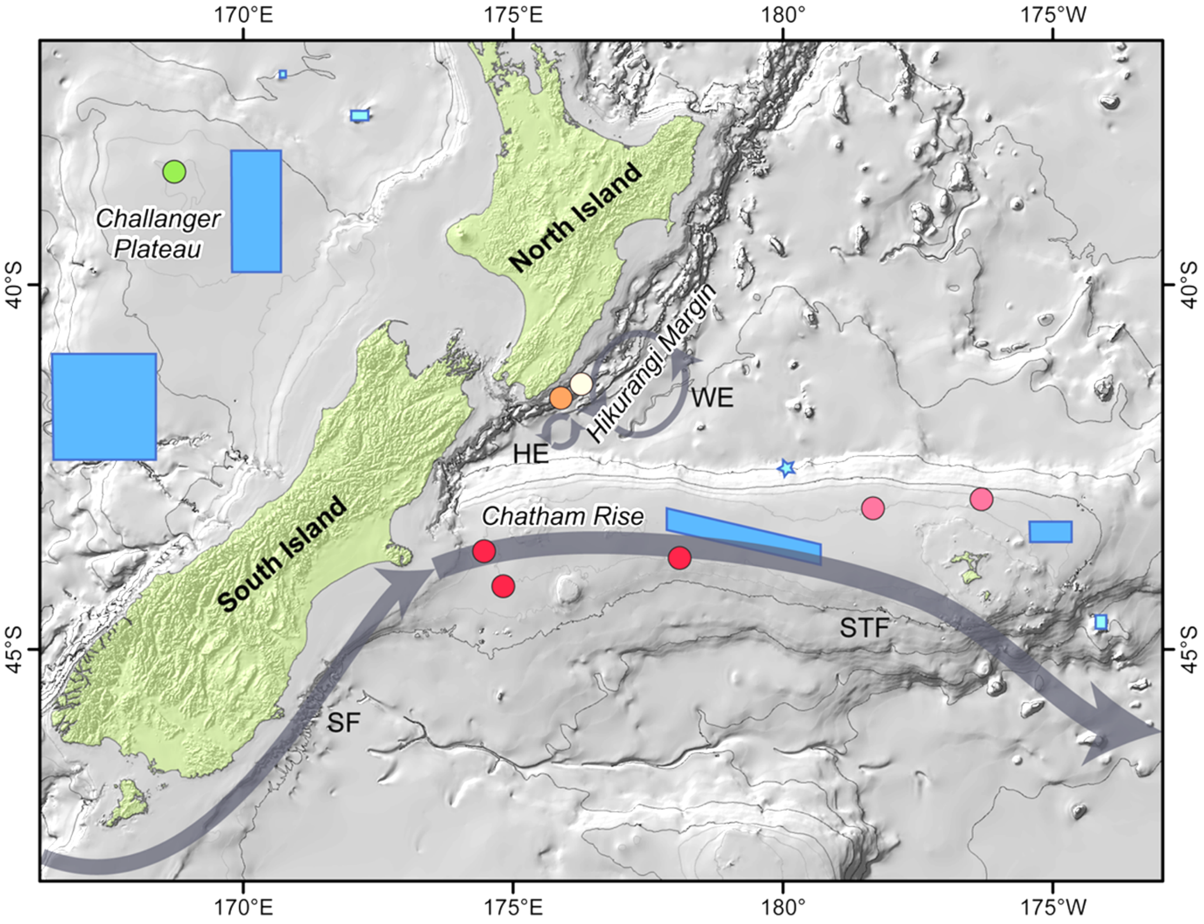
In 2007, a fishing industry-driven initiative resulted in the creation of seventeen areas within the New Zealand EEZ that were designated Benthic Protection Areas (BPAs). These areas comprise roughly 30% of the EEZ and are closed to bottom trawling, but not to other uses such as mining. Still, they are considered by some (Helson *et al.* 2010) to fulfill New Zealand's dedication to protecting at least 10% of its marine environment (New Zealand Biodiversity Strategy, 2000). The selection criteria for the BPAs included size, low fishing levels, geometrically simple boundaries, and representativeness of the Marine Environment Classification (Helson *et al.* 2010). The population connectivity of benthic organisms was not directly considered in the design of the BPAs.

A deep-water MPA process is scheduled to commence in 2013 (Ministry of Fisheries and Department of Conservation, 2008), coinciding with a review of the BPAs (Helson *et al.* 2010). To facilitate an effective review of BPAs, the closed seamounts, and the future deep-water protected area design process, the "best available" scientific information concerning the habitats and faunal communities need to be considered, as well as input from "offshore experts." To date, only a single study has challenged the efficacy of BPA design by demonstrating that BPAs located at alternate sites could be more effective at protecting biodiversity and less costly to the fishing industry (Leathwick *et al.* 2008). The results of our study provide additional information that can be used to evaluate the placement of BPAs and future deep-water MPAs.

The difference in genetic population structure between the squat lobster and the quill worm confirms that, in terms of connectivity, MPA design should consider the implications of protecting assemblages of species with different life history strategies (Airamé *et al.* 2003). The findings of our study for common species with high levels of dispersal (like *M. gracilis*), indicate that populations in different closed areas have historically been well connected, and one can reasonably presume are currently connected. However, for common species with limited dispersal capabilities (like *H. longibranchiata*), our study findings provide a framework with which one can analyze the efficacy of future MPA design. At the broadest level, our main finding is that a species with direct development has pronounced population structure across the Challenger Plateau, Hikurangi Margin, and the Chatham Rise (Figure 7). If the maintenance of genetically distinct populations is considered integral to the goal of protecting biodiversity, then large protected areas that possess isolated populations will help to further that goal. Presently, there are BPAs on both the Chatham Rise and the Challenger Plateau but there are no closed areas on the Hikurangi Margin. Other large areas in the study region may also possess populations similarly genetically isolated by current regimes such as large eddy systems, that could also be considered in the future design of deep-water MPAs.

Our results suggest that on smaller spatial scales within regions, local topography and current regimes may have a profound impact on gene flow, leading to the differentiation of populations at different habitats. Populations at slope and seamount habitat in close proximity on the Hikurangi Margin were shown to host genetically different populations. There are some protected areas in the study area that have been specifically closed to trawling in order to protect the communities on seamounts, in part because of the then perceived isolated nature of seamount fauna (Bordie, 2003). The current BPAs, because of the large size design criterion (Helson *et al.* 2010), protect multiple habitats including seamounts and hydrothermal vents that are perceived to represent vulnerable marine ecosystems (FAO, 2007 and 2008). As such, any large protected area should afford some protection to any genetically distinct populations found at different habitats within a region.

Figure 7. Map of the study area showing genetically distinct populations (colored circles) of the worm *H. longibranchiata* relative to the position of Benthic Protection Areas (blue) and seamount closures (light blue), and local currents. Populations on the Challenger Plateau, Hikurangi Margin and Chatham Rise are green, orange, and red, respectively, with different shades of the latter two colors representing within region differences in genetic population structure. Blue rectangles represent Benthic Protection Areas and Seamount Closures (in light blue). The star marks two Seamount Closures too small to be visible on the map. The approximate position of the Southland Front (SF), the Sub-Tropical Front (STF), the Hikurangi Eddy (HE), and Wairarapa Eddy (WE) are shown with grey bands and arrows. The location of the STF is based on Figure 1 of (Hayward *et al.* 2008).



The results of the within region comparison suggest that the location of BPAs on the Chatham Rise may require revision. The central BPA on the Chatham Rise is located in the middle of the crest of the rise at depths of 300–450 m and the BPA at the eastern end of the rise extends over a depth range of 300–900 m. We do not have any population data from these specific locations but we have shown the potential for genetic variation across the Rise. Given this finding and our understanding about differences in benthic communities on the north and south flank of the Rise (Nodder *et al.* 2003), and the likelihood that populations are genetically structured by depth (Cho & Shank 2010; Miller *et al.* 2011), the two BPAs arranged along the axis of the Rise at shallow depths may not be sufficient to protect the genetic variation of populations on the Chatham Rise. The Chatham Rise is one of the largest geomorphic features of the New Zealand EEZ with a complex and productive ecosystem (Nodder *et al.,* 2012), yet large areas of its seafloor are subjected to disturbance from bottom trawling, and in the future disturbance from mining for phosphorite nodules is likely. Our results suggest that further protected areas, or a re-positioning of the current BPAs, could be considered to afford greater protection to the benthic biodiversity associated with the Chatham Rise through genetic connectivity.

Future Directions

Our assessment of the genetic connectivity of two abundant benthic invertebrates found throughout a range of deep-sea habitats in the New Zealand EEZ represents a step towards understanding the spatial structure of benthic communities in the New Zealand region, and informing the future design of deep-water MPAs in the region. However, the study has raised a number of questions about the populations of *H. longibranchiata*. For example, what is the true geographic extent of populations found in the three regions—Challenger Plateau, Hikurangi Margin, and Chatham Rise? Are the Hikurangi Margin haplotypes found along the margin to the south or north? Similarly, are quill worm populations on the central north part of the Chatham Rise unique to these sites or will they resemble populations at the northeastern sites? What about populations at other sites to the west of New Zealand? Is the population of quill worms at the Challenger Plateau site different from the sites at the other study regions simply because it is on the western side of New Zealand or is the plateau in some way isolated? Such questions apply to other invertebrate species with potentially limited dispersal capabilities. Future genetic studies of population connectivity should include a greater range of study species in order to generate information useful for the design of protected areas in the deep sea.

ACKNOWLEDGEMENTS

The authors would like to thank the captain, crew, and scientific teams of the *R/V Tangaroa* for the following cruises: TAN0705, TAN0707, TAN0905, TAN1004. Particular thanks are extended to Dr. Kareen Schnabel of the NIWA Invertebrate Collection (NIC) for her ongoing advice and assistance, and to other NIC staff for facilitating access to the collection. We also acknowledge Dr. Phil Ross from the University of Waikato for assistance in data analysis and Arne Pallentin of NIWA for creating maps in GIS. We would like to recognize the critical importance of the Fulbright Fellowship program's goal of international exchange to this work. We also thank the Woods Hole Oceanographic Institution Joint Program Committee for Biological Oceanography for allowing EK Bors to pursue this research in New Zealand.

References

Airamé S, Dugan JE, Lafferty KD, Leslie H (2003) Applying ecological criteria to marine reserve design: a case study from the California Channel Islands. *Ecological Applications* 13(1): S170-S184.

Apte S, Gardner J (2002) Population genetic subdivision in the New Zealand greenshell mussel (*Perna canaliculus*) inferred from single-strand conformation polymorphism analysis of mitochondrial DNA *Molecular Ecology* 11(9): 1617-1628.

Baba K, Fujita Y, Wehrtmann IS, Scholtz G (2011) Chapter 5. Developmental biology of squat lobsters. In: Poore GCB, Ahyong ST, Taylor J, editors. The Biology of Squat Lobsters. Melbourne: CSIRO Publishing.

Baird SJ, Wood BA, Bagley NW (2011) Nature and extent of commercial fishing effort on or near the seafloor within the New Zealand 200 n. mile exclusive economic zone, 1989–90 to 2004–05. New Zealand Aquatic Environment and Biodiversity Report 73. 143 p.

Barange M, Field JG, Steffen W (2010) Introduction: Oceans in the earth system. In: Barange E, Field JG, Harris RP, Hofmann EE, Perry RI, et al., editors. Marine Ecosystems and Global Change. Oxford: Oxford University Press. pp. 1-10.

Barnes EJ (1985) Eastern Cook Strait region circulation inferred from satellite-derived, sea-surface, temperature data. N Z J Mar Freshwat Res 19(3): 405-411.

Berkenbusch K, Probert PK, Nodder SD (2011) Comparative biomass of sediment benthos across a depth transect, Chatham Rise, Southwest Pacific Ocean. *Marine Ecology Progress Series*, 425, 79–90.

Brodie S, Clark M (2003) The New Zealand Seamount Management Strategy – steps towards conserving offshore marine habitat. In: Beumer JP, Grant A, Smith DC, editors. Aquatic Protected Areas: what works best and how do we know? Proceedings of the World Congress on Aquatic Protected Areas, Cairns, Australia, August 2002. pp. 664–673.

Carrasco F (1983) Description of adults and larvae of a new deep-water species of *Hyalinoecia* (Polychaeta, Onuphidae) from the Southeastern Pacific Ocean. J Nat Hist 17: 87-93.

Chiswell SM, Booth JD (1999) Rock lobster Jasus edwardsii larval retention by the Wairarapa Eddy off New Zealand. *Marine Ecology Progress Series*, 183, 227–240.

Cho W, Shank TM (2010) Incongruent patterns of genetic connectivity among four ophiuroid species with differing coral host specificity on North Atlantic seamounts. *Marine Ecology*, 31, 121–143.

Clark MR, Rowden AA (2009) Effect of deepwater trawling on the macro-invertebrate assemblages of seamounts on the Chatham Rise, New Zealand. *Deep Sea Research Part I: Oceanographic Research Papers*, 56, 1540–1554.

Clark MR, Rowden AA, Schlacher T *et al.* (2010) The Ecology of Seamounts: Structure, Function, and Human Impacts. *Annual Review of Marine Science*, 2, 253–278.

Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*.

Cowen RK, Sponaugle S (2009) Larval Dispersal and Marine Population Connectivity. *Annual Review of Marine Science*, 1, 443–466.

Craddock C, Hoeh WR, Lutz RA, Vrijenhoek RC (1995) Extensive Gene Flow Among Mytilid (Bathymodiolus-Thermophilus) Populations From Hydrothermal Vents of the Eastern Pacific. *Marine Biology*, 124, 137–146.

Crown Minerals (2010) New Zealand petroleum basins. Ministry of Economic Development, Wellington, New Zealand. 110 p. http://www.nzpam.govt.nz/cms/pdf-library/petroleum-publications-1/2010%20NZ%20Petroleum%20Basin%20Report-WEB.pdf Accessed 2012 May.

Cryer M, Hartill B, O'Shea S (2002) Modification of marine benthos by trawling: towards a generalization for the deep ocean? *Ecological Applications*, 12, 1824–1839.

Excoffier L, Laval G (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics*.

Department for Environment Food and Rural Affairs, United Kingdom (2011) Biodiversity 2020: A strategy for England's wildlife and ecosystem services. Report PB13583. 48 p.

FAO (2007) Report and documentation of the Expert Consultation on deep-sea fisheries in the High Seas. FAO Fisheries Report No 838. 203 p.

FAO (2008) Report of the FAO Workshop on vulnerable ecosystems and destructive fishing in deep-sea fisheries. FAO Fisheries Report No 829. 18 p.

Folmer O, Black M, HoehW, Lutz RA, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular marine biology and biotechnology* 3(5): 294-299.

Gaines SD, Lester SE, Grorud-Colvert K, Costello C, Pollnac R (2010a) Evolving science of marine reserves: New developments and emerging research frontiers. *Proceedings of the National Academy of Sciences*, 107, 18251–18255.

Gaines SD, White C, Carr MH, Palumbi SR (2010b) Designing marine reserve networks for both conservation and fisheries management. *Proceedings of the National Academy of Sciences*, 107, 18286–18293.

Glasby G, Wright I (1990) Marine mineral potential in New Zealand exclusive economic zone rid B-9643-2008. Marine Mining 9(3): 403-427.

Gordon DP, Beaumont J, MacDiarmid A, Robertson DA, Ahyong ST (2010) Marine Biodiversity of Aotearoa New Zealand (SJ Goldstien, Ed,). *PLoS ONE*, 5, e10905–17.

Hayward BW, Scott GH, Crundwell MP *et al.* (2008) The effect of submerged plateaux on Pleistocene gyral circulation and sea-surface temperatures in the Southwest Pacific. *Global and Planetary Change*, 63, 309–316.

Heath RA (1985) A review of the physical oceanography of the seas around New Zealand — 1982. *New Zealand Journal of Marine and Freshwater Research*, 19, 79–124.

Hedgecock D, Barber PH (2007) Genetic approaches to measuring connectivity. *Oceanography*, 20.

Helson J, Leslie S, Clement G, Wells R, Wood R (2010) Private rights, public benefits Industry-driven seabed protection. *Marine Policy*, 34, 557–566.

Hendrson JR (1885) Diagnoses of new species of Galatheidae collected during the "challenger" expedition. Annals and Magazine of Natural History, Series V 16: 407-421.

Kojima S, Watanabe H, Tsuchida S, Fujikura K, Rowden AA, *et al*. (2006) Phylogenetic relationships of a tube worm (*Lamellibrachia juni*) from three hydrothermal vent fields in the south pacific. J Mar Biol Assoc U K 86(6): 1357-1361.

Leathwick J, Moilanen A, Francis M *et al.* (2008) Novel methods for the design and evaluation of marine protected areas in offshore waters. *Conservation Letters*, 1, 91–102.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25, 1451–1452.

Machordom A, Macpherson E (2004) Rapid radiation and cryptic speciation in squat lobsters of the genus Munida (Crustacea, Decapoda) and related genera in the South West Pacific: molecular and morphological evidence. *Molecular Phylogenetics and Evolution*, 33, 259–279.

Mackay KA, Wood BA, Calrk MR (2005) Chatham Rise Bathymetry. NIWA Miscellaneous Chart Series No. 82. National Institute of Water & Atmospheric Research, Wellington, New Zealand.

McClain CR, Etter RJ, Stuart CT, Rex MA (2010) Biogeography of the deep-sea gastropod *Oocorys sulcata* Fischer *1884*. Journal of Conchology, 40: 287-290.

McIntosh WC (1885) Report on the Annelida Polychaeta collected by H.M.S. 'Challenger' during the years 1873-76. Report of the Scientific Results of the Voyage of H.M.S. Challenger, 1873-1876, Zoology 12: 1-554.

McKnight DG, Probert PK (1997) Epibenthic communities on the Chatham Rise, New Zealand. *New Zealand Journal of Marine and*, 31, 505–513.

Miles EL (2009) On the Increasing Vulnerability of the World Ocean to Multiple Stresses. *Annual Review of Environment and Resources*, 34, 17–41.

Miller K (1997) Genetic structure of black coral populations in New Zealand's fiords. *Marine Ecology Progress Series*, 161, 123–132.

Miller KJ, Ayre DJ (2008) Protection of genetic diversity and maintenance of connectivity among reef corals within marine protected areas. *Conservation Biology*, 22, 1245–1254.

Miller KJ, Rowden AA, Williams A, Häussermann V (2011) Out of Their Depth? Isolated Deep Populations of the Cosmopolitan Coral Desmophyllum dianthus May Be Highly Vulnerable to Environmental Change (SJ Bograd, Ed,). *PLoS ONE*, 6, e19004–10.

Ministry of Fisheries and Department of Conservation (2008) Marine Protected Areas: Classification, Protection Standard and Implementation Guidelines. Wellington, New Zealand: Ministry of Fisheries and Department of Conservation. 54 p.

Munasinghe D, Murphy N, Austin C (2003) Utility of mitochondrial DNA sequences from four gene regions for systematic studies of Australian freshwater crayfish of the genus *Cherax* (Decapoda: Parastacidae). J Crust Biol 23(2): 402-417.

New Zealand Biodiversity Strategy, February 2000, ISBN O-478-21919-9, Published by The Department of Conservation.

New Zealand Ministry of Fisheries (2010) Report from the Fisheries Assessment Plenary, May 2010: stock assessments and yield estimates. 1158 p. http://www.fish.govt.nz/ Accessed 2012 May 1.

New Zealand Petroleum and Minerals, Permit database. http://www.nzpam.govt.nz/cms Accessed 2012 May 1.

Nielsen JF (2005) The molecular phylogenetics of Antarctic sea spiders (Pycnogonida) BSc Honours thesis, University of Auckland.

Nodder SD, Pilditch CA, Probert PK, Hall JA (2003) Variability in benthic biomass and activity beneath the Subtropical Front, Chatham Rise, SW Pacific Ocean. *Deep Sea Research Part I: Oceanographic Research Papers*, 50, 959–985.

Nodder SD, Bowden DA, Pallentin A, Mackay K (2012) Seafloor habitats and benthos of a continental ridge: Chatham Rise, New Zealand. In: Harris PT, Baker EK, editors. Seafloor geomorphology as benthic habitat: GeoHab Atlas of seafloor geomorphic features and benthic habitats. Elsevier Insights. pp. 763–776.

Orensanz J M (1990) The eunicemorph polychaete annelids from Antarctic and Subantarctic Seas with Addenda to the Eunicemorpha of Argentina, Chile, New Zealand, Australia, and the Southern Indian Ocean. Antarct Res Ser 52: 1-18.

Palumbi SR (2003) Population genetics, demographic connectivity, and the design of marine reserves. *Ecological Applications*.

Paxton, H (1986) Generic revision and relationships of the family Onuphidae (Annelida: Polychaeta). Records of the Australian Museum 38: 1-74.

Perrin C, Wing SR, Roy MS (2004) Effects of hydrographic barriers on population genetic structure of the sea star Coscinasterias muricata (Echinodermata, Asteroidea) in the New Zealand fiords. *Molecular Ecology*, 13, 2183–2195.

Plouviez S, Le Guen D, Lecompte O, Lallier FH, Jollivet D (2010) Determining gene flow and the influence of selection across the equatorial barrier of the East Pacific Rise in the tube-dwelling polychaete Alvinella pompejana. *BMC Evolutionary Biology*, 10, 220.

Probert PK (1999) Seamounts, sanctuaries and sustainability: moving towards deep-sea conservation. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 9, 601–605.

Probert PK, McKnight DG (1993) Biomass of bathyal macrobenthos in the region of the Subtropical Convergence, Chatham Rise, New Zealand. *Deep Sea Research Part I: Oceanographic Research Papers*, 40, 1003–1007.

Probert PK, McKnight DG, Grove SL (1997) Benthic invertebrate bycatch from a deep-water trawl fishery, Chatham Rise, New Zealand. *Aquatic Conservation*, 7: 27-40

Pruett CL, Winker K (2008) The effects of sample size on population genetic diversity estimates in song sparrows Melospiza melodia. *Journal of Avian Biology*.

Ramillien G, Wright IC (2000) Predicted seafloor topography of the New Zealand region: A nonlinear least squares inversion of satellite altimetry data. *Journal of Geophysical Research: Solid Earth*, 105, 16577–16590.

Ramirez-Llodra E, Tyler PA, Baker MC *et al.* (2011) Man and the Last Great Wilderness: Human Impact on the Deep Sea (P Roopnarine, Ed,). *PLoS ONE*, 6, e22588–25.

Read GB, Clark H (1999) Ingestion of quill-worms by the astropectinid sea-star Proserpinaster neozelanicus (Mortensen). *New Zealand Journal of Zoology*, 26, 49–54.

Ross PM, Hogg ID, Pilditch CA, Lundquist CJ, Wilkins RJ (2011) Population Genetic Structure of the New Zealand Estuarine Clam Austrovenus stutchburyi (Bivalvia: Veneridae) Reveals Population Subdivision and Partial Congruence with Biogeographic Boundaries. *Estuaries and Coasts*, 35, 143–154.

Samadi S, Bottan L, Macpherson E, De Forges BR, Boisselier M-C (2006) Seamount endemism questioned by the geographic distribution and population genetic structure of marine invertebrates. *Marine Biology*, 149, 1463–1475.

Shank TM (2010) Seamounts : deep-ocean laboratories of faunal connectivity, evolution, and endemism. *Oceanography*, 23, 108–122.

Smith PJ, McVeagh SM, Mingoia JT, France SC (2004) Mitochondrial DNA sequence variation in deep-sea bamboo coral (Keratoisidinae) species in the southwest and northwest Pacific Ocean. *Marine Biology*, 144, 253–261.

Smith P, McVeagh S, Won Y, Vrijenhoek R. (2004) Genetic heterogeneity among New Zealand species of hydrothermal vent mussels (Mytilidae: *Bathymodiolus*). Mar Biol 144(3): 537-545.

Thompson RM (1991) Gazetteer of seafloor features in the New Zealand region. New Zealand Oceanographic Institute Miscellaneous Publication 104. 64 p.

Van Dover CL, Smith CR, Ardron J *et al.* (2012) Designating networks of chemosynthetic ecosystem reserves in the deep sea. *Marine Policy*, 36, 378–381.

Vrijenhoek RC (2010) Genetic diversity and connectivity of deep-sea hydrothermal vent metapopulations. *Molecular Ecology*, 19, 4391–4411.

Zanol J, Halanych KM, Struck TH, Fauchald K (2010) Phylogeny of the bristle worm family Eunicidae (Eunicida, Annelida) and the phylogenetic utility of noncongruent 16S, COI and 18S in combined analyses. *Molecular Phylogenetics and Evolution*, 55, 660–676.

CHAPTER 3

Multiple Spatially Distinct Introductions and not Range Expansion May Explain Colonization History in an Invasive Marine Species

ABSTRACT

Biological invasions are often characterized by a phase of post-establishment expansion in which the invading species increases its range through colonization of new geographic area. Expansions are predicted to result in specific genetic signatures, most notably decreased genetic diversity with distance from the origin of the expansion, which is often the point of introduction for invasive species. The caridean shrimp, *Palaemon macrodactylus*, is an invasive species in many regions of the globe. *P. macrodactylus* has most recently invaded the US Atlantic coast, with the first report of the species in New York in 2001. This study uses both mitochondrial *cytochrome oxidase I* (COI) sequence data as well as data for 1,598 single nucleotide polymorphisms (SNPs) generated through restriction enzyme associated DNA sequencing (RAD-seq) to test two potential scenarios describing the expansion of *P. macrodactylus* north of New York: the first focuses on range expansion facilitated by ocean currents, physical environment, and life history; the second involves multiple introductions of the shrimp in different estuarine ports. In testing these two scenarios, patterns of population genomic diversity as well as population structure are described. Results do not support a range expansion scenario in which diversity decreases with distance from the point of invasion. Rather, the data suggest a scenario of multiple introductions with diversity increasing with distance from New York, and peaks of mitochondrial diversity in populations collected from New York and the Boston-Plymouth coastline. These results indicate that human-mediated dispersal may be as important—if not more important—than oceanographic and life history considerations during the colonization phases of a marine invasion.

INTRODUCTION

Biological invasions threaten biodiversity in terrestrial, aquatic, and marine ecosystems (Bax *et al.* 2003; Lowry *et al.* 2013; Thomaz *et al.* 2014), often negatively affecting ecosystem services and damaging economies (Funk *et al.* 2014; Walsh *et al.* 2016). In addition to being a serious concern for the conservation of biodiversity, invasions are excellent models for the study of dynamic evolutionary processes (Lee 2002; Rius *et al.* 2014; Barrett 2015). Many invasions present an evolutionary paradox because the invading species is ecologically successful despite a high probability of experiencing reduced genetic diversity due to an initial founder event during introduction, often thought to decrease fitness and adaptive potential (*e.g.,* Tsutsui *et al.* 2000). The initial introduction of a species is only one of four steps characterizing biological invasions: (1) transport, (2) introduction, (3) establishment, and (4) spread (Blackburn *et al.* 2011).

Post-establishment spread is often characterized by range expansion—a process that is expected to lead to further decreases in genetic diversity with increasing distance from the point of invasion due to strong genetic drift caused by repeated founder events through space and time (Excoffier *et al.* 2009). Genetic drift during range expansions can be strong enough to fix otherwise rare alleles near the leading edge of the expansion in a process known as allele surfing (Edmonds *et al.* 2004; Travis *et al.* 2007; Hallatschek & Nelson 2008). However, the genetics of range expansion during invasion can be complicated by a variety of demographic phenomena including multiple introductions, persistent human-mediated transport within the invaded range, or both. Marine examples of this include the invasions of the European green crab *Carcinus maenas* along the US Atlantic coast (Darling *et al.* 2008; 2014), the Asian violet tunicate *Botrylloides violaceus* along the US Pacific coast (Bock *et al.* 2010), and the Pacific bryozoan *Tricellaria inopinata* along the US Atlantic coast (Johnson & Woollacott 2015). These complications imposed on an otherwise seemingly clear narrative of post-establishment spread often limit our ability to discern and predict the population genetic patterns of an invasive species or the course of an invasion.

A recent invader along the US northern Atlantic coast is the caridean shrimp *Palaemon macrodactylus* (Rathbun, 1902), native to China, Japan, and Korea (Ashelby *et al.* 2013). The species has invaded regions across the globe since the mid-twentieth century, including San Francisco Bay (1957) and other parts of the US Pacific coast, Western Europe (1992), Argentina (2000), and most recently the US Atlantic coast (Ashelby *et al.* 2013). *P. macrodactylus*, also known as the Asian prawn or Asian shrimp, was first discovered on the US Atlantic coast in 2001 in the Bronx River near New York City, incidental to ichthyofaunal surveys (Warkentine & Rachlin 2010). *P. macrodactylus* is one of at least five palaemonid shrimp—both native and invasive—in the northeast region of the United States. It was not until 2010, when the discovery of *P. macrodactylus* was announced, that researchers began looking for this specific species elsewhere along the US coastline, by which time (2010) *P. macrodactylus* was found throughout southern New England in Long Island Sound and Narragansett Bay, as well as north of Cape Cod as far as Boston in 2012 (JTC, unpublished data). *Palaemon macrodactylus* was discovered in 2007 in Chesapeake Bay (Fofonoff *et al*., 2016). New England and adjacent coastal surveys specifically designed to assess the distribution of palaemonid shrimp in the summer of 2014 then documented populations of *P. macrodactylus* from central New Jersey north to Newington, New Hampshire, the northernmost documented location to date (Carlton and Weigle, 2015).

Since the discovery of the extent of *P. macrodactylus* invasion in the United States in 2010, only one population genetic study of the species has been undertaken (Lejeusne *et al.* 2014). This global survey of invasive *P. macrodactylus* populations used mitochondrial *COI* data and concluded that the invader had high genetic diversity in all invaded regions globally, indicating limited or no founder events during each invasion. However, their study includes individuals from only one of the many invaded estuaries on the US Atlantic coast, leaving unanswered questions regarding population structuring in the invaded regions of the United States. Other *Palaemon* species have shown sometimes surprising population genetic structure within specific regions (*e.g.*, *Palaemon elegans* in Europe: Reuschel *et al.* 2010), while others follow expectations set by oceanographic current patterns (*e.g.*, *Palaemon floridanus* in the Caribbean Sea: Baeza & Fuentes 2013).

In the absence of systematic surveys of palaemonid shrimp along the U.S. Atlantic coast in the decades surrounding the appearance of *P. macrodactylus* in Europe in 1992 (and thus its ready availability to be potentially transported by ships to North America) and its discovery in 2001 in New York, the precise location and timing of introduction in North America is not known. At least two colonization scenarios are possible. The first, an expansion scenario posits that *P. macrodactylus* was introduced to New York City (2001), spread south to Chesapeake Bay (2007) and north to eastern Long Island Sound (2010), and then to Boston (2012) and New Hampshire (2014), a scenario based on the dates of observations. This expansion scenario generates explicit genetic expectations, as noted above, for the invaded region based on the assumption that this spread represents repeated founder events through space and time leading to decreased diversity along the expansion axis. A second scenario involves multiple introductions, meaning that *P. macrodactylus* owes its appearance in ports and bays such as Boston, New York, and Chesapeake Bay to multiple separate introductions, a phenomenon that generates a different set of genetic expectations, including potentially distinctive, localized genetic structuring and peaks of genetic diversity.

This study seeks to examine which of the above two scenarios may best explain the nature of the establishment and distribution of *Palaemon macrodactylus* on the Atlantic coast of North America, focusing on northern populations. In testing these two possible scenarios, we describe the distribution of genetic diversity in the invaded area between New York and New Hampshire in the context of range expansion expectations. We highlight population genomic structure to examine potential patterns of local isolation and connectivity in the invaded range. We use both mitochondrial *cytochrome oxidase I* sequence data and data from 1,598 single nucleotide polymorphisms (SNPs) generated from restriction-enzyme-associated DNA sequencing (RAD-seq). This is the first use of genome-wide SNP markers in an invasive *Palaemon* species.

METHODS

Sample collection

Palaemon macrodactylus samples were collected from marina dock fouling communities throughout New England (Figure 1) with a hand-held fishing net (45.7 cm diameter ring, 1.0 cm mesh). Sites to the south of Cape Cod included Evers Marina in the Bronx River, New York, NY; Mystic Seaport in Mystic, CT; and Moby Dick Marina in Fairhaven, MA. Sites to the north of Cape Cod included Brewer Marina in Plymouth, MA; University of Massachusetts Boston, in Boston, MA; and Great Harbor Marine in Newington, NH (Table 1). Sites with more established fouling community habitats were often the locations with the greatest abundance of shrimp and were, therefore, targeted for collection. Shrimp distributions in marinas were found to be patchy, leading collection teams involved in the 2014 ShrimpEX surveys (Carlton and Weigle, 2015) to sample large areas of dock surfaces to acquire the required sample numbers. Shrimp were sorted by gross morphology and color at the sampling locations and those likely to be *P. macrodactylus* were preserved in 100% ethanol at the sampling site and kept cool until they were drained of ethanol and stored in a -20°C freezer.

Figure 1. Map of the study area with study sites indicated by white circles.

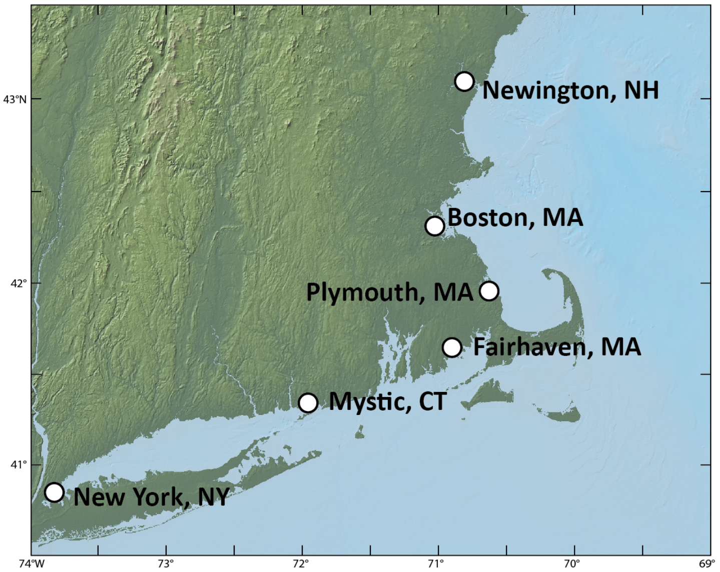


Table 1. Sample site details, latitude and longitude, and through-water distance in kilometers from the New York sampling location.

Location	**Marina Name**	**Latitude (N)**	**Longitude (W)**	**km from NYC**
New York, NY	Evers Marina	40.8442	-73.8131	0
Mystic, CT	Mystic Seaport	41.3638	-71.9644	215.74
Fairhaven, MA	Moby Dick Marina	41.6536	-70.9141	340.09
Plymouth, MA	Brewer Plymouth Marina	41.9565	-70.6596	556.59
Boston, MA	U. Massachusetts, Boston	42.3115	-71.0401	599.19
Newington, NH	Great Bay Marine	43.1160	-70.8357	652.55

P. macrodactylus were identified morphologically under a dissecting microscope using the defining characteristics of a double row of setae on the ventral side of the rostrum and typically three rostral teeth behind the posterior margin of the orbital socket (González-Ortegón & Cuesta 2006). Length and reproductive status (*i.e.*, gravid, not-gravid) were recorded for each specimen. Size distributions of all shrimp collected and of those included in population genetic analysis are presented in Appendix I.

DNA extraction and polymerase chain reaction

Genomic DNA (gDNA) was extracted from a section of abdominal muscle tissue from each shrimp individual using the Omega Insect Extraction Kit (Omega Biotek, Norcross, GA, USA), with a standard protocol including the suggested liquid nitrogen homogenization step. gDNA samples were stored in the kit's elution buffer at 4°C or -20°C until PCR reactions and RAD sequencing. Polymerase chain reactions (PCRs) to amplify *cytochrome oxidase I* (*COI*) were carried out using primers CrustCOIF (5'-TCAACAAATCAYAAAGAYATTGG-3') and DecapCOIR (5'-AATTAAAATRTAWACTTCTGG-3') (Lejeusne *et al.* 2014). The thermocycler temperature profile consisted of 95° denaturing step for 3 minutes, then 30 cycles of 95° for 45 seconds, 48° for 60 seconds, 72° for 60 seconds, followed by a final extension step at 72° for 5 minutes. PCR reactions were purified using a QIAGEN PCR Purification Kit (Qiagen GmbH, Germany) and were sequenced at Eurofins Operon Genomics (Eurofins MWG Operon LLC, Louisville, KY, USA).

During initial optimization of gDNA extraction and amplification protocols, sequencing efforts sometimes produced double peaks (two equally strong sequencing results at one nucleotide position) in the chromatograms for mitochondrial *COI*. These double peaks were replicated across individual samples in both directions of sequencing reads, and occurred reliably at specific nucleotide locations. In nuclear genes, such results would indicate heterozygosities, but for mitochondrial genes, these results are unexpected because only one copy of each mitochondrial gene is expected to be present, except in rare cases of bi-parental inheritance. We concluded that the consistent double peak results were evidence of a second copy of the gene, or a pseudogene (Williams & Knowlton 2001). In response to this result, we developed the protocol described above, which includes a different type of extraction technique (*i.e.,* different from both the original phenol-chloroform extraction attempts and the Chelex extraction used in previous studies) and a reduced number of amplification cycles during PCR. This protocol eliminated the double peaks. Details of the original methods are reported in Appendix II.

Mitochondrial sequencing analysis

Mitochondrial DNA sequences were edited and assembled using *Geneious 8.1.5* (Kearse *et al.* 2012). Consensus sequences were then aligned using the *Geneious* MAFT alignment plug-in with default settings. The 85 haplotype sequences for *P. macrodactylus* published in Lejeusne *et al.* (2014) were downloaded from the NCBI nucleotide database (GenBank Accession Numbers HG792276.1 through HG792360.1, and G792313.1) and aligned to the sequences generated for this study. All sequences were trimmed to 501 basepairs in order to allow for the inclusion of more individuals (final sequence length was shorter than in Lejeusne *et al.* (2014), in which 598 basepairs were used). Sequence trimming did not lead to the exclusion of any haplotypes present in individuals sequenced for this study (a conclusion based on the location of the polymorphisms). However, because the sequences were trimmed, haplotypes *Pm55, Pm56, Pm57, Pm58, Pm59, Pm83, and Pm84* had variable basepairs trimmed from the dataset. Every defining polymorphism of the sequences in this study fell within the trimmed region of the sequencing reads. Additionally, to test for potential effects of possible pseudogene sequencing (mentioned above) and accurately compare to previously published *COI* data (from Lejeusne *et* al, 2014), tests were run excluding nucleotides from analyses that were potentially problematic based on this study's initial methods analysis. While removal of possibly problematic bases where double peaks occurred resulted in a reduction in the number of haplotypes, from 85 haplotypes to 22, any alterations in haplotype calling did not substantially change the conclusions of the present study and all nucleotides are included in the following analyses (please see Appendix II for a summary of these tests and the results).

Summary statistics including nucleotide diversity and pairwise differences for the mitochondrial data, as well as pairwise F_{ST} values were calculated using *Arlequin* v3.5.2.2. A haplotype network was constructed using these results combined with previously published data.

Restriction Enzyme Associated DNA (RAD) sequencing

Genomic DNA samples were normalized to a concentration of 20 ng/µl as measured on a QUBIT 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). Restriction enzyme associated DNA sequencing library preparation using the *SbfI* restriction enzyme (restriction site: 5'-CCTGCAGG-3') was carried out on concentration-normalized gDNA by Floragenex Inc. (Eugene, OR, USA) in identical fashion to several other recent RAD-seq studies (*e.g.,* Reitzel *et al.* 2013; Herrera *et al.* 2015). For library preparation, gDNA was digested with the *SbfI* restriction enzyme, yielding fragments of various lengths. Barcode tags, 10 basepairs in length and specific to individual, as well as an Illumina adaptor, were ligated onto the sticky end of the cut site. Samples were then pooled, sheared, and size selected for optimal Illumina sequencing.

Libraries were then enriched through PCR and sequenced by 96-multiplex in a single lane of an Illumina Hi-Seq 2000 sequencer.

RAD data filtering, SNP calling, and population genomic analyses

Using the *process_radtags* program in *Stacks* v1.35 (Catchen *et al.* 2013), raw Illumina reads were filtered for quality with a minimum phred score of 10 in a sliding window of 15% read length (default settings) and sorted by individual barcode. Reads were truncated to 90 basepairs (bp) to remove barcodes and adaptors but leaving the six basepair restriction site intact. Putative loci were generated using the *denovo_map.pl* pipeline in *Stacks*. We used a *stack-depth parameter* (*-m*) of 3, meaning that three reads were required to generate a stack (*i.e.,* a locus); a *within-individual distance parameter* (*-M*) of 3, allowing for three SNP differences in a read; and a *between-individual distance parameter* (*-n*) of 3. The final size of the locus catalog varied as expected with different values for the *denovo_map.pl* parameters (for a complete evaluation of the sensitivity to different parameters, please see Appendix III).

Population summary statistics including allele frequencies, observed and expected heterozygosities (H_{obs} and H_{exp}), π, and F_{IS}, were calculated by the *populations* program in *Stacks*, using loci found in five of the six populations and in at least 60% of individuals per population using flags *-p* 5, *-r* 0.6. Due to the nature of the sequencing quality and coverage (see Results), the *populations* program was also run for all populations excluding New York, NY, with loci found in all five of the remaining populations, and 60% of individuals (argument flags *-p* 5, *-r* 0.6). These data were used in the principal component analysis (see below). Information on the effects of changing the *-p* and *-r* flags is available in Appendix III. For each RAD-tag, only one SNP was used from the 90 bp sequence using the flag *–write_single_snp* (specifying that if there were two or more SNPs in the sequence, *Stacks* would only use the first). Observed and expected heterozygosity (H_{obs} and H_{exp}) values were also calculated in the R Package PopGenKit (https://cran.r-project.org/web/packages/PopGenKit/index.html) to provide secondary validations of reported values. Additionally, allelic richness (A_{rich}) was calculated using PopGenKit. Genetic diversity summary statistics (H_{obs}, H_{exp}, and A_{rich}) were regressed against distance from the New York collection site using the *stats* package from *Scipy* (https://scipy.org). The least cost distance dispersal trajectories used in these regressions were calculated using the 'gdistance' package in R with a bathymetric constraint from ETOPO1 (van Etten, 2015; R Core Team, 2016).

Three methods were used to describe the genetic structure of *P. macrodactylus* populations. First, the *smartpca* program within *EIGENSOFT* (Price *et al.* 2006) was used to perform a principal component analysis (PCA) of genetic diversity. Custom scripts archived as iPython notebooks were used to convert *Stacks* PLINK output files into *EIGENSOFT* input files, and to visualize the PCA results (https://github.com/ekbors/thesis_scripts). The *smartpca* program within *EIGENSOFT* was run with one iteration of outlier removal ('*numoutlieriter'* = 1) with otherwise default parameters. To evaluate the impact of missing data on population clustering in *EIGENSOFT*, the '*missingmode'* argument was used in certain parameter runs. Missing data appeared to have a significant effect on the population clustering patterns (Appendix IV). Therefore, the results from the missing test runs led us to use the *populations* output excluding New York and analyze loci in all of the remaining five populations. The decision to exclude New York was based on the fact that the fewest loci were retained for that site, and therefore removing it resulted in less of a reduction in loci used when the new constraint that loci must be in all populations was introduced (the *-p* flag in *Stacks populations*). Second, *fastSTRUCTURE* (Pritchard *et al.* 2000; Hubisz *et al.* 2009; Raj *et al.* 2013) was run with the number of genetic lineages (the value of *k*) set to values between one and ten to assess genetic structure through a hierarchical analysis, and the program *chooseK.py* was run to select the value of *k* most consistent with the program's Bayesian structure model. Third, F_{ST} values were calculated by the *populations* program in *Stacks* using a p-value cutoff of 0.05 with a Bonferroni correction implemented by the program (using the *--fst_correction 'bonferroni_gen'* argument).

In addition to the described approaches of regressing genetic diversity measurements with distance from New York in order to capture potential range expansion signals in diversity summary statistics, a range-expansion specific analysis was also implemented based on asymmetries in allele frequency data which may indicate the directionality of expansion (Peter & Slatkin 2013). Using an R package developed by Peter & Slatkin (2013), we calculated *psi*, or the "directionality index" in order to quantify the relationships between allele frequencies and potential direction and strength of expansion.

RESULTS

Geographic distribution of mitochondrial haplotypes

A 501 basepair region of *cytochrome oxidase I* (COI) was sequenced for a total of 106 individual *P. macrodactylus* (19 from New York, NY; 22 from Mystic, CT; 10 from Fairhaven, MA; 15 from Plymouth, MA; 19 from Boston, MA; 21 from Newington, NH) (Table 2). In the six populations of *P. macrodactylus* sampled in this study, six haplotypes were identified. For ease of comparison and consistency, the haplotype names used in previously published results are used in this study (Lejeusne *et al.* 2014).

Five of the six haplotypes were previously reported in Lejeusne *et al*. (2014) and one haplotype, here named *PmU86*, was new and unique to New York, NY (n = 1). This is in contrast to previous reports of 11 haplotypes in Mystic, CT, alone, seven of which were reported to be unique (Lejeusne *et al.* 2014). Results of the tests run to investigate what the impact of pseudogene sequencing would have been these results as well as on previously published work, is presented in Appendix II. In short, tests to remove basepairs at locations where double peaks were observed reduced the number of haplotypes previously published from 85 to 22 and the number of loci in this study from 6 to 5.

In the present study, certain haplotypes were only present in one or two populations while others were present in all (Figure 2). The most common haplotypes were *Pm18* (n = 70) and *Pm3* (n = 22). *Pm18* and *Pm3* are also the most common haplotypes globally (Lejeusne *et al.* 2014). *Pm27*, a haplotype previously only reported in Yamaguchi, Japan, in the native range, was located in Boston and Plymouth, MA. *Pm1* was also only previously reported in the native range, but in this study, it was found in Fairhaven and Boston, MA. *Pm67*, a haplotype previously described only in the invaded population in the US Northeast, was observed in New York, NY, but not in Mystic, CT, as previously reported (a possible result of sampling effort or an unobserved fluctuation in population size and genetic composition). Individuals sampled from Fairhaven, MA, and Newington, NH, were all one of two main haplotypes, *Pm18* and *Pm3*. The two major haplotypes had only one nucleotide difference between them, and all other haplotypes only differed by one nucleotide from either of the major haplotypes (Figure 3). The one newly reported haplotype sequence in this study was uploaded to GenBank, and is referred to as *PmU86* throughout the rest of this study. As noted above, all other sequenced haplotypes from this dataset were present in data previously reported by Lejeusne *et al*. (2014). Haplotype diversity peaked in New York and Boston while nucleotide diversity and pairwise differences were highest in Plymouth (Table 2).

There were two samples from Mystic, CT, for which ambiguous bases were present across different sequencing attempts of amplified PCR products. The genomic location of this disagreement in the *COI* sequences was one for which there was no variability reported in previous studies or in other samples in this study. Therefore, for these two samples, we considered the most parsimonious result to be that the sequences that were consistent with all other samples were correct and that the sequences that would indicate a new polymorphism were, in fact, errors. However, the ambiguity is noteworthy, especially in light of the sequencing variation mentioned in the Methods (above) and described further in Appendix II.

Figure 2. Map of study locations with proportion of sampled individuals with each haplotype.

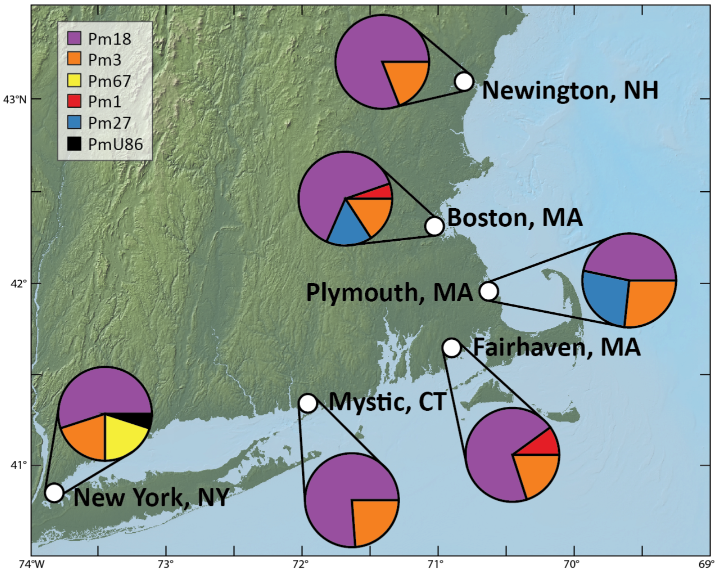
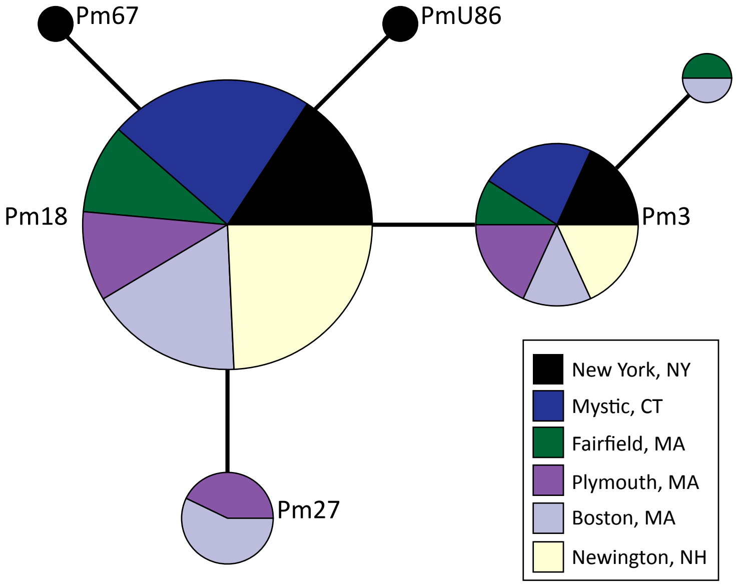


Table 2. *mtCOI* diversity statistics for *Palaemon macrodactylus* at the study sites (*n* = number of individuals, *h*= haplotype diversity).

Location	*n*	*h*	Nucleotide diversity	Number pairwise differences
New York, NY	19	4	0.0015	0.7370
Mystic, CT	21	2	0.0008	0.3810
Fairhaven, MA	10	3	0.0013	0.6667
Plymouth, MA	15	3	0.0017	0.8381
Boston, MA	19	4	0.0015	0.7368
Newington, NH	21	2	0.0007	0.3680

Figure 3. Haplotype network for *cytochrome oxidase I* sequenced for *Palaemon macrodactylus*. The area of each circle reflects the relative proportion of individuals with each haplotype, with *PmU86* and *Pm67* representing one shrimp each.



RAD-seq efficacy and identification of loci

Illumina sequencing of RAD libraries yielded 122,722,404 raw reads. After processing raw reads with the *Stacks* program *process_radtags*, 13.43% of reads were removed due to a missing or ambiguous barcode, 7.16% of reads were filtered due to an ambiguous restriction site and 15.32 % of reads were filtered out due to low sequencing quality, leaving 64.10% of reads retained for further processing. The percent removed due to low sequencing quality can be considered high (especially when compared to Chapters 4 and 5), and was driven mostly by poor quality in a number of sequencing tiles in the middle of the sequencing reads. The number of reads discarded due to poor quality or ambiguous data varied slightly by individual sample (Appendix III).

The final catalog used in this study contained 1,598 loci. The average depth of sequencing coverage per locus across all individuals was 21.43 with an average standard deviation of sequencing depth of 224.21, which is higher than in some other RAD-seq studies (Chapter 4). For the *populations* run including five of the six populations and 60% of individuals per population, the number of loci included per population varied greatly (from 969 for New York to1,598 in Newington, NH), with New York having the fewest loci (Table 3 and Appendix III). For the *populations* run in which New York was excluded and loci were required to be in all five of the remaining populations (which was required to mitigate the effect of missing data in the New York population), 1,092 loci were retained.

Population genomic diversity patterns throughout the invaded range

Genetic diversity statistics either remained consistent or increased with distance from New York (Figure 4). H_{obs} increased with distance from New York (Figure 4A). H_{exp} and A_{rich} had consistent values throughout the sampled region with the exception of small increases in A_{rich} in Boston and Plymouth, MA (Figure 4B, C).

Figure 4. (A) Observed Heterozygosity, (B) Expected Heterozygosity, and (C) Allelic Richness plotted against distance between population locations. For (A) Observed Heterozygosity, $R^2 = 0.822$ and the p-value = 0.013. The regressions for other summary statistics were not significant.

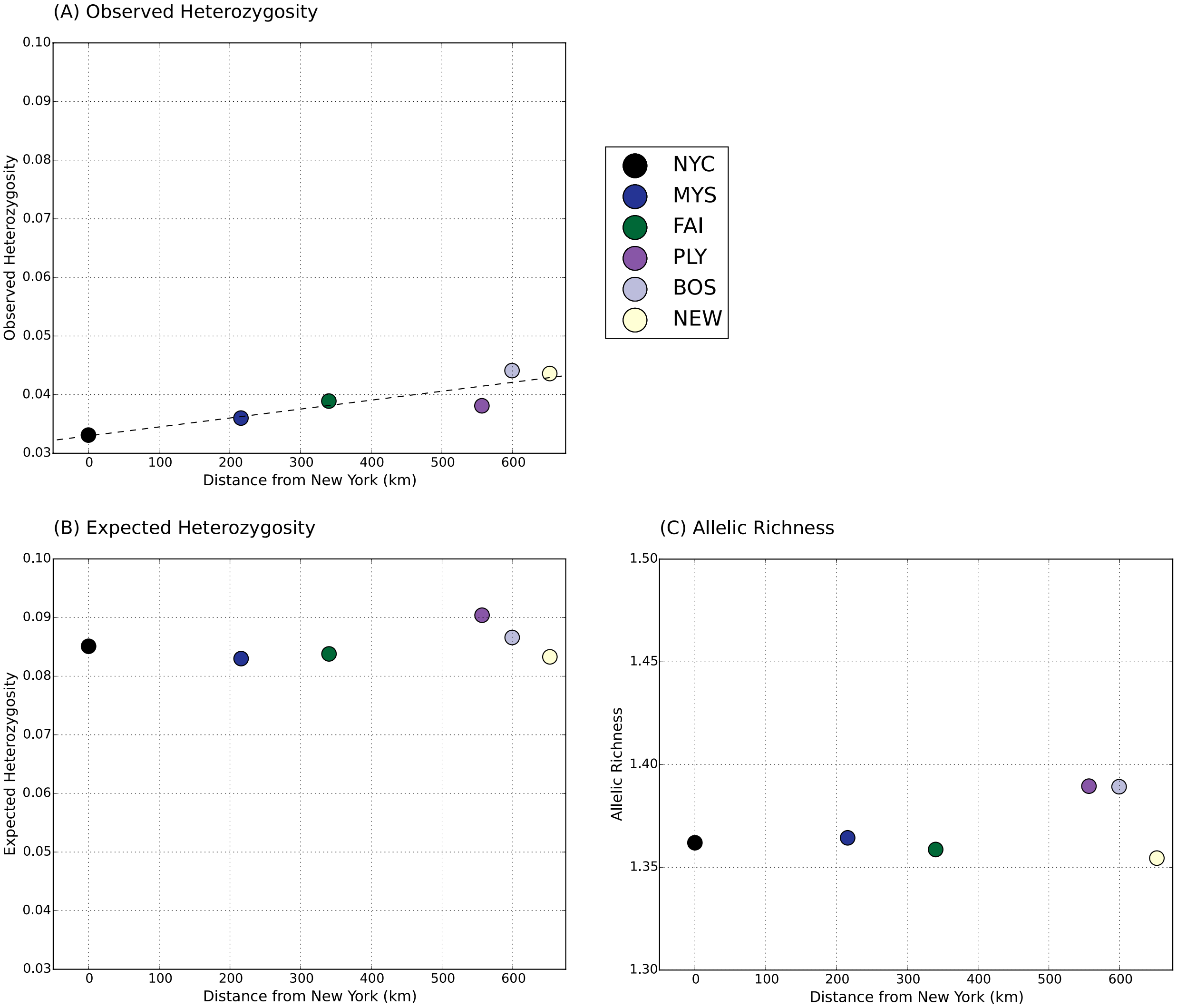
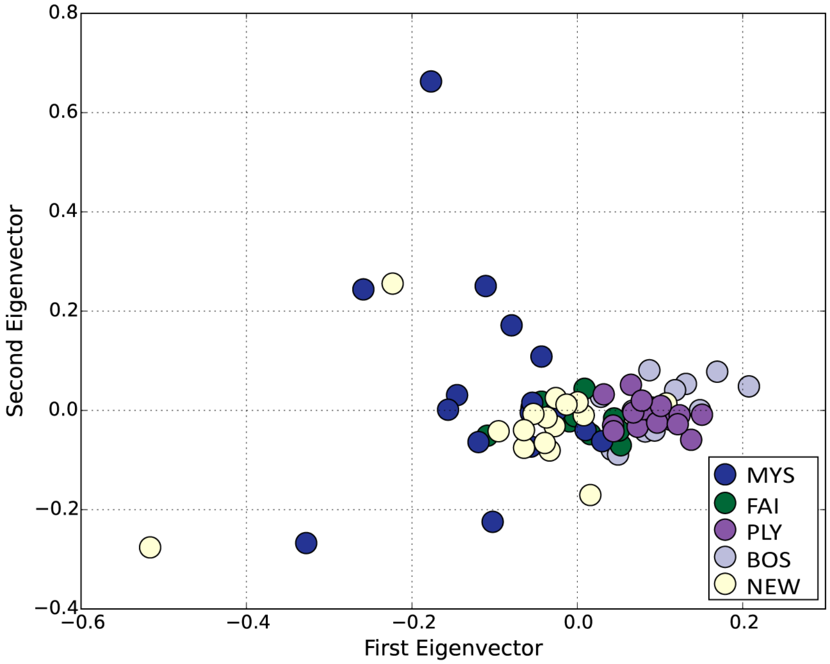


Table 3. Average *Stacks* summary statistics, all shrimp populations. For *populations* run of *-p* 5, *-r* 0.6 including all populations in the initial filter. "N (total)" is the number of individuals sequenced while "n (avg)" is the average number of individuals used by Stacks for the loci included in the analysis.

Location	**N (total)**	**n (avg)**	**# loci**	**P**	**H_{obs}**	**H_{exp}**	**Pi**	**F_{IS}**
New York, NY	17	12.1682	969	0.9451	0.0331	0.0851	0.0888	0.2181
Mystic, CT	17	12.6765	1459	0.9472	0.036	0.083	0.0865	0.2103
Fairhaven, MA	13	9.3401	1479	0.9448	0.0389	0.0838	0.0887	0.1656
Plymouth, MA	16	11.5302	1422	0.9416	0.0381	0.0904	0.0947	0.2066
Boston, MA	14	10.8371	1541	0.9443	0.0441	0.0866	0.091	0.1775
Newington, NH	17	13.1729	1544	0.9461	0.0436	0.0833	0.0866	0.1685

In the PCA for which NYC was omitted and loci were required to be in all populations (in order to reduce the effects caused by missing data, of which the New York samples were significant contributors), sampling locations demonstrated some clustering (Figure 5) with 11.6% of the variation described by the first eigenvector, 11% described by the second, and 10.4% the third (10 eigenvectors were calculated by *EIGENSOFT* indicating that the diversity was almost spread evenly across all 10). In addition to structuring highlighted by PCA, the Bayesian probability program *fastSTRUCTURE* also indicated some regional population genomic structuring. The *chooseK.py* program in *fastSTRUCTURE* identified a *k* value of 5 (indicating the existence of 5 genetic lineages) as being both the value that maximizes likelihood in the model and explains the structure, and the populations were clearly differentiated by genetic lineage (Appendix V). This initial *fastSTRUCTURE* analysis was run on all of the populations using the *populations* output in which loci were required to be in 5 of the 6 populations and 60% of individuals. Population genetic structure analyses were also run in *fastSTRUCTURE* using the other datasets generated from other runs of *populations* in *Stacks.* When requiring loci to be in all populations (*-p* 6), chooseK.py identified a *k* value of 2 to maximize likelihood and a value of 5 to explain structure. That analysis did not yield geographic patterns in the data. This could be because missing data are excluded, and also due to a loss of power in the dataset by using too few loci. In the analysis excluding New York and requiring loci to be in all other individuals, the value of *k* identified that maximizes model likelihood is 2 and the value that explained structure was 6. Again, no geographic structure was evident in the visualized data. Due to the variation in structure analysis results across different *populations* runs, it is not possible to draw a definitive conclusion from the *fastSTRUCTURE* results (Appendix V).

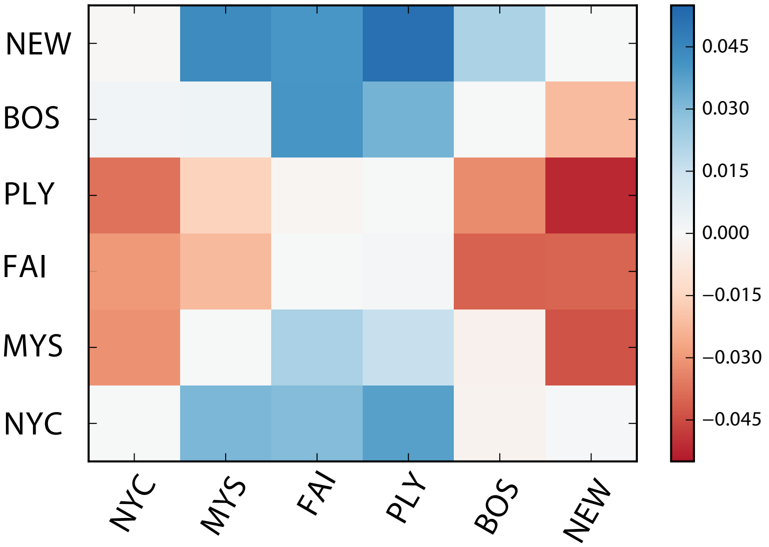
Figure 5. Omitting New York, PCA of select sites made with *EIGENSOFT* for loci included in the remaining five populations and 60% of the individuals in each (MYS = Mystic, FAI = Fairhaven, BOS = Boston, NEW = Newington, PLY = Plymouth).



Despite the apparent population structuring detected by *EIGENSOFT* and *fastSTRUCTURE*, the F_{ST} values—when using a Bonferroni correction—were almost all statistically no different from zero. This could indicate that structuring is based more on regional variation in the distribution of diversity, rather than driven by isolation and inbreeding, as F_{ST} often measures. The only values of F_{ST} that were different for zero were the following pairwise values: Fairfield-to-Mystic, 7.7×10^{-4}; Newington-to-Mystic, 4.6×10^{-4}; Newington-to-Plymouth, 1.4×10^{-3}; Newington-to-Boston, 4.8×10^{-4}.

The values of *psi,* or the directionality index ranged from -0.0518 to 0.0375 on one axis. The directionality index revealed that two locations were almost equally likely to be near the "origin" of expansion: New York and Boston (Figure 6). This indicates that both could be origins of expansion and that other locations may be receiving individuals from those areas. Also notable is the fact that Boston and Newington have *psi* values with the opposite signs to the others, indicating that they are perhaps in an opposite direction of expansion than the other populations.

Figure 6. A heatmap of *psi* values for all 6 populations of *P. macrodactylus*. These values indicate the "directionality index," a statistic meant to capture asymmetries in allele frequencies indicating directional expansion. As is seen in the heatmap, both Boston and New York have values at or very near zero, with Boston's value slightly positive when others are negative. Note that the values are symmetrical above and below the diagonal axis, except for sign.



Discussion

Potential support for multiple introductions rather than a spatial expansion

Results in this study do not support a range expansion scenario in which invading populations of *Palaemon macrodactylus* expanded northward along the coast from New York City, the first-reported location for the species in the United States in 2001. Instead, results suggest evidence for an alternative scenario in which multiple introductions account for the spatial spread of *P. macrodactylus* in the estuarine waters of the northern US Atlantic coast, with a second introduction possible in the Boston area. This conclusion is based on the existence of two haplotypes previously only described in the native range of Asia (Lejeusne *et al.* 2014) in Boston, Plymouth, and Fairhaven, MA. The peaks in nucleotide diversity in the *COI* data in Boston-Plymouth, and New York also add to the evidence suggesting multiple introductions in those two areas.

Further support for the scenario in which there is at least one additional introduction in the north comes from the documented increase in observed heterozygosity across the 1,598 loci with increasing distance from New York. This trend is contrary to the expectation during a range expansion, in which diversity would be expected to decrease with increasing distance from the point of introduction. These results may demonstrate how the expectations of range expansion can be overwhelmed by the specific context of the invasion—in this case, what may appear to be spatial range expansion may have actually been driven by multiple introductions, changing the distribution of genetic diversity in invasive populations. Additionally, the calculations of the directionality index, or *psi*, identified Boston/Newington and New York as potential locations of the origin of expansion, pointing to the existence of two locations of distinct introductions before spread.

Analysis of 1,092 loci generated with RAD-seq and used for PCA (excluding the samples from New York), revealed slight population genetic structuring by location in the invaded range. F_{ST} values for the RAD-seq dataset, however, were extremely low and/or statistically indistinguishable from zero. F statistics are used to quantify the level of inbreeding that results from isolation of subpopulations in a structured metapopulation (Hartl & Clark 2007). From these measurements, population structuring is inferred. In cases of range expansion—a clear violation of Hardy-Weinberg equilibrium—one must consider that this framework for evaluating structure may not be appropriate. While the population structure detected using PCA and *fastSTRUCTURE* may be driven by isolation of the individual estuaries in which shrimp were collected, it could also be driven by differential introduction patterns across the invaded range. Specifically, shrimp of different genetic identity might exist in Boston and Plymouth than those in Newington and Mystic because invaders of that genotype may have been introduced more recently in some places and not yet introduced into others. This may not necessarily mean that Newington and Mystic are isolated populations. Rather, it may indicate that the population genetic structure is driven by the unfolding of the invasions in a non-linear manner. Similarly, Mystic and Newington may not be genetically connected oceanographically, but instead, may be experiencing similar impacts of post introduction expansion patterns. Perhaps the two sites were both colonized by the first introduction and underwent bottlenecks *or* perhaps Newington was colonized from Boston's invasion and Mystic was colonized from New York's, and thus they both have decreased diversity and therefore apparently cluster together in measures of structure. With time, population homogenization may actually weaken the structure signal, or isolation may strengthen it.

Population genetic structure could also be driven by intracoastal human-mediated dispersal after introduction. Recreational boats, commercial fishing vessels, ferries, cargo ships, and touring yachts could all participate in the reshuffling of shrimp around New England. Many such vessels could transport shrimp in ballast water along the coast, or, for example, in water of tires or other structures used as hull bumpers (Lejeusne *et al.* 2014). This could explain, for example, why some of the mitochondrial haplotypes in Boston were found in Fairhaven and not Plymouth while others were found in Plymouth and not Fairhaven. In many invasions, multiple vectors are responsible for dispersal (Richardson *et al.* 2016). Intracoastal dispersal by humans has been shown to be important in other marine invasions, including that of the tunicate *Styela clava* in the northeastern Pacific Ocean (Darling *et al.* 2012). Patterns of haplotype distribution could also be due to insufficient sampling of diversity driven by sampling numbers, or sampling of segments of the population that are distributed in patches. For example, if shrimp were only collected from one patch within a marina or estuary from one sampling location but those from another sampling location were collected from five different patches within a marina or river basin, then the second set of shrimp might be expected to be more diverse depending on the nature of the patchiness.

Oceanographic and life history contexts of the Palaemon macrodactylus invasion

Oceanographic currents are major drivers of larval dispersal in the oceans, with oceanographic patterns sometimes dictating population genetic structure patterns (White *et al.* 2010). For example, the coastal worm *Clymenella torquata*, has a discontinuity in population genetic structure just to the south of Cape Cod—a location hypothesized to be a phylogeographic boundary—that was attributed to converging water masses and not by the physical barrier of the Cape itself (Jennings *et al.* 2008). However, physical oceanographic studies have revealed that the coastal morphology and current flow patterns in the region are such that the estuaries of Long Island Sound, Boston Harbor, and Cape Cod Bay are all largely tidally forced, with wind, river inputs, and tides driving patterns in Long Island Sound (Whitney *et al.* 2016), and some more persistent but still variable conditions prevailing in Massachusetts waters north of Cape Cod (Jiang *et al.* 2007). In Cape Cod Bay, seasonal variation has been shown to change the retention patterns of planktonic crustaceans, such as the copepod *Calanus finmarchicus* (Jiang *et al.* 2007). Considering these daily and seasonal fluctuations in oceanographic patterns, we expect that instead of persistent oceanographic currents, tidal forcing may play a more important role in the dispersal of larval shrimp in and out of estuaries, making predicting the patterns difficult because of the stochasticity of such processes.

Reproductive strategies, dispersal capabilities, and other life history traits play a crucial role in determining population genetic connectivity and geographic spread in many marine species (Selkoe *et al.* 2016). In its native range, *P. macrodactylus* reproduces seasonally from April to October, with two cohorts produced per season throughout their two year lifespan (Omori & Chida 1988). The shrimp are often found in brackish water in coastal estuaries but there is some evidence that developing larvae and gravid females may migrate to higher salinity waters (Vázquez *et al.* 2015). In the invaded region of France, for example, evidence suggests that different life stages are spatially segregated within estuaries suggesting that migration occurs (Béguer *et al.* 2011). Further evidence of offshore mixing of larvae has been reported in the western Mediterranean Sea (Torres *et al.* 2012). These processes could lead to high levels of mixing in *P. macrodactylus*.

Environments with high levels of oceanographic mixing and species with longer-distance larval dispersal often drive populations towards lower levels of population structure and, notably, towards genetic patterns that do not correlate with Euclidian distance (Cowen *et al.* 2007; White *et al.* 2010). However, *the P. macrodactylus* populations exhibit slight genetic structure, so despite the potential combination of human-mediated dispersal and oceanographic mixing, we acknowledge other processes may be responsible for observed patterns, or that simply not enough time has passed for mixing to diminish the signals of multiple introductions.

Consistency with other invasions on the US Atlantic coast

Multiple introductions are common in marine invasions (Rius *et al.* 2014). One of the most well-known crustacean along the US Atlantic coast, the European green crab *Carcinus maenas*, has been introduced multiple times (Roman 2006; Darling *et al.* 2008; 2014). Green crabs were discovered on the Atlantic coast of the United States in 1817, likely through transport in ship fouling assemblages or in solid ballast of cargo or transport vessels (Carlton & Cohen 2003; Roman 2006). The expansion of *C. maenas* to the Gulf of Maine took about half a century, eventually stalling near the Canadian border. Then, in the 1980s, *C. maenas* was recorded in northern Nova Scotia. The invaders in the northern part of Nova Scotia were not, however, individuals expanding their invasive range from further south (as might be suspected due to climate change) but proved to be new arrivals from Europe (Roman 2006; Darling *et al.* 2008). Without genetic data, this second introduction of more diverse European green crabs would not have been distinguishable from a northward expansion—similar to what is reported in this study. Following this secondary invasion, northern crabs began spreading southward (as evidenced by haplotype data), contrary to the previous direction of invasion but following coastal currents. Over the course of just five years—between 2002 and 2007—microsatellite genetic markers revealed southward movement and genetic mixing of the new invasive lineages (Darling *et al.* 2014).

In recent years, the Asian shore crab, *Hemigrapsus sanguineus*, another invasive crustacean, surpassed *C. maenas* as the most abundant intertidal crab in the US Atlantic coast (Lord & Williams 2016). Introduced in ballast water, *H. sanguineus* was first collected in 1988 in Cape May County, Delaware, and now occupies a range from South Carolina to Maine (Epifanio 2013). Two genetic studies of *H. sanguineus* exist: one in the native range (Yoon *et al.* 2011) and one in the invaded region of the US Atlantic coast (Lord & Williams 2016). Both studies use the mitochondrial *COI* gene. Lord and Williams (2016) documented an increase in *H. sanguineus* density of up to a factor of almost 70 times between 2005 and 2015, showing population growth as well as expansion. Northern populations, however, did not increase in density as substantially as those closer to the center of the invaded range. Contrary to the expectations for an introduction followed by a range expansion, *COI* data for *H. sanguineus* did not show any specific clines or discontinuity in the population genetics of the invaded range. Although there were other haplotypes present, the populations along the Northeast were, like those of *P. macrodactylus*, dominated by one haplotype.

Another invader undergoing range expansion along the northern US Atlantic coast is the Asian violet tunicate, *Botrylloides violaceus*. The violet tunicate is invasive on both coasts of the US, having first invaded the US Pacific Coast in the mid 20th century and the US Atlantic Coast after that, likely in the late 1970s (Bock *et al.* 2010). The spread of the tunicate reportedly differed on each coast. On the west coast, genetic data point towards punctuated, spatially discontinuous dispersal, likely human-mediated, or to potentially multiple introductions. On the east coast, however, the tunicate followed an isolation-by-distance invasion pattern with a gradient of genetic diversity in microsatellite data from south to north (Bock *et al.* 2010). This juxtaposition of the same species invading two different coasts with two apparently different expansion patterns—one driven by human-mediated dispersal and one driven by oceanography and life history—highlights how each invasion can be different depending on dispersal vectors and context. The US Atlantic coast introduction and genetic patterns match the expectations of range expansion.

Amending the paradigm for genetic structure in invasive species using a temporal perspective

When a species is newly introduced—like the shrimp *Palaemon macrodactylus*—the ways in which introduction patterns distribute genetic diversity could be most important for genetic structuring. The prevalence of multiple introductions and their potential ability to alter the evolutionary trajectory of an invasive species by dramatically changing the nature and distribution of genetic diversity highlights the need to amend the conception of classical marine population genetics as being driven primarily by physical oceanographic drivers and life history traits. A new framework should include the dispersal by humans as a driver of population structure, the more so since anthropogenic dispersal of marine species has been in play for centuries if not millennia (Carlton, 2009). This is obvious in the case of a secondary invasion but can also occur within a species range. For example, inside what is considered to be the native range of *Palaemon elegans*—including the Mediterranean, Black, and Baltic Seas—population structure and a recent apparent expansion may be driven by human-mediated dispersal and not oceanographic or biological factors (Reuschel *et al.* 2010). Additionally, at different times in the history of a metapopulation, different processes are likely to be more important to the shaping of genetic diversity than others.

In the case of *Carcinus maenas* as well as *Palaemon macrodactylus*, comparing molecular data from multiple time points is crucial for detecting multiple invasions. In order to test the hypotheses generated by our understanding of dynamic population processes, repeated genetic assessment of invasive populations (along with knowledge of the quality of baseline survey data) improves our understanding of how an observed pattern that might appear, at a coarse-grained level, to be standard range expansion, may actually be a second introduction (as was observed in *P. macrodactylus*). The previously published data used in comparison to results of this study are based on samples collected at just one site, making definite inferences of temporal processes difficult. Broader coverage of invaded areas may facilitate more concrete conclusions. In the case of *P. macrodactylus*, artifacts of reporting and searching for otherwise largely unstudied crustaceans inadvertently led to an assumption of sequential spread along the coastline. For these reasons, genetic monitoring, or at least periodic genetic studies throughout an invasion will prove highly valuable. While the data presented here suggest that a second introduction is possible, and the previous information regarding haplotype distributions corroborates this explanation of the data, it is still not possible to say exactly when this introduction took place. It is also not possible to determine how long the current population structure will persist as the invasion progresses or if *P. macrodactylus* will appear to expand northward and southward, or be subjected to yet more new introductions from overseas in the future.

Conclusions

Similar to other invasive species, *Palaemon macrodactylus* has likely been introduced multiple times in the invaded range along the US northern Atlantic coast, as is reflected in patterns of genetic diversity. This study represents the first use of genome-wide population genomic markers generated by RAD-sequencing in an invasive *Palaemon* shrimp. Data presented here indicate that invasion context is crucial to making predictions about genetic diversity. In addition to oceanographic dispersal mechanisms and life histories, human-mediated dispersal may play an important role in shaping the diversity of marine species. Furthermore, the age of invasive metapopulations may partially determine which force plays the most important role in driving population structure. At the current time, our understanding of the population dynamics of invasion are not generalizable across species. It is only through more intense study, continued research through time, and comparisons among multiple invasive species, that our understanding of marine invasions will improve.

ACKNOWLEDGEMENTS

We would like to acknowledge the exuberant field assistance provided by the many members of the "ShrimpEX2014" research team, based out of the Williams College - Mystic Seaport Maritime Studies Program in Mystic, CT. Support was provided for the ShrimpEX2014 research program by the Massachusetts Office of Coastal Zone Management and the Northeast Sea Grant Consortium. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374. Funding for the RAD-sequencing was provided in part by the Woods Hole Sea Grant Program, and the Woods Hole Oceanographic Institution Biology Department. Additional support for this research was provided by the National Science Foundation (OCE-1131620 to TMS) and the James Education Fund for Ocean Exploration within the Ocean Exploration Institute of the Woods Hole Oceanographic Institution.

Literature Cited

Ashelby CW, Johnson ML, De Grave S (2013) The global invader *Palaemon macrodactylus* (Decapoda, Palaemonidae): an interrogation of records and a synthesis of data. *Crustaceana*, **86**, 594–624.

Baeza JA, Fuentes MS (2013) Phylogeography of the shrimp *Palaemon floridanus* (Crustacea: Caridea: Palaemonidae): a partial test of meta-population genetic structure in the wider Caribbean. *Marine Ecology*, **34**, 381–393.

Barrett SCH (2015) Foundations of invasion genetics: the Baker and Stebbins legacy. *Molecular Ecology*, **24**, 1927–1941.

Bax N, Williamson A, Aguero M, Gonzalez E, Geeves W (2003) Marine invasive alien species: a threat to global biodiversity. *Marine Policy*, **27**, 313–323.

Béguer M, Bergé J, Martin J *et al.* (2011) Presence of *Palaemon macrodactylus* in a European estuary: evidence for a successful invasion of the Gironde (SW France). *Aquatic Invasions*, **6**, 301–318.

Blackburn TM, Pyšek P, Bacher S *et al.* (2011) A proposed unified framework for biological invasions. *Trends in Ecology & Evolution*, **26**, 333–339.

Bock DG, Zhan A, Lejeusne C, MacIsaac HJ, Cristescu ME (2010) Looking at both sides of the invasion: patterns of colonization in the violet tunicate *Botrylloides violaceus*. *Molecular Ecology*, **20**, 503–516.

Carlton JT (2009) Deep invasion ecology and the assembly of communities in historical time, pp. 13-56, in: G Rilov and JA Crooks, editors, Biological Invasions in Marine Ecosystems. Springer-Verlag, Berlin.

Carlton JT, Cohen AN (2003) Episodic Global Dispersal in Shallow Water Marine Organisms: The Case History of the European Shore Crabs Carcinus maenas and C. aestuarii. *Journal of Biogeography*, **30**, 1809–1820.

Carlton JT, Weigle S (2015) Final Report. Shrimp Expedition 2014 (*ShrimpEx14*): A Rapid Assessment Survey of Non-Indigenous Marine and Estuarine Shrimp Species in the Northeastern United States, 23 pp. On file at the Williams College - Mystic Seaport Maritime Studies Program, Mystic CT.

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

Cowen RK, Gawarkiewic G, Pineda J, Thorrold SR, Werner FE (2007) Population Connectivity in Marine Systems An Overview. *Oceanography*, **20**, 14–21.

Darling JA, Bagley MJ, Roman J, Tepolt CK, Geller JB (2008) Genetic patterns across multiple introductions of the globally invasive crab genus *Carcinus*. *Molecular Ecology*, **17**, 4992–5007.

Darling JA, Herborg L-M, Davidson IC (2012) Intracoastal shipping drives patterns of regional population expansion by an invasive marine invertebrate. *Ecology and Evolution*, **2**, 2557–2566.

Darling JA, Tsai YHE, Blakeslee AMH, Roman J (2014) Are genes faster than crabs? Mitochondrial introgression exceeds larval dispersal during population expansion of the invasive crab *Carcinus maenas*. *Royal Society Open Science*, **1**, 140202–140202.

Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences*, **101**, 975–979.

Epifanio CE (2013) Invasion biology of the Asian shore crab *Hemigrapsus sanguineus*: A review. *Journal of Experimental Marine Biology and Ecology*, **441**, 33–49.

Excoffier L, Foll M, Petit RJ (2009) Genetic Consequences of Range Expansions. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 481–501.

Fofonoff PW, Ruiz GM, Steves B, Carlton JT. (2016) *Palaemon macrodactylus,* in California Non-native Estuarine and Marine Organisms (Cal-NEMO) System. http://invasions.si.edu/nemesis/. Accessed December 13, 2016.

Funk JL, Matzek V, Bernhardt M, Johnson D (2014) Broadening the case for invasive species management to include impacts on ecosystem services. *BioScience*, **64**, 58–63.

González-Ortegón E, Cuesta JA (2006) An illustrated key to species of Palaemon and Palaemonetes (Crustacea: Decapoda: Caridea) from European waters, including the alien species Palaemon …. *Biological Association of the United Kingdom*.

Hallatschek O, Nelson DR (2008) Gene surfing in expanding populations. *Theoretical Population Biology*, **73**, 158–170.

Hartl DL, Clark AG (2007) *Principles of Population Genetics*. Sinauer Associates Incorporated.

Herrera S, Watanabe H, Shank TM (2015) Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Molecular Ecology*, **24**, 673–689.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

Jennings RM, Shank TM, Mullineaux LS, Halanych KM (2008) Assessment of the Cape Cod

Phylogeographic Break Using the Bamboo Worm Clymenella torquata Reveals the Role of Regional Water Masses in Dispersal. *Journal of Heredity*, **100**, 86–96.

Jiang M, Brown MW, Turner JT *et al.* (2007) Springtime transport and retention of *Calanus finmarchicus* in Massachusetts and Cape Cod Bays, USA, and implications for right whale foraging. *Marine Ecology Progress Series*, **349**, 183–197.

Johnson C, Woollacott R (2015) Analyses with newly developed microsatellite markers elucidate the spread dynamics of *Tricellaria inopinata* d'Hondt and Occhipinti-Ambrogi, 1985 - a recently established bryozoan along the New England seashore. *Aquatic Invasions*, **10**, 135–145.

Kearse M, Moir R, Wilson A *et al.* (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Lee CE (2002) Evolutionary genetics of invasive species. *Trends in Ecology & Evolution*, **17**, 386–391.

Lejeusne C, Saunier A, Petit N *et al.* (2014) High genetic diversity and absence of founder effects in a worldwide aquatic invader. *Scientific Reports*, **4**, 1–9.

Lord JP, Williams LM (2016) Increase in density of genetically diverse invasive Asian shore crab (*Hemigrapsus sanguineus*) populations in the Gulf of Maine. *Biological Invasions*, 1–16.

Lowry E, Rollinson EJ, Laybourn AJ *et al.* (2013) Biological invasions: a field synopsis, systematic review, and database of the literature. *Ecology and Evolution*, **3**, 182–196.

Omori M, Chida Y (1988) Life history of a caridean shrimp *Palaemon macrodactylus* with special reference to the difference in reproductive features among ages. *NIPPON SUISAN GAKKAISHI*, **54**, 365–375.

Peter BM, Slatkin M (2013) Detecting range expansions from genetic data. *Evolution*, **67**, 3274–3289.

Price AL, Patterson NJ, Plenge RM *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

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

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raj A, Stephens M, Pritchard JK (2013) *Variational Inference of Population Structure in Large SNP Datasets*. Cold Spring Harbor Labs Journals.

Rathbun MJ (1902) Japanese stalk-eyed crustaceans. Proceedings of the United States National Museum **26**: 23-55. http://dx.doi.org/10.5479/si.00963801.26-1307.23

Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM (2013) Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology*, **22**, 2953–2970.

Reuschel S, Cuesta JA, Schubart CD (2010) Marine biogeographic boundaries and human introduction along the European coast revealed by phylogeography of the prawn *Palaemon elegans*. *Molecular Phylogenetics and Evolution*, **55**, 765–775.

Richardson MF, Sherman CDH, Lee RS, Bott NJ, Hirst AJ (2016) Multiple dispersal vectors drive range expansion in an invasive marine species. *Molecular Ecology*, **25**, 5001–5014.

Rius M, Turon X, Bernardi G, Volckaert FAM, Viard F (2014) Marine invasion genetics: from spatio-temporal patterns to evolutionary outcomes. *Biological Invasions*, **17**, 869–885.

Roman J (2006) Diluting the founder effect: cryptic invasions expand a marine invader's range. *Proceedings of the Royal Society B: Biological Sciences*, **273**, 2453–2459.

Selkoe KA, D'Aloia CC, Crandall ED *et al.* (2016) A decade of seascape genetics: contributions to basic and applied marine connectivity. *Marine Ecology Progress Series*, **554**, 1–19.

Thomaz SM, Kovalenko KE, Havel JE, Kats LB (2014) Aquatic invasive species: general trends in the literature and introduction to the special issue. *Hydrobiologia*, **746**, 1–12.

Torres AP, Santos Dos A, Cuesta JA *et al.* (2012) First record of *Palaemon macrodactylus* Rathbun, 1902 (Decapoda, Palaemonidae) in the western Mediterranean. *Mediterranean Marine Science*, **13**, 278–282.

Travis JM, Munkemuller T, Burton OJ *et al.* (2007) Deleterious mutations can surf to high densities on the wave front of an expanding Population. *Molecular Biology and Evolution*, **24**, 2334–2343.

Tsutsui ND, Suarez AV, Holway DA, Case TJ (2000) Reduced genetic variation and the success of an invasive species. *Proceedings of the National Academy of Sciences*, **97**, 5948–5953.

van Etten, J (2015) gdistance: Distances and Routes on Geographical Grids. R package version 1.1-9. https://CRAN.R-project.org/package=gdistance

Vázquez MG, Bas CC, Kittlein M, Spivak ED (2015) Effects of temperature and salinity on larval survival and development in the invasive shrimp *Palaemon macrodactylus* (Caridea: Palaemonidae) along the reproductive season. *Journal of Sea Research*, **99**, 56–60.

Walsh JR, Carpenter SR, Vander Zanden MJ (2016) Invasive species triggers a massive loss of ecosystem services through a trophic cascade. *Proceedings of the National Academy of Sciences*, **113**, 4081–4085.

Warkentine BE, Rachlin JW (2010) The First Record of *Palaemon macrodactylus* (Oriental Shrimp) from the Eastern Coast of North America. *Northeastern Naturalist*, **17**, 91–102.

White C, Selkoe KA, Watson J *et al.* (2010) Ocean currents help explain population genetic structure. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 1685–1694.

Whitney MM, Ullman DS, Codiga DL (2016) Subtidal exchange in eastern long island sound. *Journal of Physical Oceanography*, **46**, 2351–2371.

Williams ST, Knowlton N (2001) Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus Alpheus. *Molecular Biology and Evolution*, **18**, 1484–1493.

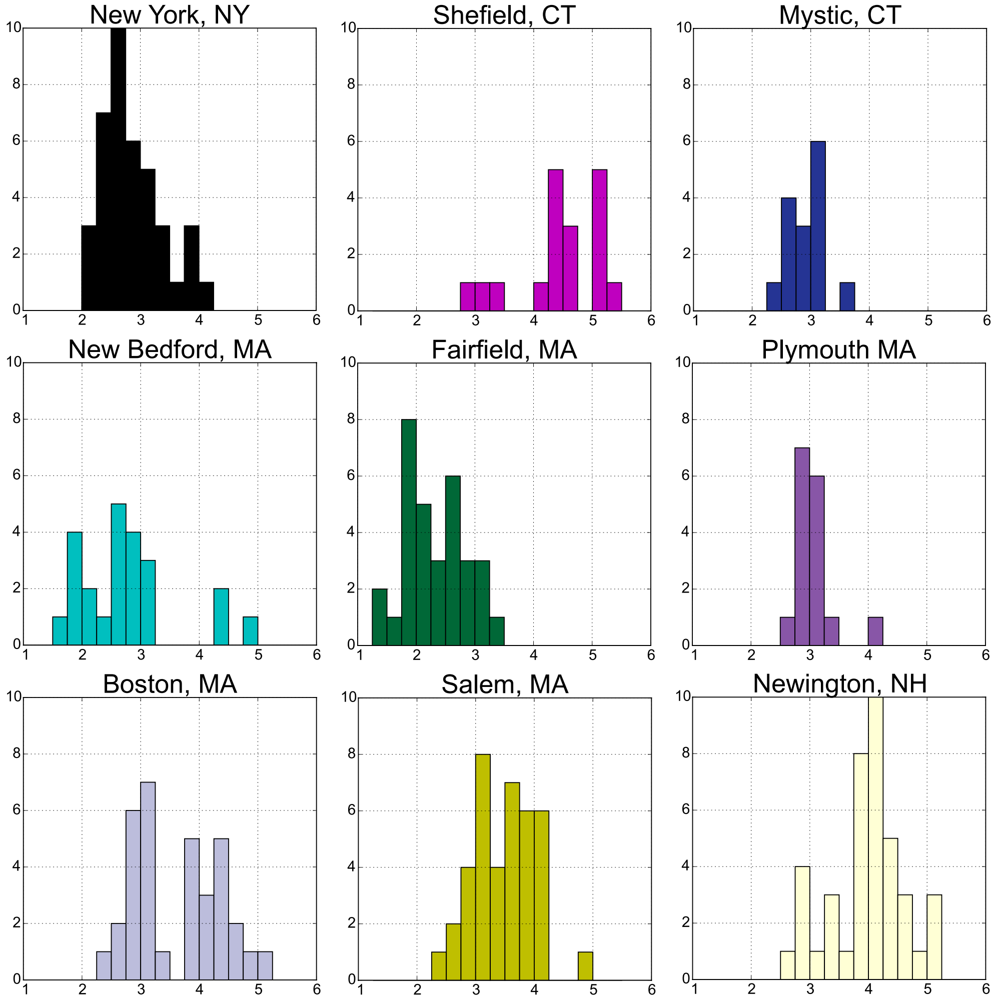
Yoon M, Hong S-E, Kwon Nam Y, Kim DS (2011) Genetic diversity of the Asian shore crab, *Hemigrapsus sanguineus*, in Korea and Japan inferred from mitochondrial cytochrome coxidase subunit I gene. *Animal Cells and Systems*, **15**, 243–249.

Chapter 3 Appendices

Appendix I: Sample size distributions of *Palaemon macrodactylus* individuals

More shrimp were collected from each site included in this study than were used in the genetic analysis. Additionally, shrimp were collected from sites not included in this study as part of the ShrimpEX2014 sampling effort (these sites were Shefield, CT, New Bedford, MA, and Salem, MA). The length distributions of all the shrimp collected are reported here (Figure AI.1) in order to capture the overall distribution of sizes in the invaded range. The length distributions of shrimp used in the genetic analysis are also reported for comparison to overall size distribution from each site (Figure AI.2).

Figure I.1. Length (cm) (from the end of the telson to the base of the rostrum) distributions for all samples collected. The X-axis is length and the Y-axis is the number of shrimp with a specific length.



New York, NY

Shefield, CT

Mystic, CT

New Bedford, MA

Fairfield, MA

Plymouth MA

Boston, MA

Salem, MA

Newington, NH

Figure I.2. Length (cm) distributions (from the end of the telson to the base of the rostrum) for samples included in genetic analyses. The X-axis is length and the Y-axis is the number of shrimp with a specific length.

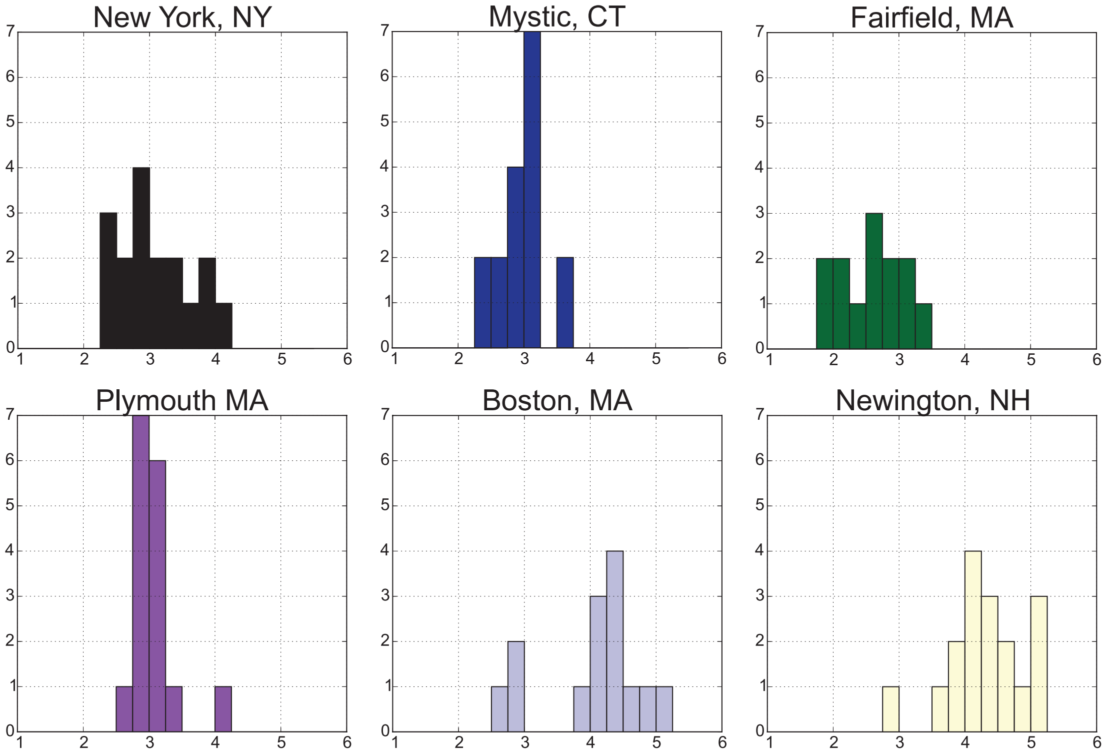


Table I.1. Average length of shrimp (end of telson to base of rostrum) in total sample and in RAD sample (NYC = New York, SHE = Shefield CT, MYS = Mystic CT, FAI = Fairhaven MA, POP = New Bedford MA, PLY = 2.94, BOS = Boston MA, SAL = Salem MA, NEW = Newington, NH).

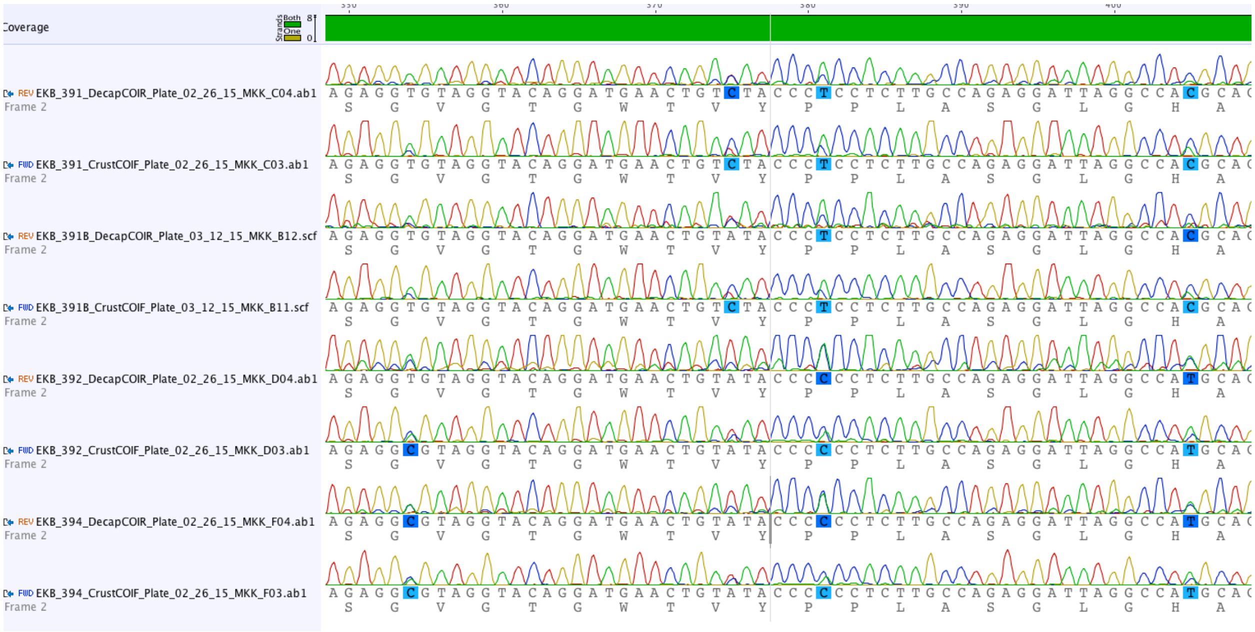
Location	Average length Overall (cm)	Average length for RAD samples (cm)
NYC	2.74	2.96
SHE	4.35	--
MYS	2.80	2.85
FAI	2.20	2.46
POP	2.63	--
PLY	2.94	2.94
BOS	3.50	3.93
SAL	3.37	--
NEW	2.90	4.19

Appendix II: Preliminary protocols used for mitochondrial sequencing and subsequent results

II.A. Original extraction protocol and double-peak Sanger sequencing issues

Initial extractions were executed using Phenol-Chloroform extraction technique as described in (Herrera *et al.* 2015). In the first rounds of sequencing effort, our PCR program included 40 replication cycles of the thermocycler profile (reported in the Methods) in order to generate enough amplified gene product to visualize on a gel with the primers. With the knowledge that some extraction methods result in differential extraction of nuclear DNA and mitochondrial DNA (Guo *et al.* 2009), we attempted to extract the DNA using different methods and reduce the number of PCR cycles to 30. Ultimately, the extraction method reported in the Methods section proved to eliminate our problems with double-banding. This reliably eliminated the double peak phenomenon observed in the raw sequences. The double peaks in our sequences align directly with polymorphic sites published in previous population genetic studies on this species (Lejeusne *et al.* 2014). It is difficult to affirm whether they were true polymorphisms or the results of sequencing a duplicated gene.

Figure IIA.1. Chromatogram data from mt*COI* sequences for each four pairs of sequences, representing three individual *Palaemon* samples (the first four lines of the alignment are two different sampling runs of the same individual). As can be seen for each of the four polymorphic sites in the sequences, there are ambiguous peaks—often in both sequencing directions—that make certain basepair calling impossible.



```
REV EKB_391_DecapCOIR_Plate_02_26_15_MKK_C04.ab1   A G A G G T G T A G G T A C A G G A T G A A C T G T C T A C C C T C C T C T T G C C A G A G G A T T A G G C C A C G C A C
Frame 2                                              S     G     V     G     T     G     W     T     V     Y     P     P     L     A     S     G     L     G     H     A
FWD EKB_391_CrustCOIF_Plate_02_26_15_MKK_C03.ab1   A G A G G T G T A G G T A C A G G A T G A A C T G T C T A C C C T C C T C T T G C C A G A G G A T T A G G C C A C G C A C
Frame 2                                              S     G     V     G     T     G     W     T     V     Y     P     P     L     A     S     G     L     G     H     A
REV EKB_391B_DecapCOIR_Plate_03_12_15_MKK_B12.scf  A G A G G T G T A G G T A C A G G A T G A A C T G T A T A C C C T C C T C T T G C C A G A G G A T T A G G C C A C G C A C
Frame 2                                              S     G     V     G     T     G     W     T     V     Y     P     P     L     A     S     G     L     G     H     A
FWD EKB_391B_CrustCOIF_Plate_03_12_15_MKK_B11.scf  A G A G G T G T A G G T A C A G G A T G A A C T G T C T A C C C T C C T C T T G C C A G A G G A T T A G G C C A C G C A C
Frame 2                                              S     G     V     G     T     G     W     T     V     Y     P     P     L     A     S     G     L     G     H     A
REV EKB_392_DecapCOIR_Plate_02_26_15_MKK_D04.ab1   A G A G G T G T A G G T A C A G G A T G A A C T G T A T A C C C C C C T C T T G C C A G A G G A T T A G G C C A T G C A C
Frame 2                                              S     G     V     G     T     G     W     T     V     Y     P     P     L     A     S     G     L     G     H     A
FWD EKB_392_CrustCOIF_Plate_02_26_15_MKK_D03.ab1   A G A G G C G T A G G T A C A G G A T G A A C T G T A T A C C C C C C T C T T G C C A G A G G A T T A G G C C A T G C A C
Frame 2                                              S     G     V     G     T     G     W     T     V     Y     P     P     L     A     S     G     L     G     H     A
REV EKB_394_DecapCOIR_Plate_02_26_15_MKK_F04.ab1   A G A G G C G T A G G T A C A G G A T G A A C T G T A T A C C C C C C T C T T G C C A G A G G A T T A G G C C A T G C A C
Frame 2                                              S     G     V     G     T     G     W     T     V     Y     P     P     L     A     S     G     L     G     H     A
FWD EKB_394_CrustCOIF_Plate_02_26_15_MKK_F03.ab1   A G A G G C G T A G G T A C A G G A T G A A C T G T A T A C C C C C C T C T T G C C A G A G G A T T A G G C C A T G C A C
Frame 2                                              S     G     V     G     T     G     W     T     V     Y     P     P     L     A     S     G     L     G     H     A
```

II.B. Tests for possible effects of mtDNA pseudo-gene sequences on analysis results

In order to assess the possible implications of unintentional pseudogene sequencing, we executed preliminary tests excluding nucleotide sites where the double-peaks described above existed (*i.e.,* where it was impossible to call a base and instead an "N" was used in the sequence). The initial sequencing run with the protocol described above (Phenol-Chloroform extraction) included 66 individuals from New York, Fairhaven, and Newington. Out of the 66 individuals in the initial studies, we set a threshold for 6 occurrences of an "N" at a specific basepair positions to consider it potentially at risk of being a pseudogene. This resulted in 12 basepair positions flagged as potentially polymorphic due to the sequencing of pseudogene. An alignment of the initial *COI* data, the *COI* data included in the study (using the Biotek Extraction kit) and the published haplotypes from genbank was made and the 12 basepair positions were removed.

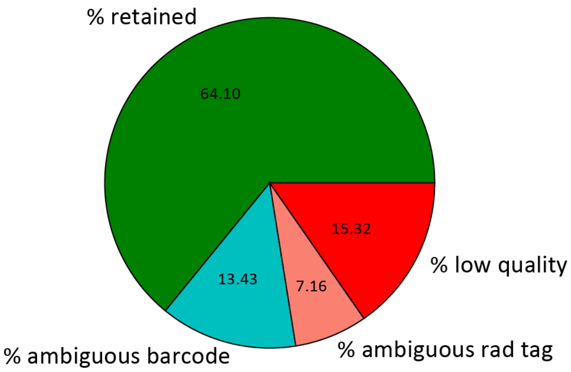
Removal of the polymorphic nucleotide positions potentially based on pseudogene sequencing resulted in a significant reduction of haplotypes from the original study (Lejeusne *et al.* 2014): 85 haplotypes were reduced (merged) to 22 haplotypes. Some haplotypes stayed unmerged while others underwent significant merging. For example, Haplotype *Pm18*, the most common haplotype, absorbed 21 other haplotypes. Thirteen haplotypes did not change due to the omission of certain basepairs. Of these 13 haplotypes, 7 were found only in the native range. In many cases, the reduction of haplotypes through merging would have led to different results for the previously published global population genetic study. For example, Mystic CT—a site included here—would have only 4 haplotypes, one of which was unique to the site instead of 22 haplotypes, with 11 unique. For the data included in this study, the only difference if excluding certain basepairs would be the merging of haplotypes *Pm1* and *Pm3*.

It is impossible to know exactly what factors were responsible for the reported polymorphisms in previously published studies, but it is clear that the first round of our extractions and sequencing resulted in unusable sequences that had uncallable bases at 12 sites in otherwise unambiguous sequences. These 12 sites were major drivers of diversity reported in previous studies. Changes in the levels of diversity in Mystic observed in this study as compared to previous data could be driven by biological factors and not due to differences in sequencing results. However, it is not possible to definitively determine the source of the differences at this time.

Appendix III: Quality and quantity of RAD-sequencing throughout the *Stacks* pipeline

III.A. Process_radtags detailed information

Figure III.A.1. Percent reads retained by *process_radtags*, and percent filtered out for low quality, ambiguous rad-tags, or ambiguous barcodes.



III.B. Detailed results of multiple runs of denovomap.pl

The program *Stacks* allows for sorting and alignment of sequence reads into "stacks," or loci, based on the number of polymorphisms allowed in each "stack" of sequences. Determining the appropriate parameters for this sorting in the absence of information regarding the genomic sequence of a species almost certainly involves splitting some sequenced loci that should be considered one locus into multiple loci and, conversely, combining multiple loci into one locus when in fact multiple loci exist (described in Rodríguez-Ezpeleta *et al.* 2016). The two main parameters that determine how loci are constructed are the *–n* parameter, or the *within-individual distance parameter* which determines how many polymorphisms are allowed at one locus within an individual, and *–M*, or the *between-individual distance parameter* which determines how many polymorphisms are allowed between individual when building a catalog of loci. Altering these values affects the catalog size. *Denovomap.pl* was run 16 times with different permutations of the *–M* and *–n* parameters and the resulting catalog sizes ranged from about 3,820,000 to 380,000 (Figure III.B.1). For this study, the *–M* and *–n* flags were set to three, which represented a catalog size close to the middle of all possible catalog sizes.

The sequencing depth of each stack (locus) averaged across all loci for each individual ranged from 16.03 to 27.72. The mean depth of sequencing increased with an increase in the number of reads, meaning that as more reads were used per individual, coverage increased and not the number of loci being sequenced (Figure III.B.2). The standard deviation of the depth of coverage was relatively high and also increased with the total number of utilized reads. Finally, the number of repetitive reads discarded increased with the total number of utilized reads (Figure III.B.2.

Figure III.B.1. Effect of changing the values of the *-n* and *-M* arguments while running *denovomap.pl*.

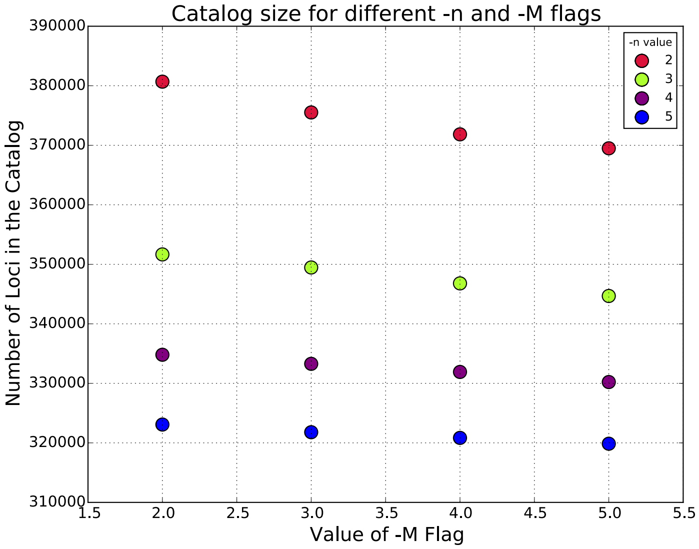
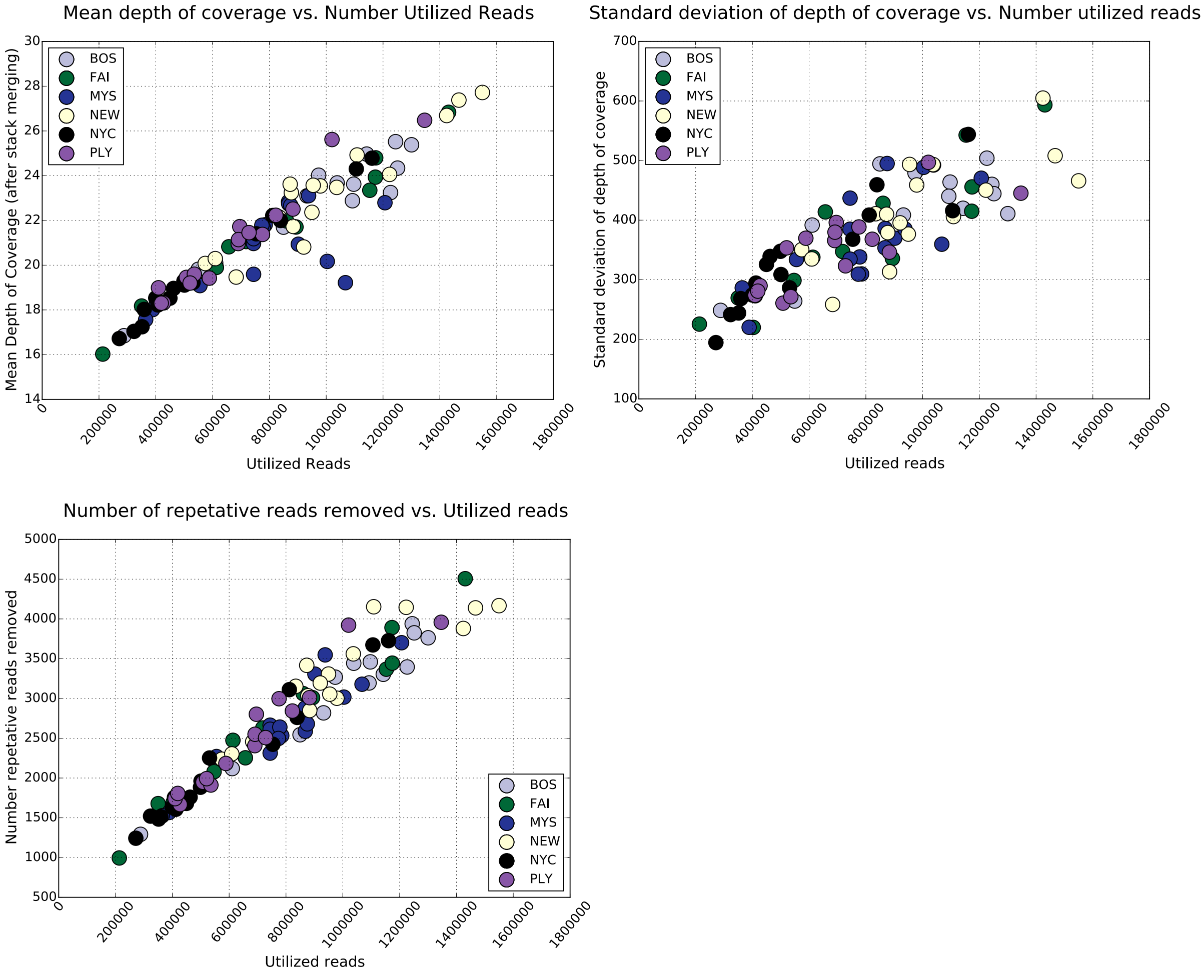


Figure III.B.2. The mean depth of coverage averaged across all loci for each individual (upper left panel), the standard deviation of the depth of coverage across all loci for each individual (upper right panel), and the number of repetitive reads removed vs. the total number of utilized reads (lower left panel). Color key indicates sampling population of each individual.



III.C. Stacks populations program parameter results

As population parameters are relaxed (*i.e.,* made less stringent in terms of how many populations and individuals a locus is present in to be considered). One might expect the number of loci used to converge on a value close to the "true" number of loci, or *SbfI* cut sites in the genome. However, in these data, as requirements are relaxed, and more individuals with unique SNPs are added to analysis, the number of loci does not converge, but rather grows (Figure III.C.1). This is an indication that *most of the loci sequenced are not in most of the individuals*. In other words, loci were not sequenced evenly across all individuals. This could be the result of a poor digest or some problem in the library preparation perhaps related to enzyme efficiency.

Figure III.C.1. Number of polymorphic sites generated with different flags run in the *populations* program in *Stacks* for a different denovomap run. These analyses were executed with slightly different parameters for the denovomap run included in the chapter but are reflective of the final dataset.

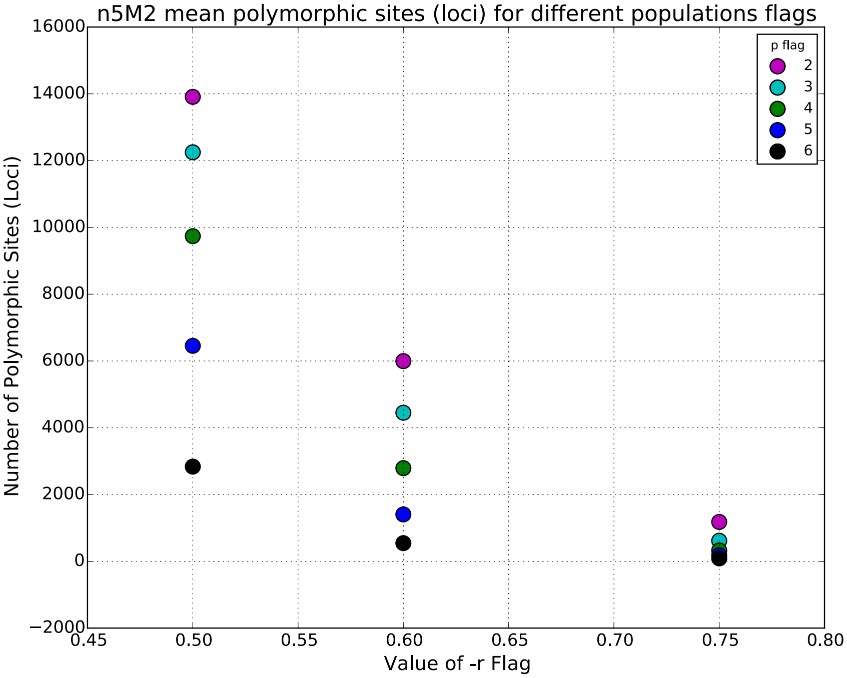


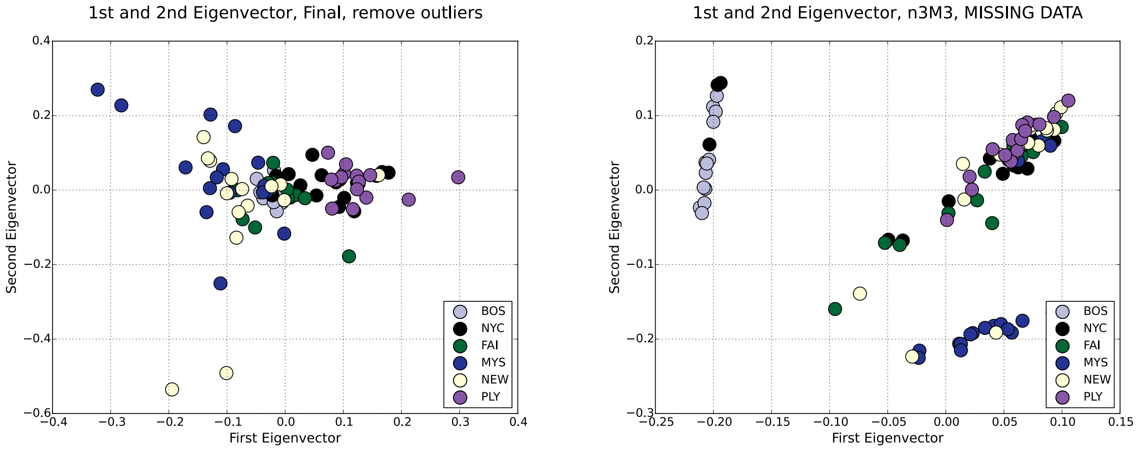
Table AIII.1. Number of loci used in *popluations* analysis with specific argument combinations. These values are for downstream analysis of denovomap parameters –n = 3 and –M = 3. Note that these are different from the data in Figure AIII.2 (above). The patterns, however, were the same although the absolute values varied.

Population parameters	**New York**	**Mystic**	**Fairfield**	**Plymouth**	**Boston**	**Newington**	**All**
-p 6 -r 0.6	-	-	-	-	-	-	498
-p 5 -r 0.6	969	1459	1479	1422	1541	1544	1598
-p 5 -r 0.6, no NYC	-	-	-	-	-	-	1092
-p 4 -r 0.6, no NYC	NA	2732	2827	2516	3091	3111	3329

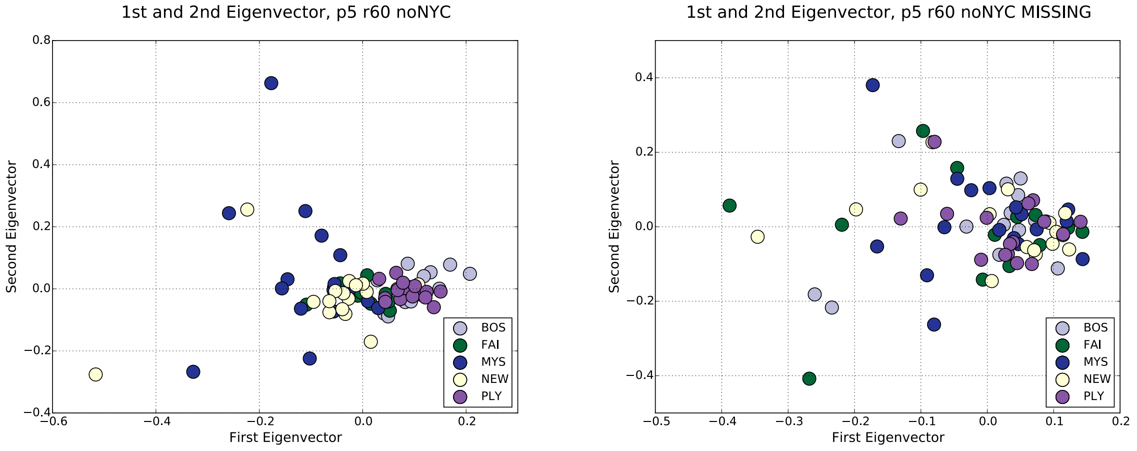
APPENDIX IV: Evaluation of the effects of missing data on principal component analysis

The *smartpca* program within *EIGENSOFT,* which was used to complete the principal component analysis (PCA) for this study has a built-in option to assess the PCA results based solely on missing data denoted by a *'-missingmode'* flag in the parameters file. In the three case studies below, it is evident that in any case when loci are being included in the analysis that are not in all populations (*i.e.*, a *–p* flag that is smaller than the total number of populations) that there is some clustering due to missing data. The *smartpca* program in *EIGENSOFT* replaces missing data with the average value across all populations, therefore pulling the individual or population with the missing data towards the center of the parameter space. This is likely what is occurring with the NYC population in the Case 1 below.

CASE 1: Requiring loci to be in 5 out of 6 populations (using all experimental populations):

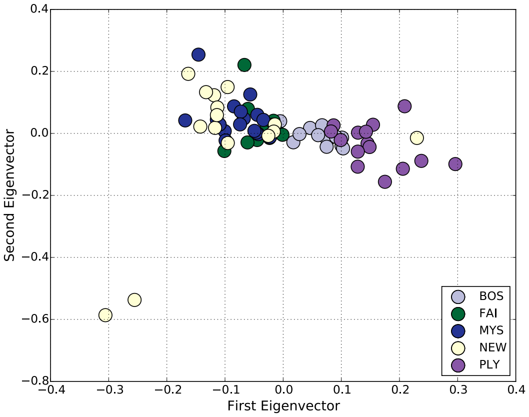


CASE2: Requiring loci to be in 5 out of 5 populations, excluding NYC (the population with fewest loci)

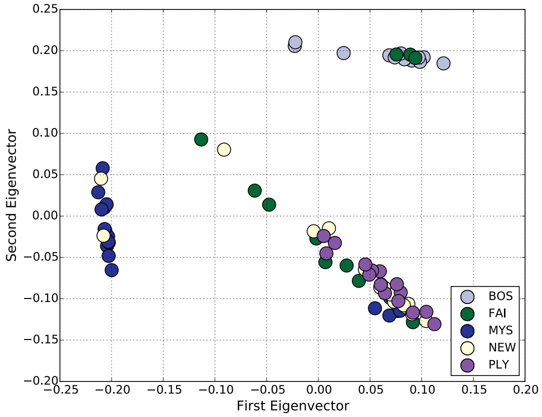


CASE3: Requiring loci to be in 4 out of 5 populations, excluding NYC (the population with fewest loci)

1st and 2nd Eigenvector, p4 r60 noNYC



1st and 2nd Eigenvector, p4 r60 noNYC remove outiers MISSING



Appendix V. fastSTRUCTURE results for multiple *populations* outputs

Two examples of *fastSTRUCTURE* plots below highlight the variability in outcome depending on which *Stacks populations* output is used. The first contains some missing data for each population, possibly driving the structure pattern (Figure AV.1). The second does not have missing data sorted by population but may lose power due to fewer loci (Figure AV.2).

Figure V.1. *FastSTRUCTURE* plot for k=5, the value of genetic lineages that both maximized model likelihood and explained genetic structure for the *populations* run of *Stacks* in which all populations were used but loci were required to be in only 5 and 6 populations and 60% of individuals within a population.

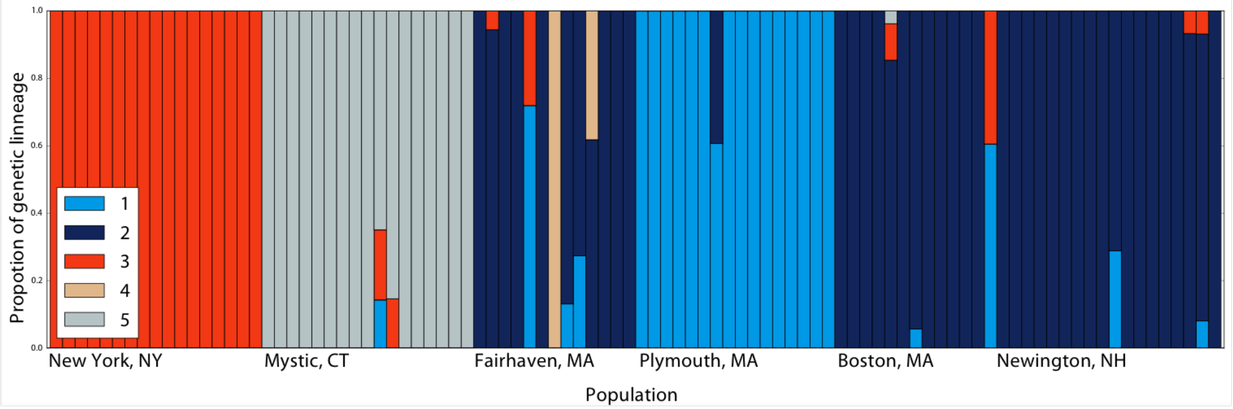
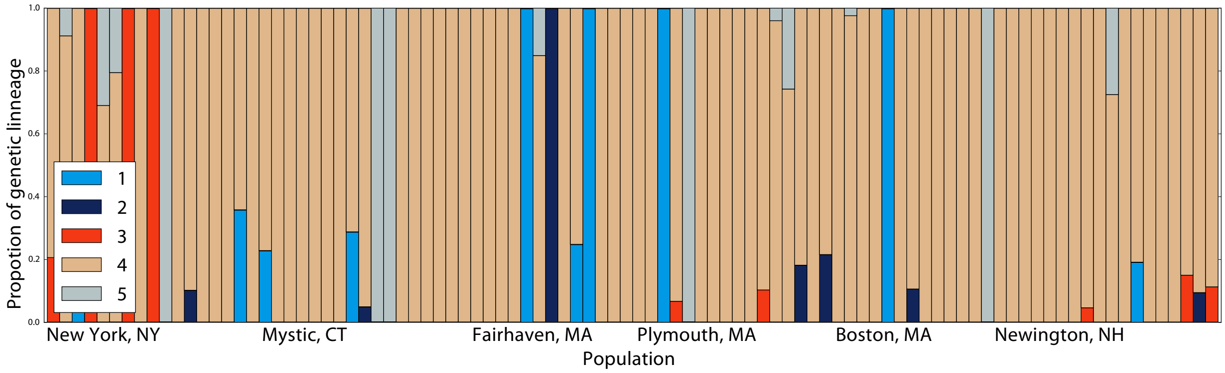


Figure AV.2. FastSTRUCTURE plot for k=5 for the *populations* run of *Stacks* in which loci were required to be in all 6 populations and 60% of individuals within a population.



Literature cited in Chapter 3 Appendices

Guo W, Jiang L, Bhasin S, Khan SM, Swerdlow RH (2009) DNA extraction procedures meaningfully influence qPCR-based mtDNA copy number determination. *Mitochondrion*, **9**, 261–265.

Herrera S, Watanabe H, Shank TM (2015) Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Molecular Ecology*, **24**, 673–689.

Lejeusne C, Saunier A, Petit N *et al.* (2014) High genetic diversity and absence of founder effects in a worldwide aquatic invader. *Scientific Reports*, **4**, 1–9.

Rodríguez-Ezpeleta N, Bradbury IR, Mendibil I *et al.* (2016) Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection. *Molecular Ecology Resources*, **16**, 991–1001.

CHAPTER 4

Non-Equilibrium Population Genomics of the Rapidly Invading Lionfish, *Pterois volitans,* Reveals Expansion Signals Without Spatial Metapopulation Structure

ABSTRACT

Describing the genomic legacies of range expansions is a critical step towards predicting the evolutionary and ecological outcomes of shifting species distributions due to global climate change and species invasions. The invasion of the Indo-Pacific lionfish, *Pterois volitans*, into waters off the US East Coast, Gulf of Mexico, and Caribbean Sea provides a natural model to study rapid range expansion in an invasive tropical marine fish with high dispersal capabilities. During range expansions, strong genetic drift characterized by repeated founder events can result in decreased genetic diversity with increased distance from the center of the historic range, or the point of invasion. We report results from 12,759 loci sequenced by restriction enzyme associated DNA sequencing (RAD-seq) as well as mitochondrial control region D-loop data for nine *P. volitans* populations throughout the invaded range (with one additional site for the mitochondrial analyses). While genome-level analyses are consistent with previous findings of low to no spatially explicit metapopulation genetic structure in the Caribbean Sea, genetic diversity of the lionfish throughout the invaded range is not homogeneous. In fact, patterns of genomic diversity correlate with the expansion pathway. Observed heterozygosity decreases with distance from Florida while expected heterozygosity stays mostly constant, indicating population genetic disequilibrium correlated with distance from the point of invasion. Using an F_{ST} outlier analysis (LOSITAN) and a Bayesian environmental correlation analysis (BayEnv 2.0), we identified 256 and 616 loci, respectively, that are putatively experiencing selection or strong genetic drift. Of these, 24 loci were shared between the two outlier methods, three of which may be involved in movement, growth, and reproduction pathways, potentially indicating adaptation in the invaded range.

INTRODUCTION

The distributions of species are perpetually changing over multiple temporal and spatial scales. For example, re-colonization of high latitudes following glacial retreat has been reported in both terrestrial (Hewitt 1999; 2000) and marine species (Silva *et al.* 2014; Shum *et al.* 2015). The genetic signature of poleward expansions is often described as decreasing genetic diversity with increasing latitude, with occasional examples of habitat refugia altering expansion pathways (Maggs *et al.* 2008). More recently, species distributions have been shifting due to anthropogenic climate change (Parmesan & Yohe 2003; Perry 2005; Harley *et al.* 2006; Pinsky *et al.* 2013) and other sources of environmental change, such as habitat alteration (Bradshaw *et al.* 2014), and non-native species introductions (Lowry *et al.* 2013). Range shifts can be dramatic and rapid, as in the case of a disease epidemic (Lessler *et al.* 2016); or slow and steady, as in the case of some land animals (Gracia *et al.* 2013).

Invasive species are frequently utilized in evolutionary biology as “natural experiments” or models to investigate adaptation to new environments (Barrett 2015). Being able to predict the evolutionary dynamics of invasion is not only important for managing invasions themselves, but is also essential for anticipating impacts of climate-driven range shifts, and for conserving species undergoing distributional shifts due to other anthropogenic disturbances. Despite the importance of these processes to management and conservation, fundamental questions about the genetic impacts of invasion and range expansion remain. Gaps persist in our understanding of the accumulation of mutations during expansion, the drivers of invasion success, and the ways in which genetic diversity changes during invasion following a strong initial bottleneck (Bock *et al.* 2015). Here, we use the invasion of the Indo-Pacific lionfish, *Pterois volitans* [Linnaeus, 1758] as a model for rapid range expansion on a decadal time scale in a marine species with high dispersal capabilities. The use of next generation sequencing and other emerging genomic tools to address these issues is widely recognized as the frontier in invasion genetics research—promising a synergy between previously intractable questions and burgeoning technologies (Chown *et al.* 2014; Rius *et al.* 2015).

The invasion of *Pterois volitans* and *Pterois miles* [Bennett, 1828] in the Western Atlantic and Caribbean Sea is unprecedented in both rate of geographic spread and ecological damage (Hixon *et al.* 2016). *P. volitans* is the most common species in the invasion, with *P. miles* less common and mostly restricted to the northern part of the invaded range (Freshwater *et al.*

2009a). For this reason, the present study focuses only on *P. volitans*. In their native range in the western Pacific and Indian Oceans, lionfish populations appear to be well controlled by predators and competitors (Kulbicki *et al.* 2012), but in their invaded range, lionfish are prolific breeders, insatiable predators, and habitat generalists (Morris & Akins 2009; Morris 2009), resulting in a stark contrast between the ecological role of the fish in its native and invaded range. First reported off Dania, Florida in 1985, the lionfish invasion in the western Atlantic likely originated from an introduction (or several introductions) in southern Florida followed by a long incubation period and an immense post-establishment expansion (Betancur-R *et al.* 2011). In the late 1990s and early 2000s, lionfish began their northward expansion, and by 2004 sightings of juveniles were reported as far north as Cape Cod, although no known breeding populations have been established north of North Carolina. In 2004, lionfish spread to the Bahamas, and in the years since have been reported throughout the Caribbean Sea to Brazil, southeast to the coast of South America, in Panama, Belize, Mexico, and in the Gulf of Mexico (Schofield 2010). Lionfish have most recently been reported south of the Amazon River (Ferreira *et al.* 2015). The timing of the invasion is well characterized by observational data (Schofield 2009) (summarized in Figure 1).

Lionfish present a significant ecological threat to Caribbean reef biodiversity (Hixon *et al.* 2016). In Bahamian reefs, for example, invasive lionfish caused decreases in density, biomass, and species richness of native reef fish communities by approximately 46%, 31%, and 21% respectively (Albins 2015). In many areas, lionfish prey on other reef fish and are known as generalist feeders and opportunistic predators. However, geographic variation has been observed in specific fish diets (Eddy *et al.* 2016). The broad diet of lionfish, their rapid proliferation, and the vulnerability of Caribbean reefs has led some to predict a "worst case scenario" in which the Caribbean will experience "depauperate reef-fish communities and degraded coral reefs" as a direct outcome of the lionfish invasion (Albins & Hixon 2011).

Range expansions, similar to that of the lionfish, are known to result in specific genetic consequences, which have been presented in a body of simulation and theoretical research that has grown dramatically since the early 2000s. The process that dominates much of the literature is known as "allele surfing" (alternatively called "gene surfing" or "mutation surfing"), in which an otherwise rare allele or new mutation rises to high frequency near a range margin because of repeated founder events through space and time (Edmonds *et al.* 2004; Klopfstein 2005; Hallatschek & Nelson 2008; Peischl *et al.* 2013). Allele surfing can vary in strength, leaving either strong or subtle gradients in allele frequencies. In cases of strong allele surfing, the mutation or allele in question may become fixed at the range edge, even when the allele is disadvantageous to the population (Travis *et al.* 2007; Peischl *et al.* 2013; Peischl & Excoffier 2015). These processes could act during an invasion to decrease genetic diversity with increasing distance from the point of introduction (Excoffier *et al.* 2009), as has been observed in humans with distance from Africa (Ramachandran *et al.* 2005). If extreme, such decreases in diversity could lead to a reduction in adaptive potential (Volis *et al.* 2014). The pattern of allele surfing can resemble that of strong selection along the expansion axis.

The potential for population genomics to provide insight into the evolutionary impacts of range expansions and invasion dynamics in non-model systems has been widely acknowledged (Kirk *et al.* 2013; Barrett 2015; Bock *et al.* 2015). Still, while many authors have used genetic techniques to approach range expansion questions in, for example, volcano barnacles (Dawson *et al.* 2010) and green crabs (See & Feist 2009; Darling *et al.* 2014), only a handful of studies to date have explicitly attempted to harness the potential power of next generation sequencing (NGS) for use in range expansion analyses focused on drift and selection (White *et al.* 2013; Tepolt & Palumbi 2015). White *et al.* (2013) found both genetic drift and natural selection in populations of an invasive bank vole in Ireland, including patterns of decreased expected and observed heterozygosity, and allelic richness. While these patterns indicate that the forces described in theory and simulations are likely realized in some invasive species, any generalizations must be based on multiple systems spanning the spectrum of demographic contexts.

For lionfish in the invaded range, recent genetic studies have focused on mitochondrial sequencing to describe population genetic connectivity and population structure. To date, studies have identified just nine haplotypes of the mitochondrial D-loop gene in the invaded range but have not traced these directly back to a specific source in the native range, where genetic diversity is much greater (Freshwater *et al.* 2009a; Betancur-R *et al.* 2011; Butterfield *et al.* 2015). While north-to-south (*i.e.,* Western Atlantic to Caribbean) population differentiation exists in the invaded range, overall, a lack of metapopulation genetic structure has been reported within oceanic basins (Freshwater *et al.* 2009b; Betancur-R *et al.* 2011; Toledo-Hernández *et al.* 2014; Butterfield *et al.* 2015), with some local population structure reported in Puerto Rico (Toledo-Hernández *et al.* 2014). Only four of the nine total haplotypes have been reported in the

Caribbean (Betancur-R *et al.* 2011; Toledo-Hernández *et al.* 2014; Butterfield *et al.* 2015; Johnson *et al.* 2016). Most recently, the expansion into the Gulf of Mexico has resulted in a bottleneck between Caribbean populations and the Gulf of Mexico populations, evidenced by the existence of only three of the four haplotypes found in the Caribbean region (Johnson *et al.* 2016). These course-scale patterns are congruent with large-scale barriers to dispersal between oceanic basins and with the timing of the expansion.

The present study contains the first genome-wide single nucleotide polymorphism (SNP) data for the invasive lionfish throughout the Caribbean Sea using 12,759 loci across nine populations. SNP data are analyzed from a range expansion perspective, identifying changes in genetic diversity with distance from the point of invasion. Our initial prediction was that range expansion would lead to decreased genetic diversity (allele frequency, allelic richness, and heterozygosity) with increased distance from Southeastern Florida. We predicted that signals of allele surfing would be detectable in the SNP data, in line with the theoretical predictions outlined above. Counter to these predictions, patterns of decreased diversity in the form of decreased average allele frequency or allelic richness were not observed in the data. However, decreases in average observed heterozygosity was observed, indicating higher levels of disequilibrium near the range edge, with more central populations tending towards an equilibrium state, as populations become more established and the front moves farther away. Specific loci were identified as outliers in both Bayesian and F_{ST} analyses. These loci could be under selection or experiencing strong genetic drift.

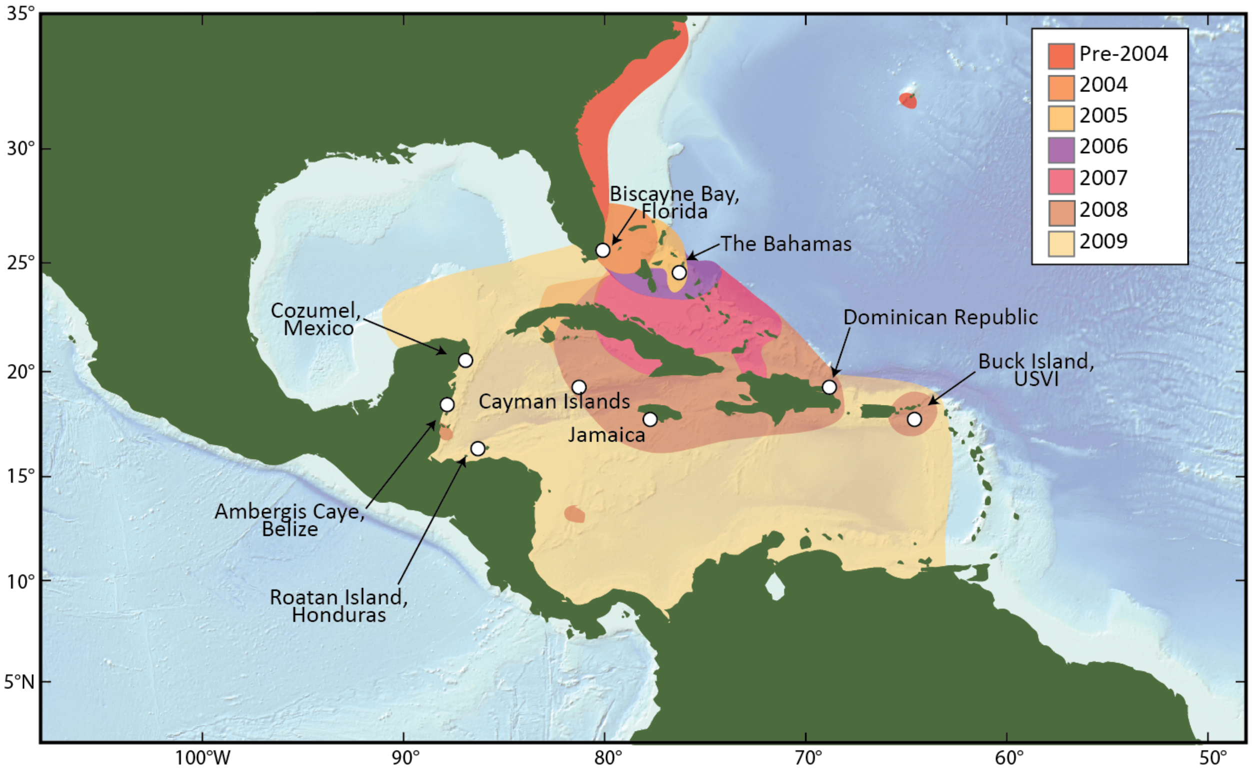
METHODS

Sample collection

Pterois volitans individuals were collected from nine Caribbean sites for genomic analysis (Figure 1). Additional individuals were collected from Trinidad and included in the mitochondrial analysis but not in the RAD-seq analysis. *P. volitans* from Biscayne Bay, Florida, were collected by SCUBA divers from the U.S. National Park Service in August and September of 2013 as part of ongoing collection programs. Fin clips were subsampled from each fish and stored in ethanol in a -20°C freezer. Similarly, samples from the US Virgin Islands were collected from Buck Island by divers from the University of the Virgin Islands between May of 2013 and February of 2014 and fin clips were subsampled. Samples from The Bahamas, the

Dominican Republic, Jamaica, the Cayman Islands, Cozumel Mexico, Belize, Honduras, and Trinidad (for mitochondrial analysis only) were collected by divers throughout 2013 and tissue subsamples were archived in the U.S. National Oceanic and Atmospheric Administration (NOAA) Beaufort Laboratory, Beaufort, North Carolina. Sections of muscle tissue from archived filets were subsampled at NOAA. Fish were identified to species when possible through meristics (*i.e.,* morphological traits) at the collection site and later confirmed through molecular barcoding. If provided by collectors, latitude and longitude, depth, date of collection, sex, and standard or total length for each sample can be found in Appendix I. The latitude and longitude of the most common collection site per country was used in subsequent spatial analyses (Appendix I). Tissue samples were shipped to the Woods Hole Oceanographic Institution in ethanol or frozen and then were stored at -80°C until genomic DNA extraction.

Figure 1. Map of the study region showing the nine sampling locations used for RAD-sequencing (samples from Trinidad, not shown here, were used in the mitochondrial analyses). Colored contours on the map show the extent of the invasion in the years from 2004-2009, by which point all of the nine sites had been invaded (see legend for dates).



To estimate the age of each sampled individual, and therefore the likely time of recruitment of the individuals used in this study, we calculated age from total length using a von Bertalanffy growth curve (Barbour *et al.* 2011). For samples that lacked a standard length measurement but had a total length measurement, we utilized a conversion function to estimate standard length (Fogg *et al.* 2013). Distributions of estimated fish age and recruitment year are presented in Appendix I.

DNA extraction and mitochondrial DNA PCR, sequencing, and analysis

Genomic DNA (gDNA) was extracted from muscle or fin clip tissue using a CTAB and proteinase K digest, a phenol-chloroform purification, and an ethanol precipitation as described in (Herrera *et al.* 2015b). gDNA was stored in AE buffer from a QIAGEN DNeasy Blood and Tissue Extraction Kit (Qiagen GmbH, Germany) at 4° C or -20° C until gene amplification and sequencing.

Polymerase chain reactions (PCRs) were performed targeting the mitochondrial control region D-loop with primers LionfishA-H (5'-CCATCTTAACATCTTCAG TG-3') and LionfishB-L (5'-CATATCAATATGATCTCAGTAC-3') (Freshwater *et al.* 2009b). The thermocycler temperature profile consisted of 95° denaturing step for 3.5 minutes, then 30 cycles of 95° for 30 seconds, 51° for 45 seconds, 72° for 45 seconds, followed by a final extension step at 72° for 5 minutes. PCR reactions were purified using a QIAGEN PCR Purification Kit (Qiagen GmbH, Germany) and were sequenced using Sanger sequencing at Eurofins Operon Genomics **(**Eurofins MWG Operon LLC, Louiseville, KY, USA). Sequences were edited and aligned using *Geneious 8.1.5* (http://www.geneious.com, Kearse *et al.* 2012) and were compared to the previously published haplotypes. Mitochondrial sequence data were generated for a total of 217 individual *P. volitans* samples (23 from Florida, 17 from The Bahamas, 16 from the Dominican Republic, 25 from Jamaica, 15 from the Cayman Islands, 24 from Mexico, 18 from Belize, 24 from Honduras, 23 from the US Virgin Islands, and 32 from Trinidad). Genome-wide single nucleotide polymorphism (SNP) data were generated for a subset of 120 of those samples.

Restriction enzyme associated DNA Sequencing

Restriction enzyme associated DNA sequencing (RAD-seq) library preparation using the *SbfI* restriction enzyme (restriction site: 5'-CCTGCAGG-3') was carried out on concentration-normalized gDNA by Floragenex Inc. (Eugene, OR, USA) in identical fashion to several other recent RAD-seq studies (Reitzel *et al.* 2013; Herrera *et al.* 2015b). A subset of samples were prepared for paired-end Illumina sequencing following the library prep protocol described in Baird *et al.* (2008), in order to generate longer sequencing assemblies for future analyses as well as provide possible comparisons of methods.

In brief, gDNA was digested with the *SbfI* restriction enzyme, yielding fragments of many different lengths. Barcode tags 10 basepairs in length that were specific to each individual, and an Illumina adaptor, were ligated onto the sticky end of the cut site. Samples were then pooled, sheared, and size selected for optimal Illumina sequencing. For the paired-end sample library prep, a second adaptor was ligated to the second end of the read. Libraries were then enriched through PCR and sequenced by 96-multiplex in a single lane of an Illumina Hi-Seq 2000 sequencer (one lane for the single end sequencing, one for the paired end). For the samples sequenced in a paired-end Illumina run, each sample was loaded twice to achieve a standard coverage (*i.e.,* for one individual, two libraries were generated from two aliquots of gDNA with two barcodes).

RAD-seq data processing and population genomic analyses

Using the *process_radtags* program in *Stacks* v1.19, raw Illumina, reads were filtered for quality with a minimum phred score of 10 in a sliding window of 15% read length (default settings) and sorted by individual-specific barcode. Reads were truncated to 90 basepairs (bp), including the 6 basepair restriction site. For the data generated with paired-end sequencing, only the first read was used. Putative loci were generated using the *denovo_map.pl* pipeline in *Stacks* v1.35 (references to *Stacks* from this point forward will all be to this version). We used a *stack-depth parameter* (*-m*) of 3, such that 3 reads were required to generate a stack (*i.e.,* a locus); a *within-individual distance parameter* (*-M*) of 3, allowing for three SNP differences in a read; and a *between-individual distance parameter* (*-n*) of 3, allowing for three fixed differences between individuals to build a locus in the catalog. In initial exploratory analyses, altering the values of the *within-individual* and *between-individual* parameters did not significantly impact the number or identity of downstream loci called by *Stacks* (not reported).

Population summary statistics (allele frequencies, observed and expected heterozygosities, π, and F_{IS}) were calculated by the *populations* program in *Stacks*, using loci found in eight of the nine populations and in at least 80% of individuals per population (command flags *-p* 8, *-r* 0.8). Information on the effect of changing the *-p* and *-r* flags is available in Appendix II. For each RAD-tag, only one SNP was used from 90 bp sequence using the program flag *–write_random_snp* (if there were two or more SNPs in the sequence, *Stacks* would randomly choose one to analyze). Heterozygosity (observed and expected) values were also calculated in the R Package PopGenKit (https://cran.r-project.org/web/packages/PopGenKit/index.html) to provide secondary validations of reported values. Allelic richness was calculated using PopGenKit.

Three methods were used to describe the genetic structure of lionfish populations in the study area: principal component analysis (PCA), a Structure analysis, and F_{ST} calculations. The *smartpca* program in *EIGENSOFT* (Price *et al.* 2006) was used to perform a PCA of genetic diversity. Custom iPython notebooks used to convert *Stacks* PLINK output files into *EIGENSOFT* input files, and for the visualization of the PCA are available at the author's GitHub (https://github.com/ekbors/thesis_scripts). *Smartpca* was run with four iterations of outlier removal (*'numoutlieriter'* = 4) with otherwise default parameters. In addition to the PCA analysis, *fastSTRUCTURE* (Pritchard *et al.* 2000; Hubisz *et al.* 2009; Raj *et al.* 2013) was run with the number of genetic lineages (the value of *k*) set to values between one and ten to assess genetic structure through a hierarchical analysis, and the program *chooseK.py* was run to select the value of *k* most consistent with the program's spatial structure model. F_{ST} values were calculated by the *populations* program in *Stacks* using a p-value cutoff of 0.05 and a Bonferroni correction (using the *'bonferroni_gen'* flag in the *populations* program). In addition to these analyses, frequency spectra of the major alleles and of F_{IS} reported by *Stacks* were plotted in iPython. F_{IS} is calculated as $F_{IS} = \frac{H_S - H_I}{H_S}$ where H_S is the heterozygosity in the subpopulation and H_I is the heterozygosity of the individual. The output from *Stacks* reports F_{IS} values of zero when the H_S is equal to zero ($p = 1$), but in these cases, the numerical value of F_{IS} is actually undefined. In order to remove these values, only F_{IS} values that were calculated when $H_{EXP} > 0$ were used.

Genetic diversity summary statistics were regressed against distance from the southern Florida collection site of Biscayne Bay using the *stats* package from *Scipy* (https://scipy.org). The least cost distance dispersal trajectories used in these regressions were calculated using the 'gdistance' package in R with a bathymetric constraint from ETOPO1 (van Etten, 2015; R Core Team, 2016) with an additional requirement that pathways to sites to the west of Cuba first went around the east side of Cuba, a reasonable alteration considering the direction and strength of the Florida Current, as well as existing literature about the difficulty of dispersal of lionfish across that current (Johnston & Purkis 2015). Other methods of measuring distance were explored, including Euclidian distance and a non-modified least-cost ocean distances (that did not require pathways to go around Cuba) that result in slightly different regressions but ultimately the same conclusions (Appendix III).

In addition to the described approaches of regressing genetic diversity measurements with distance from Florida, we also implemented range-expansion specific analyses (Peter & Slatkin 2013). Using an R package developed by Peter and Slatkin (2013), we calculated *psi*, or the "directionality index," which measures asymmetries in allele frequency data to evaluate the likely direction of expansion in a set of populations and the relative distance of a site to the center of the range.

Genome size estimation

To predict the size of the *Pterois volitans* genome based on the observed number of restriction sites (*i.e.,* half the number of observed RAD loci), we used the linear model and parameter estimates for the *SbfI* enzyme described by Herrera *et al.* (2015a) as implemented in the program *PredRAD* (Herrera *et al.*, 2015a). To generate a range for the number of restriction cut sites for *SbfI*, we ran the *Stacks* pipeline and *populations* program with several different permutations of parameters (Appendix II) and then used a range of the number of total RAD loci generated by the different program runs.

Blast2GO and locus identification

To annotate the RAD loci and infer possible links to gene function, we aligned the sequences to the non-redundant sequence database (restricted to teleost bony fishes) of NCBI using the BLASTx (Basic Local Alignment Search Tool) program as implemented in Blast2GO v2.5.1 (Conesa *et al.* 2005). We used an e-value threshold of $1x10^{-3}$, a word size of three and a HSP length cutoff of 33. BLAST results were used to map Gene Ontology (GO) and annotate RAD loci.

Locus-specific diversity analyses

Custom scripts were developed to identify groups of loci in the data with unique diversity patterns (https://github.com/ekbors/thesis_scripts). Loci were identified for which (1) the major allele switched to the minor allele in at least one of the nine populations (*i.e.*, "*p*" drops below 0.5), or (2) the difference between the maximum and minimum value of the overall major allele among the populations exceeded a defined value (measured at values of 0.5, 0.6, 0.7. 0.8, and 0.9). Loci identified by these filtering techniques were used in analyses of site frequency spectra and F_{IS} to determine if specific loci were driving and/or breaking patterns in the dataset, meaning that the forces driving those loci might be dominating the overall population data.

LOSITAN and BayEnv outlier analyses

To detect genomic outliers potentially under selection or strong genetic drift driven by expansion (which will yield similar diversity patterns), we used two analysis programs. *LOSITAN* (Antao *et al.* 2008) utilized data-wide F_{ST} values to identify loci that were outliers in their F_{ST} values. We ran 1,000,000 simulations in *LOSITAN* for all nine populations with the options for "Forced mean F_{ST}" and "Neutral F_{ST}" selected. The false detection rate was set to 0.01 and a correction was implemented by the program.

The second program used for outlier analysis was *BayEnv 2.0* (Coop *et al.* 2010; Günther & Coop 2013), a program based on a Bayesian analysis that first develops a covariant matrix as a null model and then generates a linear model of relationship between diversity and an environmental factor. We used the calculated ocean distance from Florida as an environmental gradient against which to test patterns of diversity in the data. The method implemented by *BayEnv* is intended to control for underlying population structure by generating a Bayes Factor for each locus indicating its relative goodness-of-fit to the linear model related to the environmental gradient. To interpret these Bayes Factors, loci were binned in decimal intervals (randomly choosing *p* or *q* for each locus). Within each bin, each locus was ranked by its Bayes Factor and that rank was divided by the number of loci in the bin. This creates the empirical distribution from which loci in the top 5% and 1% of Bayes Factor values were identified, as described in Coop *et al.* (2010) and (Hancock *et al.* 2010).

Traditionally, these analyses are used to identify regions of the genome under selection, however, as described in the introduction, signals of allele surfing and strong genetic drift in the case of a range expansion could lead to allele frequency patterns correlating with distance or with expansion in ways that resemble the patterns of selection, meaning that in some cases, the loci showing correlation to the gradient of distance may just as likely be the result of drift as selection (White *et al.* 2013).

RESULTS

Mitochondrial control region population analysis

Mitochondrial haplotypes consisting of 679 basepairs of the mitochondrial control D-loop region were sequenced for 217 samples. Only 5 of the 9 known haplotypes previously described were identified in these samples (Freshwater *et al.* 2009b; Betancur-R *et al.* 2011; Butterfield *et al.* 2015). These haplotypes correspond to previously-named haplotypes H01, H02, H03, H04, and H06. Mitochondrial data do not indicate any new introductions of genetic material since the first publication of mitochondrial population genetic data in 2009. Also in line with previous studies, distributional patterns and haplotype relationships largely corresponded to those described in Butterfield *et al.,* 2015. For most locations, only 2 or 3 haplotypes were present in the tested sample, but all 5 haplotypes were found in the Bahamas samples. For a complete summary of the mitochondrial results, see Appendix IV.

RAD-seq and single nucleotide polymorphism calling

Processing of raw Illumina data by the program *process_radtags* in *Stacks* resulted in the removal of less than 1% of the data due to poor sequencing quality (a low phred score), about 20% of the data due to ambiguity in the restriction site, and between 9% and 16% of reads due to ambiguous barcodes (inability to attribute a sequencing read to an individual). The number of reads removed varied slightly by sequencing type (single end *vs.* paired end) and by population (Appendix II). The mean depth of reads for each individual, averaged over loci, was 24.5 reads and the average of the standard deviations for each individual was 28.3. More in-depth information on the depth of coverage is provided in the supplemental information (Appendix II). *Cstacks* generated a catalog of 1,376,469 putative loci, 12,759 of which were used by the *populations* program and in all subsequent analyses. The overall patterns of genetic diversity and genetic structure were not altered significantly in different parameter runs of *Stacks*. When more loci were included in analyses, heterozygosity increased overall—trends held the same shape but shifted upwards. This filtering-diversity relationship is consistent with what is generally known about RAD-sequencing approaches specifically under-reporting diversity (Arnold *et al.* 2013) and being more conservative in the *populations* filtering for loci.

Genome size estimation

The number of RAD loci identified in multiple populations ranged from 9,502 to 48,079 with the majority of values between 30,000 and 50,000 (data are estimates from one catalog of loci generated by *denovomap.pl*, reviewed in Appendix II). Given that there are two RAD "loci" at each cut site (sequencing in both directions away from the site), we generated estimates for genome size for 15,000, 20,000, and 25,000 cut sites, representing the majority of putative values for cut sites (Table 1). Estimates ranged from 370,725,631 basepairs to 680,784.288 basepairs. Considering these results, the 12,759 loci used in this study represent between 0.17% and 0.31% of the total lionfish genome.

Table 1. Cut site and genome size estimates as generated by PredRAD.

No. of Cut Sites	Lower estimate of genome size	Upper estimate of genome size
15,000	370,725,631	477,646,515
20,000	452,612,181	583,149,943
25,000	528,391,137	680,784,288

Spatial population genomic analyses

Observed heterozygosity decreased linearly with distance from Florida (Figure 2A; Table 2) even though both allelic richness (average number of alleles per locus) and expected heterozygosity (calculated by *Stacks* as *2pq* from the Hardy Weinberg equation) remained almost the same throughout the sampled range (Figure 2B, 2C). The difference between the expected and observed heterozygosity—a measure of deviation from Hardy Weinberg equilibrium—increased with distance from Florida (Figure 3). All measures of distance explored here resulted in similar regressions for observed heterozygosity (Appendix III). In general, site frequency spectra (SFS) followed expectations for the shape of the distribution and distributions for each population were similar to each other (Figure 4). However, the SFS for The Bahamas demonstrated a lower peak to the left of the curve (lower proportion of alleles with major allele frequency at 1) with a thicker tail, which could indicate a shift from having many rare alleles—common in a rapidly growing population—to a more stable distribution. Mexico also has a slightly thicker tail than other populations (Figure 4). F_{IS} distributions in Florida and The Bahamas were closer to an equilibrium expectation of zero than F_{IS} distributions from populations closer to the moving range edge, which showed a thicker tail in the distribution skewing towards 1 (*e.g.*, the Cayman Islands, and Mexico) (Figure 5). These range-expansion patterns were observed despite a notable lack of spatial metapopulation genetic structure.

Figure 2. Summary statistics plotted against the "modifiied" ocean distance, measured from Florida. (A) Observed heterozygosity ($R^2 = 0.744$, p-value = 0.003), (B) Expected heterozygosity (no significant regression), and (B) Allelic richness (no significant regression).

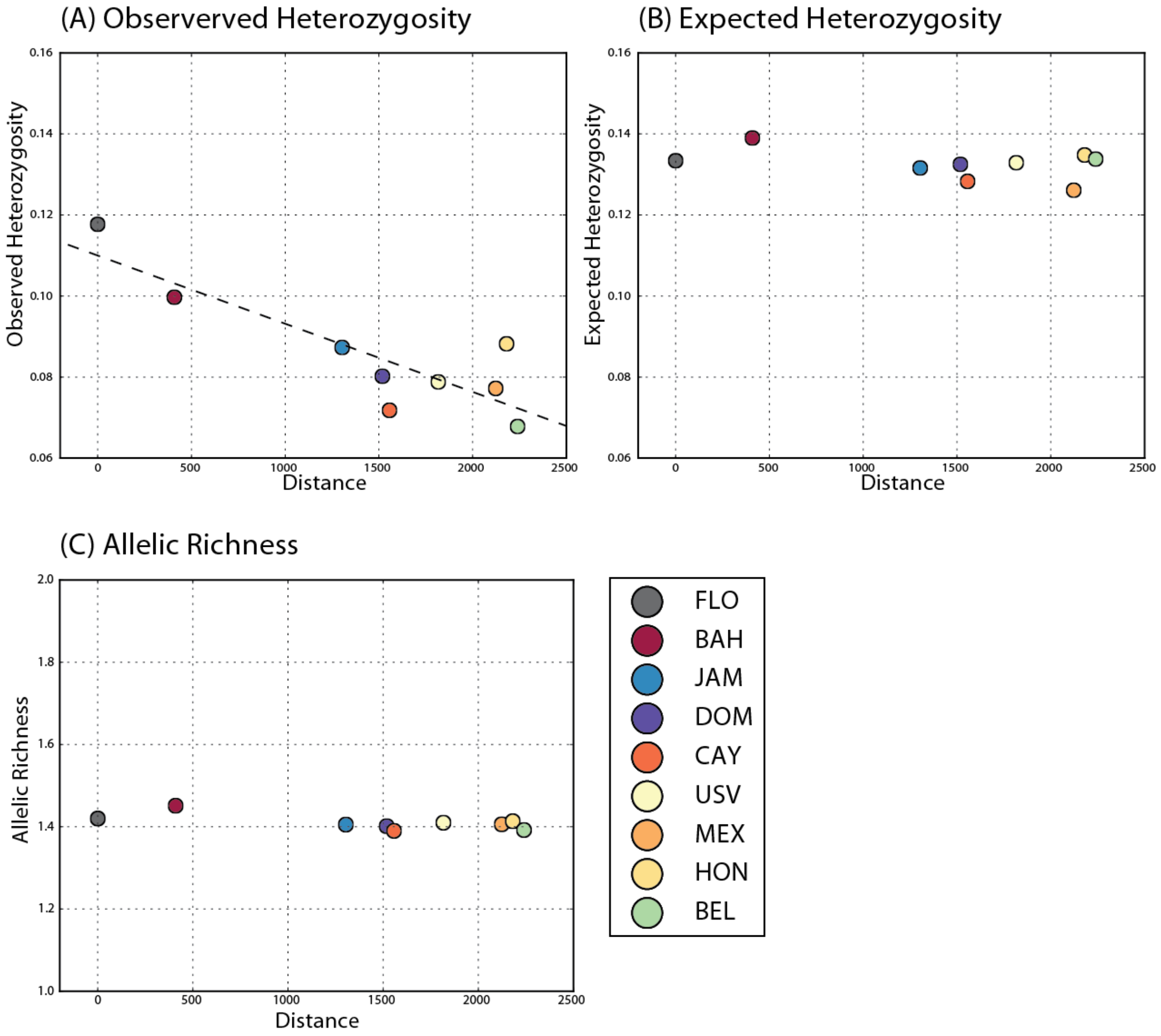


Figure 3. The difference of H_{obs} from H_{exp} vs. distance from Florida ($R^2 = 0.69$, p-value = 0.005).

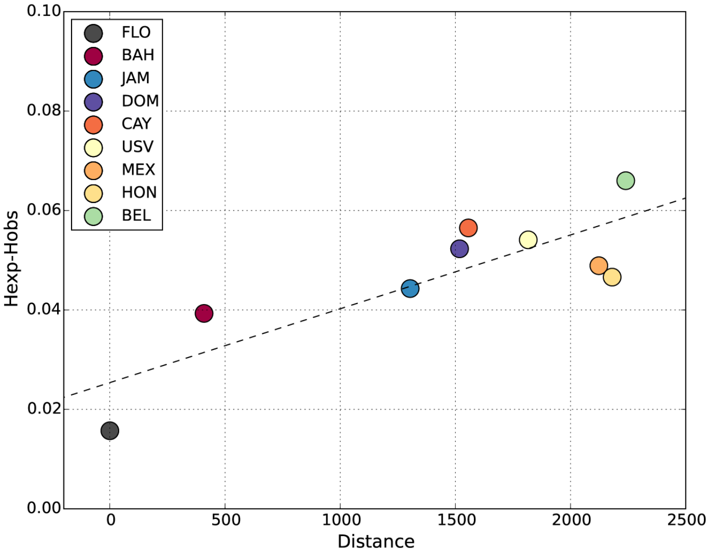


Table 2. Population genetic summary statistics averaged over all loci and by population, as generated by the *Stacks populations* program.

Pop ID	N	N (Stacks)	Private	P	Obs Het	Exp Het	Pi	Fis
FLO	11	10.4786	421	0.9053	0.1177	0.1334	0.1402	0.0643
BAH	9	8.6021	620	0.9024	0.0997	0.139	0.1476	0.1304
CAY	11	10.0495	478	0.909	0.0718	0.1283	0.1351	0.1786
JAM	14	13.0478	746	0.9077	0.0873	0.1316	0.1369	0.1498
DOM	16	14.6043	656	0.907	0.0802	0.1325	0.1373	0.1779
MEX	7	6.5103	381	0.9096	0.0772	0.1261	0.1367	0.1474
BEL	20	17.7624	839	0.9057	0.0678	0.1338	0.1377	0.2272
HON	15	14.6673	781	0.9056	0.0882	0.1348	0.1397	0.1564
USV	16	13.7671	857	0.9069	0.0788	0.1329	0.1379	0.1828

There was no obvious spatial metapopulation genetic structuring among the nine populations in the study region. Principal component analysis in which outliers were removed by the *smartpca* program in *EIGENSOFT* (Figure 6) revealed no clustering of defined populations with the first, second, and third components (*e.g.*, eigenvectors) accounting for 11.03%, 10.44%, and 10.30% respectively of the variation in the dataset. The three outliers removed by the program were from The Bahamas. When included in the analysis, the first, second, and third components accounted for 12.7%, 11.29%, and 10.49% respectively, only slightly higher than when they are removed.

Figure 4. Site frequency spectra for each population showing the proportion of loci in each frequency bin (number of bins = 20) for the major allele (*p* as calculated by *Stacks*).

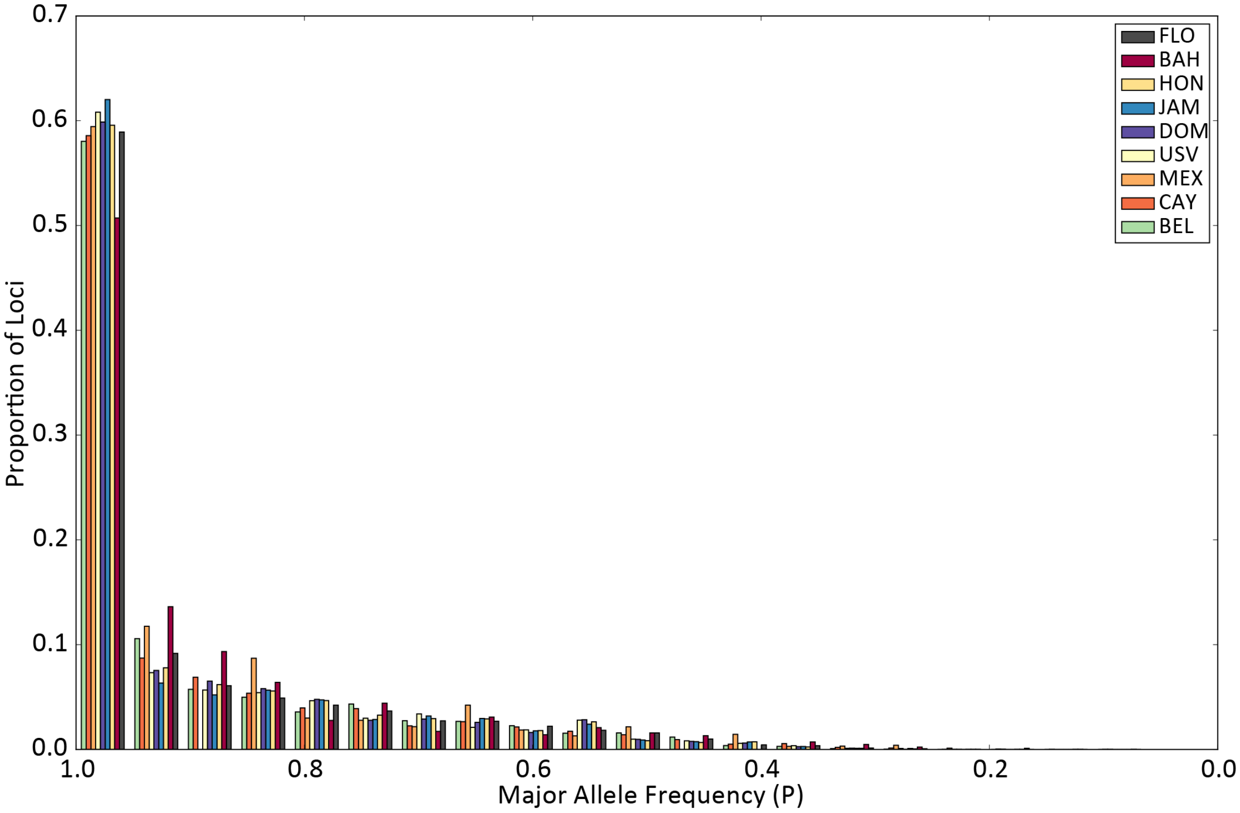


Figure 5. F_{IS} distributions showing the proportion of loci with F_{IS} values within each bin (number of bins = 10) for values between -1 and 1. Values of 0 reported by *Stacks* for loci for which the expected heterozygosity was 0 were removed from the data as described in the Methods.

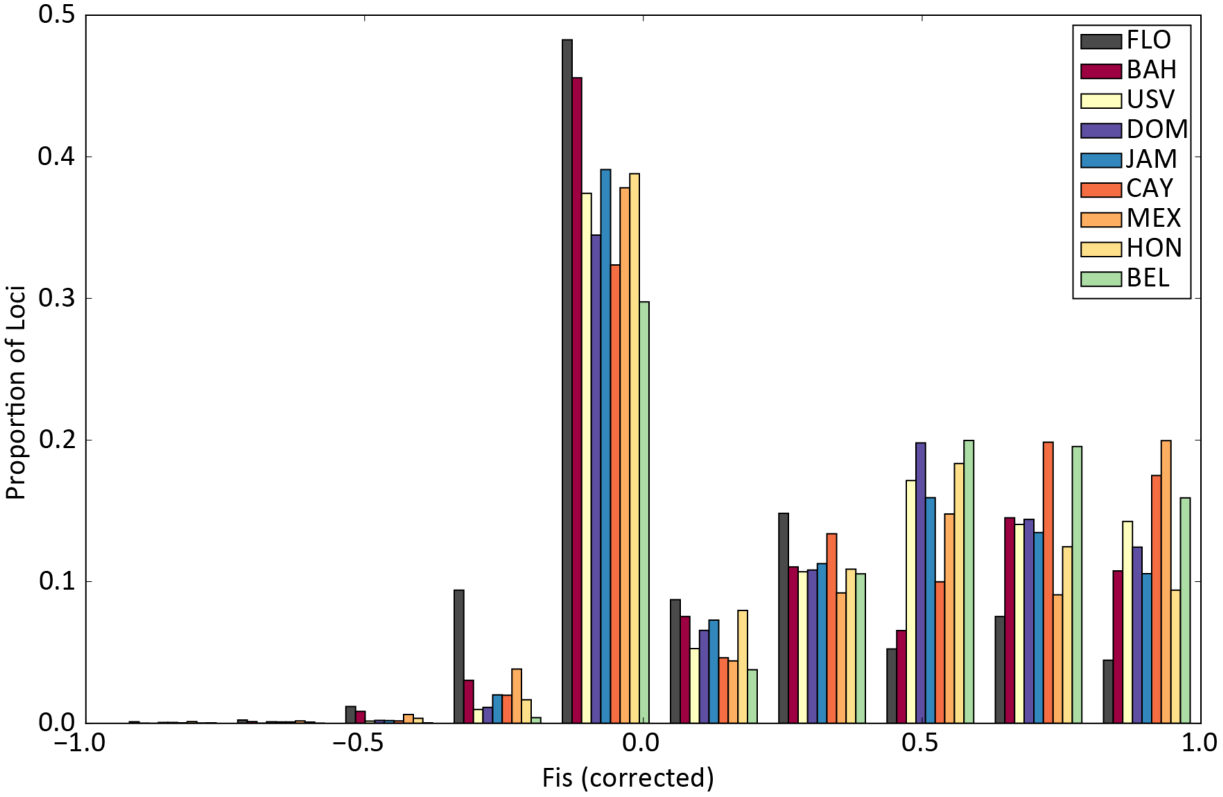
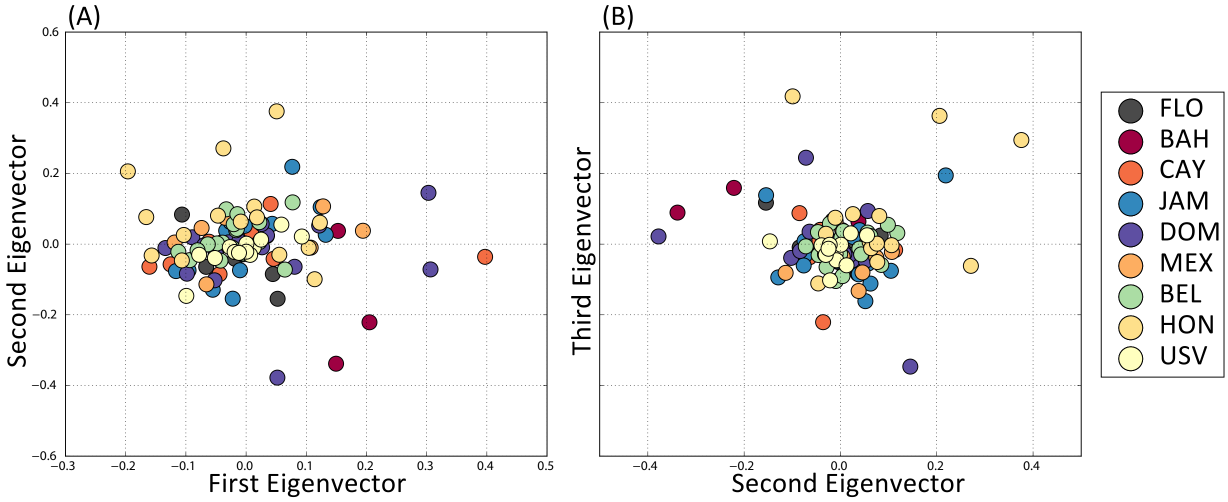
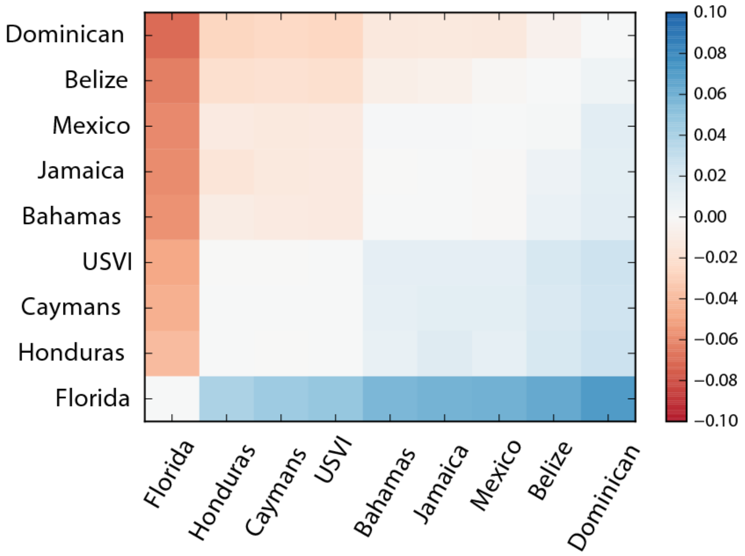


Figure 6. PCA generated by *smartpca* in *EIGENSOFT* here shown for the run with outliers removed.



In order to determine the most likely number of genetic lineages (the value of *k*), or subpopulations, the *chooseK.py* program from *fastSTRUCTURE* was run for values of *k* between 1 and 10. The value of *k* that maximized marginal likelihood and that best explained the structure in the data (two different program metrics for assessing the appropriate value of *k*) was 1, indicating that the *fastSTRUCTURE* analysis fit the data best with just one genetic lineage. After a Bonferroni correction, many pairwise F_{ST} values calculated by *Stacks* were not statistically different from zero. For those that were, F_{ST} values showed very slight genetic differentiation among populations with significant values only for 5 pairings: Bahamas-Belize = 6.91 x 10 $^{-5}$; Caymans-Mexico = 1.1 x 10 $^{-4}$; Jamaica-Dominican Republic = 6.10 x 10 $^{-5}$; Jamaica-Honduras = 1.2 x 10 $^{-4}$; Dominican Republic-Honduras = 1.1 x 10 $^{-}$4. These results indicate that populations closer to the edge of the invaded range are not genetically distinct from those at the center of the range.

Figure 7. Directionality index heatmap. The directionality index, *psi,* measures asymmetries in allele frequencies. Here, values of *psi* have been arranged from lowest to highest—intended to parallel the ordering of sites from closest to the origin of expansion to furthest.



The directionality index indicates another possible concept of distance from the point of invasion based on asymmetries of allele frequencies (Figure 7). The ordering of the index from lowest to highest (for both signs) indicates the “distance” in terms of the expansion from the center of the range. These data are ranked in the following order: Florida, Honduras, the Cayman Islands, US Virgin Islands, Jamaica, The Bahamas, Mexico, Belize, and the Dominican

Republic. This “order” of distance, or invasion directionality is different from an expectation based solely on geographic proximity. Specifically, the results indicate that the Dominican Republic is more isolated from the center of the invasion than all other sites, and that Honduras is much more connected to the core of the range even though it is geographically distant.

Blast2GO and Locus Identification

Blast2GO queries against all existing fish genome databases resulted in matches for 2,766 of the 12,759 loci (21.7%). In most cases, two RAD-tag sequences matched to a BLAST result, which is consistent with having two “loci” sequenced in each direction away from the restriction site. These results could be used in concert with future draft and scaffold assemblies of the lionfish genome to confirm identity and location or RAD loci.

Locus-specific patterns of diversity

There were 1,207 loci for which the value of *p*, or the major allele, defined as the allele most frequent across all the 120 samples, dropped below 0.5 in at least one population, meaning that for those loci, the major allele overall became the minor allele locally (these are referred to in figures as “flip-flop loci”). There were 290 loci with a difference in the minimum and maximum allele frequency of at least 0.5, 55 with a difference of at least 0.6, 3 with a difference of at least 0.7, and 1 with a difference of at least 0.8. There were no loci with a minimum-maximum difference of 0.9 or greater. Of the loci that switched from major allele overall to minor allele in at least one population, 243 were also present in the 0.5 difference list. Therefore, 964 of the loci that switched between being a major and minor allele never had a maximum difference that exceeded 0.5. These loci are likely oscillating around a frequency of 0.5, not demonstrating dramatic changes throughout the invaded range. Such loci are sometimes attributed to balancing selection. The 243 loci with larger differences between their minimum and maximum values, however, could be driven by specific forces such as drift and directional selection.

Pairwise comparisons of allele frequencies in center and edge populations were used to detect specific loci for which frequencies were greater in the core of the invaded range than closer to the edge. From the list of loci that had a difference of 0.5 or more between maximum and minimum allele frequency, 115 had greater allele frequencies in Florida than in the USVI, 127 had greater allele frequencies in Florida than in Honduras, 106 had greater allele frequencies in the Bahamas than in the USVI and 122 had a greater allele frequency in the Bahamas than in Honduras. Additional pairwise comparison results showing counts of loci that overlap with different filtering requirements, including the outlier analyses described below are presented in Table 3. Site frequency spectra (Figure 8) and F_{IS} plots (Figure 9) of the alleles that changed from major to minor allele showed strikingly different patterns than other remaining alleles. These alleles dramatically break the expectation of neutral distribution.

Outlier analyses using LOSITAN and BayEnv

Different loci were identified as outliers using different methods. *LOSITAN* analyses identified 256 loci as possible targets of directional selection (having an F_{ST} outside the upper bound of the 95% confidence interval, with a correction for multiple tests). *BayEnv 2.0* generated Bayes factors for the 12,759 analyzed loci. The binning of loci according to Coop et al. (2010) did not evenly distribute loci across bins. Instead, there were significantly more loci in bins 1 and 10, likely because the majority of loci had allele frequencies very close to one (Appendix V). Taking the top 1% of loci from each bin captured 120 loci considered to have high enough Bayes Factors to be considered correlated to the linear regression model generated by the program *BayEnv*; taking the top 5% captured 616 loci. The top 5% of loci identified by *BayEnv* were then compared to the list of F_{ST} outliers generated by *LOSITAN* and were also compared to lists generated by the locus-specific diversity analyses described above, including the loci with a change from major to minor allele, and those with large differences between their maximum and minimum frequencies (Table 3).

Table 3. Comparisons of the loci identified as outliers by the two outlier analyses and loci identified through different filtering methods through custom analysis presented in this paper.

Overlap of BayEnv top 5% with other filtering methods (total #loci = 616)					
	Lositan	**Flip-flop**	**0.5 diff**	**0.6 diff**	**0.7 diff**
Number shared	24	58	23	5	0
Overlap of Lositan Loci with other filtering methods (total #loci = 256)					
	BayEnv 1%	**Flip-flop**	**0.5 diff**	**0.6 diff**	**0.7 diff**
Number shared	5	100	118	43	3

Only five loci were present in both the top 1% of *BayEnv* Bayes factors and in the *LOSITAN* directional selection results. Of these, four were located in regions without a BLAST hit, and one was putatively identified as the "KN motif and ankyrin repeat domain-containing 4-like" with no GO terms associated with it. Of the 615 loci in the top 5% of BayEnv analysis, however, 24 were also identified as outliers by *LOSITAN* analysis. Of these 24 loci, 7 were putatively identified by Blast2GO. Several of these loci were identified by GO terms as being membrane proteins or involved in membranes (Table 4).

Of the loci identified as outliers in both *BayEnv* and *LOSITAN* (Table 4), four were more closely scrutinized. In a BLAST-n query of the National Center for Biotechnology Information (NCBI) nucleotide database, the identity of locus 48803, a putative glutamate receptor, was highly supported with no gaps in the alignment and between 93% and 95% identity matches with glutamate receptor sequences for the Asian sea bass, *Lates calcarifer*; the bicolor damselfish, *Stegastes partitus*; and the turquoise killifish, *nethobranchius furzeri*. The conserved domain analysis resulted in the identification of this gene region as being a periplasmic binding protein, type I, which is consistent with glutamate receptor proteins. For locus 11751, putatively identified by Blast2GO as a progestin receptor, the BLAST-n alignment yielded a maximum identity of 90% for a progestin receptor sequence from the Asian sea bass, *Lates calcarifer*. BLAST-nr results were the same for both loci. There was no conserved domain identified for locus 11751. Locus 54375, a potential antigen-like protein, was identified as part of the CLECT conserved domain, which includes c-type lectin-like protein domains found across a broad range of proteins, including those found in human dendritic cells and some antigen-like proteins. This locus was not able to be more specifically identified. Finally, there was high support for locus 15012 as a tyrosine kinase with a max identify score of 96%.

The site frequency spectra and F_{IS} results for outlier loci that were identified by *LOSITAN* and those identified by *BayEnv 2.0* were markedly different (Figures 8, 9), indicating that these sites in the genome are likely exhibiting different locus-specific patterns of genetic diversity from each other. Therefore, we also consider the loci that were *not* overlapping in the two datasets to be of interest (those that were in just the *BayEnv* results and those that were in just the *LOSITAN* results), and have flagged further investigation of those loci as an important next step in this research.

Figure 8. Site frequency spectra for different loci filtering methods showing the proportion of loci in each frequency bin (number of bins = 20) for the major allele (*p* as calculated by *Stacks*).

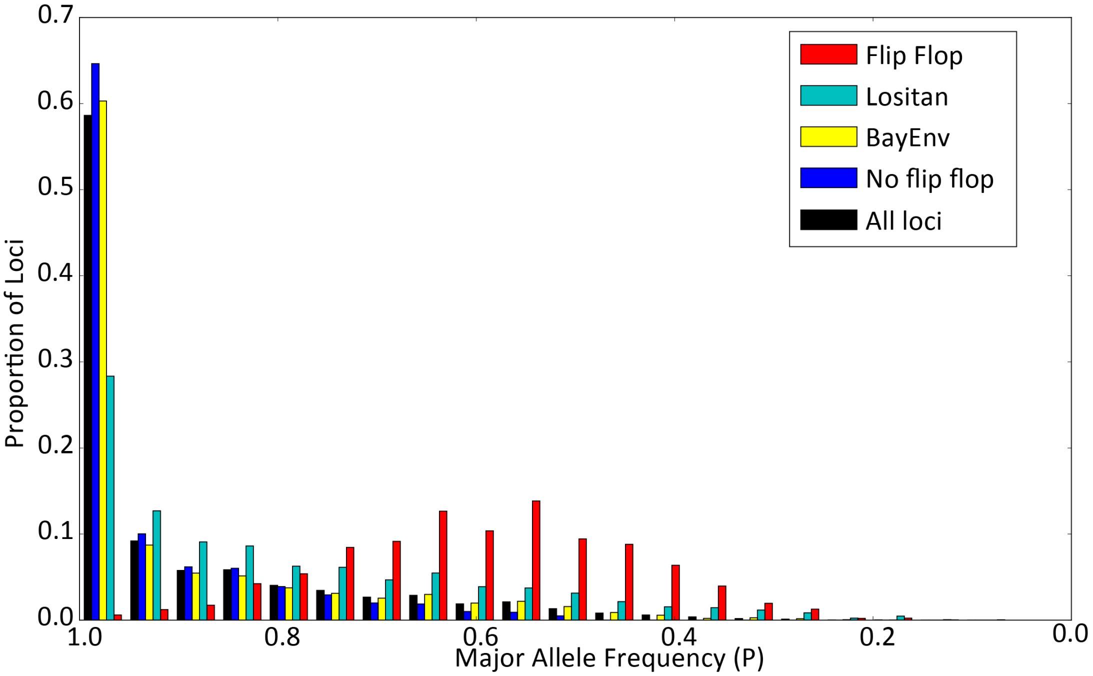


Figure 9. F_{IS} distributions showing the proportion of loci with F_{IS} values for different filtering methods. Values were corrected as described in the Methods section.

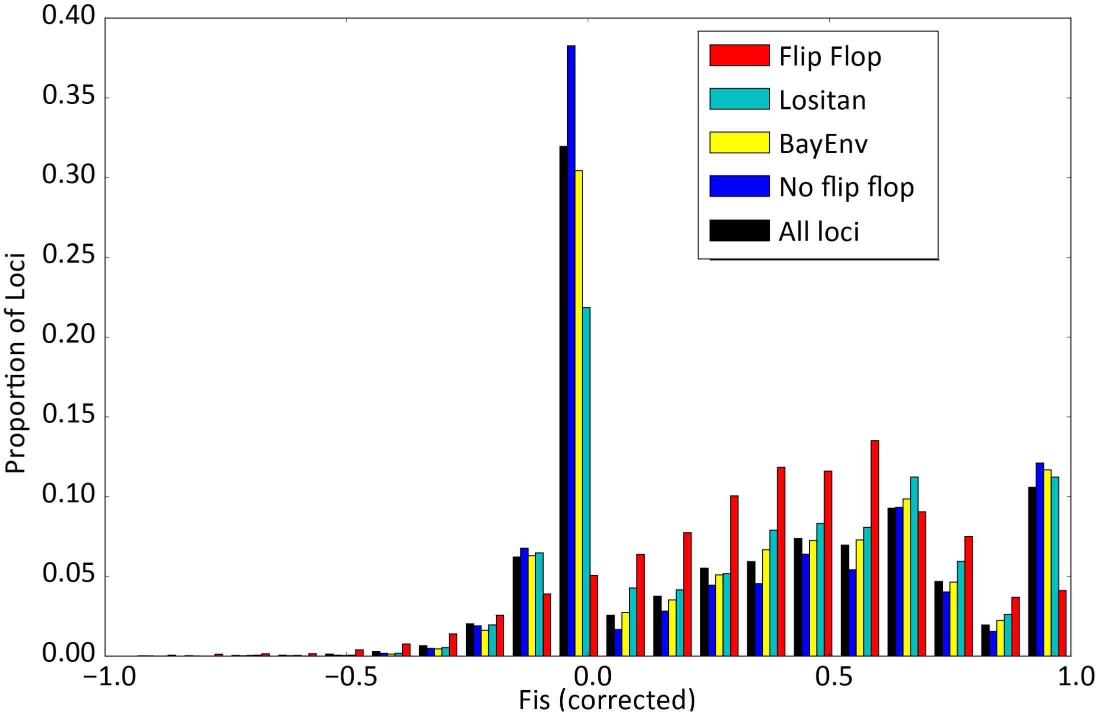


Table 4. Blast2GO results for those loci that overlapped between the BayEnv top 5% and the Lositan outlier results.

Locus	BLAST ID	GO terms
48803	"glutamate receptor NMDA 2B"	C:postsynaptic membrane; P:ion transmembrane transport; C:integral component of membrane; C:cell junction; P:ionotropic glutamate receptor signaling pathway; F:ionotropic glutamate receptor activity; F:extracellular-glutamate-gated ion channel activity
15012	"proto-oncogene tyrosine-kinase Src isoform X1"	F:ATP binding; P:peptidyl-tyrosine phosphorylation; F:non-membrane spanning protein tyrosine kinase activity; P:response to yeast
80176	"coiled-coil domain-containing KIAA1407 homolog"	No GO Terms
20821	"KN motif and ankyrin repeat domain-containing 4-like"	No GO Terms
11751	"membrane progestin receptor beta-like"	C:integral component of membrane
54375	"CD209 antigen E isoform X2"	C:membrane
75133	"PREDICTED: uncharacterized protein LOC103354480"	No GO Terms

DISCUSSION

Genetic disequilibrium detected closer to the moving range boundary

This study contains the first population genomic data generated using RAD-seq for the invasive lionfish, *Pterois volitans*. Using 12,759 loci, we observed geographic patterns correlating diversity with distance from the point of invasion despite a lack of spatial metapopluation genetic structure. The most important of these patterns is the decrease of observed heterozygosity with distance from the point of invasion, despite almost constant values of expected heterozygosity, indicating a relationship between distance from the point of invasion and increased levels of disequilibrium. No geographic metapopulation genetic structure was observed in either a principal component analysis or *fastSTRUCTURE* analysis and only minor differences in F_{ST} values were observed across nine populations in the Caribbean Sea. Slight differences in F_{ST} between some sites could indicate population structuring with more differentiation in The Bahamas and Mexico than other sites; however, the increased F_{ST} at these sites may be more a function of limited sample numbers than of genetic isolation. Mitochondrial data as well as RAD-seq genetic structure analyses were consistent with previous lionfish genetic results in which a strong initial bottleneck was followed by mixing in Caribbean currents has likely led to low levels of population differentiation (Butterfield *et al.* 2015).

Elevated F_{IS} values in populations further from Florida could indicate cryptic structure in the sampled populations (*e.g.*, the Wahlund Effect, Hartl and Clark, 1997). Population densities closer to the edge of the invasion could be lower than those in populations closer to the center of the invaded range, which could lead to signals of cryptic structure. While F_{IS} values are not specifically elevated at sites where fish were sampled from multiple reefs, it is also possible that reef patchiness in different locations, or other sources of habitat heterogeneity could contribute to differences in F_{IS}. In three spined stickleback, elevated patterns of F_{IS} have been linked to cryptic structure in newly colonized freshwater populations (Catchen *et al.* 2013).

Invasion pathways are often either based on field observations or inferred from genetic data. For the locations included in this study, observational data indicate that lionfish invaded first in Florida (1985), then The Bahamas (2004), and then sequentially in the Cayman Islands (February 2008), Jamaica (March, 2008), the Dominican Republic (May 2008), the US Virgin Islands (June 2008 or November 2008), Belize (December 2008), Mexico (January 2009), and Honduras (May 2009) (Schofield 2009; 2010). Observational data, however, can be heavily biased by density of observations being taken, or even by the dissemination of knowledge about an invasion. For example, in the case of lionfish, the observations could be skewed based on the number of recreational and scientific divers in an area, or prior knowledge of the invasion. From the patterns in observed heterozygosity and the directionality index, contrasting hypotheses of invasion pathways can be generated. When ranked by decreasing observed heterozygosity, the invasion pathway would be hypothesized to follow the order: Florida, The Bahamas, Honduras, Jamaica, Dominican Republic, US Virgin Islands, Mexico, Cayman Islands, then Belize. When ranked by the value of the directionality index which is meant to indicate genetic distance in an invasion, the invasion pathway is hypothesized to be: Florida, Honduras, Cayman Islands, US Virgin Islands, Jamaica, Bahamas, Mexico, Belize, and Dominican Republic. In this case, the order deviates strongly from geography of the invaded range and from the observational data, with Honduras being possibly “closer” to the invasion center than, for example, The Bahamas. It is possible that the sightings database could still be the most accurate, with deviations to the ordering with heterozygosity or directionality index arising from other genetic processes.

Differences in these invasion hypotheses, like the possible oceanographic connection indicated by the directionality index between Honduras and Florida and the higher ranking of Honduras in the heterozygosity values, could be the result of oceanographic current patterns or of biological and environmental peculiarities of the lionfish or the sites being colonized.

Outlier analyses identify loci primarily without BLAST IDs

Using over ten thousand loci, we identified sites in the genome that break with equilibrium expectations. Identifying the types of loci that are outliers can give an indication of the proportion of the genome experiencing selection *vs.* expansion-driven drift. We identified 24 loci that are likely undergoing selection or strong genetic drift during expansion, including seven with BLAST hits, several of which were identified as membrane proteins by Blast2GO analysis. The majority of the identified outliers did not have BLAST hits, possibly meaning that these outliers are signals of strong genetic drift acting on neutral portions of the genome. However, there are several other explanations for this ratio of putative genic to non-coding SNPs acting as outliers. First, the SNPs could reside in neutral DNA, which could indicate that these are signals of strong genetic drift and not selection. However, they could also be in regions of DNA that are linked to genes under selection, they could be in genes lacking annotation within the genomes queried, or they could be within genes that are more divergent in lionfish, preventing detection using BLASTx. Because the knowledge of gene identity of our RAD-tags is only cursory, we are unable to specifically disentangle signals for beneficial or deleterious mutations or alleles; however, we are able to infer that signals of surfing are not strong enough to effect average allele frequencies throughout the range. Further analysis of assembled paired-end RAD-seq data could aid in better identifying the location of these loci in fish genomes because longer sequences could improve genome query results (Bourgeois *et al.* 2013).

In analyzing selection outlier results, it is common practice to identify loci that are shared among multiple outlier identification software programs and consider them to be stronger candidates for selection than those found only by one program. This is often done because it is widely acknowledged that each method has its own limitations and biases that may skew the data when only one program is used (Lotterhos & Whitlock 2014). Outliers identified by these programs represent estimates, or hypotheses, based on specific models of selection. The loci identified in this study by *LOSITAN* and *BayEnv* have markedly different genetic patterns (F_{IS}

and site frequency spectra, Figures 8 and 9), which indicate that they may be experiencing different evolutionary forces. *BayEnv* and *LOSITAN* use different metrics to find loci of interest, therefore it is not surprising that their identified loci exhibit different patterns of site frequency and other metrics (*e.g.*, F_{IS}) than each other.

Of the loci that were identified as outliers by both *BayEnv* and *LOSITAN* and further confirmed through more extensive BLAST analysis, three in particular stand out as being potentially important for lionfish success in the invaded range because of the functional roles of these proteins. The first is the glutamate receptor, locus 48803. Glutamate receptors, especially the NMDA receptor regions, play a role in learning and memory (Riedel *et al.* 2003). In marine fish, the receptor has been identified as important for spatial working memory (Partridge *et al.* 2016), and exploration-related behaviors that could be important to finding food or habitat (Pujolar *et al.* 2014a). Glutamate receptors have been shown to be under selection in other teleost fish species, including the European eel and small yellow croaker (Pujolar *et al.* 2014b; Liu *et al.* 2016). There is some evidence that behavioral movement patterns are changing as the lionfish invasion progresses (Benkwitt 2016), which could be related to changes in the pathways involving spatial memory and exploring behaviors. The second locus of interest is locus 11751, which is potentially a membrane progestin receptor. Progestin receptors have been shown to play a role in gamete maturation in Atlantic croaker (Tubbs *et al.* 2010), oocyte maturation in spotted seatrout (Zhu *et al.* 2013), as well as sperm hypermotility in southern flounder (Tan *et al.* 2014). Selection in progestin receptors could be related to the extremely high fecundity observed in the invaded range. The final locus of interest is the proto-oncogene tyrosine kinase, locus 15012. Tyrosine kinases play a role in cell division and growth and as such have garnered attention as oncogenes in cancer research (Vivanco & Sawyers 2002). However, they have also been shown to be related to growth in fish, including rainbow trout (Newsted & Giesy 2000). As such, the detection of this locus as an outlier could indicate adaptive change in growth pathways in the invasive lionfish.

Oceanographic considerations of range expansion

In marine environments, the concept of distance is complicated by a dynamic circulation patterns in which currents are responsible for the movement of marine larvae across large distances. Of the three measures of distance used in this work, the fit of the linear regression of observed heterozygosity was the best for the modified ocean distance measurements—the ones that require the distance pathway to go around the east end of Cuba, the first in a long list of possible modifications of distance measurements to match oceanographic currents. Modifying these estimates of distance further based on oceanography could lead to even better fits of genetic signals. Hypotheses of invasion pathways generated by our genetic data could be further tested by coupled biophysical models that consider specific dispersal patterns by year (Cowen *et al.* 2006; Galindo *et al.* 2006). These models help to account for specific life histories and dispersal traits of the marine species and their environment.

Dispersal capability likely influences the genetic patterns during a range expansion. In theoretical range expansion literature, relatively limited and constant dispersal is often assumed; however, many invasive species have the capacity to disperse long distances. For example, long distance dispersal has been shown to counteract the impacts of repeated founder events during expansion in European starlings in South Africa (Berthouly-Salazar *et al.* 2013). Even in marine systems that have relatively high local retention and self-recruitment, stochastic long-distance dispersal events are thought to be major drivers of genetic patterns (Simpson *et al.* 2014). Many marine species have the capacity to disperse hundreds to thousands of kilometers during their pelagic larval phase via ocean currents (Cowen *et al.* 2007), and in that way, currents shape population genetic structure from shallow water (White *et al.* 2010) to the deep sea (Bors *et al.* 2012). In fact, in many invasive species, selection has been demonstrated for traits that confer the ability for long-distance dispersal, or increased dispersal speed or capacity, as has been seen in ladybirds (Lombaert *et al.* 2014), cane toads (Phillips *et al.* 2010), and crickets (Thomas *et al.* 2001). In lionfish, the invasion of the Caribbean is thought to have been facilitated by long-distance dispersal events driven by hurricanes and other severe disruptions to standard current flow (Johnston & Purkis 2015).

Study design and genetic signals

The year of recruitment of individual fish is likely to affect genetic outcomes. For example, the fish sampled from the Cayman Islands—while collected in 2013—putatively recruited to the reef as early as 2005/2006, which would make them the oldest fish in the study (see Appendix I). The observed heterozygosity of the Cayman Islands population falls below the regression line generated for the data. This finding could be a result of the fact that the sampled

Cayman fish are from an older age bracket, potentially representing a genetic cohort from earlier in the invasion—one of lower diversity than in expected in 2013. Therefore, the age and recruitment date of samples in population genetic studies of range expansion can impact results. Population genetic papers usually assume a sufficiently long time scale of genetic change that the specific age class of individuals sampled is unimportant to the genetic conclusions (Bors *et al.* 2012); however, in range expansions when rapid genetic change is expected, differences of one or two years could change the expected genetic signals.

This study is the first to employ RAD-seq to describe range expansion genetics in lionfish. While the development of NGS has accelerated the generation and analysis of large amounts of reduced representation genomic data, and the subsequent resolution of questions in the field of non-model species genomics (Reitzel *et al.* 2013; Therkildsen *et al.* 2013; Merz *et al.* 2013), limitations to our analysis without whole genome sequencing remain. Specifically, allele surfing may be too difficult to definitively detect using the methods in this paper because identifying and measuring the frequency of rare alleles requires specific sampling considerations not always possible when relying on field sample collection. Our sites were distributed throughout the known invaded range, recognizing that perfect sampling of the whole area of invasion in the Caribbean is nearly impossible. Strong allele surfing could still be taking place and remain undetected by our analyses. The use of reduced representation libraries still only yields data for less than one percent of the lionfish genome. Therefore, the likelihood of capturing loci that are experiencing allele surfing—unless there are many such loci—is low.

Conclusions and implications

Range expansions, while an undeniably important force in shaping genetic diversity across the planet, have limited signatures in some species due to the specific context of the expansion. Here we have demonstrated that while not all the predicted patterns of expansion manifest themselves in the lionfish populations sampled, the range expansion process has led to disequilibrium closer to the range front. Caribbean populations of lionfish are well mixed and dispersal among sites is high, potentially precluding the detection of predicted decreases in allele frequency along the expansion axis in populations sampled in 2013.

Ultimately, the lack of obvious decreases in average allele frequency or allelic richness suggests that the process of expansion is unlikely to cause long-lasting limits to the adaptive potential of lionfish in their invaded range. It could also be inferred that signals of disequilibrium dissipate over time and space for the lionfish. Time series comparisons of spatial patterns of diversity will be crucial to fully understand how a rapid invasion like that of the lionfish affects adaptive potential and the evolution of the species. In addition to temporal research, having better coverage of the lionfish genome and an understanding of exactly what part of the genome is being sequenced will help to clarify how the process of range expansion effects genomes.

ACKNOWLEDGEMENTS

I would like to thank Dr. Timothy Shank, Dr. Santiago Herrera and Dr. James Morris, who are all coauthors on the manuscript in preparation based on this chapter. Thank you also to the many participants of the Gulf and Caribbean Fisheries Institute for providing lionfish samples from around the Caribbean region, as well as Dr. Bernard Castillo at the University of the Virgin Islands and Kristian Rogers at the Biscayne Bay National Park. I would like to acknowledge Alex Bogdanoff at NOAA, Beaufort NC, for assistance with sample acquisition; Camrin Braun at WHOI, for assistance with the calculation of oceanic distances between sites; and Dr. Tom Schultz at Duke Marine Lab and Dr. Margaret Hunter at USGS for discussions concerning ongoing population genetic projects. An additional thanks to Jack Cook of the WHOI Graphics department for his assistance in generating the maps of the study area. I would like to extend a special thank you to Dr. John Wakeley of Harvard University for assistance in the analysis and interpretation of data presented in this chapter as a dissertation committee member. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374. Sequencing funding was provided in part by the PADI Foundation Grant No. 14904. Additional research support was provided by the Woods Hole Oceanographic Institution (WHOI) Ocean Ventures Fund, the Coastal Ocean Institution at WHOI, the National Science Foundation (OCE-1131620 to TMS), and the James Education Fund for Ocean Exploration within the Ocean Exploration Institute at WHOI.

LITERATURE CITED

Albins MA (2015) Invasive Pacific lionfish *Pterois volitans* reduce abundance and species richness of native Bahamian coral-reef fishes. *Marine Ecology Progress Series*, **522**, 231–243.

Albins MA, Hixon MA (2011) Worst case scenario: potential long-term effects of invasive predatory lionfish (*Pterois volitans*) on Atlantic and Caribbean coral-reef communities. *Environmental Biology of Fishes*, **96**, 1151–1157.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: A workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics*, **9**, 323–5.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

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

Barbour AB, Allen MS, Frazer TK, Sherman KD (2011) Evaluating the potential efficacy of invasive lionfish (*Pterois volitans*) Removals (H Browman, Ed,). *PLoS ONE*, **6**, e19666–7.

Barrett SCH (2015) Foundations of invasion genetics: the Baker and Stebbins legacy. *Molecular Ecology*, **24**, 1927–1941.

Benkwitt CE (2016) Invasive lionfish increase activity and foraging movements at greater local densities. *Marine Ecology Progress Series*, **558**, 255–266.

Berthouly-Salazar C, Hui C, Blackburn TM *et al.* (2013) Long-distance dispersal maximizes evolutionary potential during rapid geographic range expansion. *Molecular Ecology*, **22**, 5793–5804.

Betancur-R R, Hines A, Acero P A *et al.* (2011) Reconstructing the lionfish invasion: insights into Greater Caribbean biogeography. *Journal of Biogeography*, **38**, 1281–1293.

Bock DG, Caseys C, Cousens RD *et al.* (2015) What we still don't know about invasion genetics. *Molecular Ecology*, **24**, 2277–2297.

Bors EK, Rowden AA, Maas EW, Clark MR, Shank TM (2012) Patterns of deep-sea genetic connectivity in the New Zealand region: implications for management of benthic ecosystems. **7**, e49474.

Bourgeois YXC, Lhuillier E, Cezard T *et al.* (2013) Mass production of SNP markers in a nonmodel passerine bird through RAD sequencing and contig mapping to the zebra finch genome. *Molecular Ecology Resources*, **13**, 899–907.

Bradshaw CJA, Brook BW, Delean S *et al.* (2014) Predictors of contraction and expansion of area of occupancy for British birds. *Proceedings of the Royal Society B: Biological Sciences*, **281**, 20140744–20140744.

Butterfield JSS, Díaz-Ferguson E, Silliman BR *et al.* (2015) Wide-ranging phylogeographic structure of invasive red lionfish in the Western Atlantic and Greater Caribbean. *Marine Biology*, **162**, 773–781.

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

Chown SL, Hodgins KA, Griffin PC *et al.* (2014) Biological invasions, climate change and genomics. *Evolutionary Applications*, **8**, 23–46.

Conesa A, Gotz S, Garcia-Gomez JM *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Cowen RK, Gawarkiewic G, Pineda J, Thorrold SR, Werner FE (2007) Population connectivity in marine systems: an overview. *Oceanography*, **20**, 14–21.

Cowen RK, Paris CB, Srinivasan A (2006) Scaling of connectivity in marine populations. *Science*, **311**, 522–527.

Darling JA, Tsai YHE, Blakeslee AMH, Roman J (2014) Are genes faster than crabs? Mitochondrial introgression exceeds larval dispersal during population expansion of the invasive crab *Carcinus maenas*. *Royal Society Open Science*, **1**, 140202–140202.

Eddy C, Pitt J, Morris JA Jr *et al.* (2016) Diet of invasive lionfish (Pterois volitans and P. miles) in Bermuda. *Marine Ecology Progress Series*, **558**, 193–206.

Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences*, **101**, 975–979.

Excoffier L, Foll M, Petit RJ (2009) Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 481–501.

Ferreira CEL, Luiz OJ, Floeter SR *et al.* (2015) First record of invasive lionfish (Pterois volitans) for the Brazilian coast (B Ruttenberg, Ed,). *PLoS ONE*, **10**, e0123002–5.

Fogg AQ, Hoffmayer ER, Driggers WB III (2013) Distribution and length frequency of invasive lionfish (Pterois sp.) in the northern Gulf of Mexico. *Gulf and Caribbean ….*

Freshwater DW, Hines A, Parham S *et al.* (2009) Mitochondrial control region sequence analyses indicate dispersal from the US East Coast as the source of the invasive Indo-Pacific lionfish Pterois volitans in the Bahamas. *Marine Biology*, **156**, 1213–1221.

Galindo HM, Olson DB, Palumbi SR (2006) Seascape genetics: A coupled oceanographic-genetic model predicts population structure of Caribbean corals. *Current Biology*, **16**, 1622–1626.

Gracia E, Botella F, Anadon JD *et al.* (2013) Surfing in tortoises? Empirical signs of genetic structuring owing to range expansion. *Biology Letters*, **9**, 20121091–20121091.

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195,** 205-220.

Hallatschek O, Nelson DR (2008) Gene surfing in expanding populations. *Theoretical Population Biology*, **73**, 158–170.

Hancock AM, Witonsky DB, Ehler E *et al.* (2010) Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences*, **107**, 8924–8930.

Harley CDG, Randall Hughes A, Hultgren KM *et al.* (2006) The impacts of climate change in coastal marine systems. *Ecology Letters*, **9**, 228–241.

Herrera S, Reyes-Herrera PH, Shank TM (2015a) Predicting RAD-seq Marker numbers across the eukaryotic tree of life. *Genome Biology and Evolution*, **7**, 3207–3225.

Herrera S, Watanabe H, Shank TM (2015b) Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Molecular Ecology*, **24**, 673–689.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hewitt GM (1999) Post-glacial re-colonization of European biota. *Biological journal of the Linnean Society*, **68**, 87–112.

Hixon MA, Green SJ, Albins MA, Akins JL, Morris JA Jr (2016) Lionfish: a major marine invasion. *Marine Ecology Progress Series*, **558**, 161–165.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

Johnson J, Bird CE, Johnston MA, Fogg AQ, Hogan JD (2016) Regional genetic structure and genetic founder effects in the invasive lionfish: comparing the Gulf of Mexico, Caribbean and North Atlantic. *Marine Biology*, **163**, 1–7.

Johnston MW, Purkis SJ (2015) Hurricanes accelerated the Florida-Bahamas lionfish invasion. *Global Change Biology*, **21**, 2249–2260.

Kearse M, Moir R, Wilson A *et al.* (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Kirk H, Dorn S, Mazzi D (2013) Molecular genetics and genomics generate new insights into invertebrate pest invasions. *Evolutionary Applications*, **6**, 842–856.

Klopfstein S (2005) The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, **23**, 482–490.

Kulbicki M, Beets J, Chabanet P *et al.* (2012) Distributions of Indo-Pacific lionfishes Pterois spp. in their native ranges: implications for the Atlantic invasion. *Marine Ecology Progress Series*, **446**, 189–205.

Lessler J, Chaisson LH, Kucirka LM *et al.* (2016) Assessing the global threat from Zika virus. *Science*, **353**, aaf8160–aaf8160.

Liu B-J, Zhang B-D, Xue D-X, Gao T-X, Liu J-X (2016) Population structure and adaptive divergence in a high gene flow marine fish: the small yellow croaker (Larimichthys polyactis) (Y-G Yao, Ed,). *PLoS ONE*, **11**, e0154020–16.

Lombaert E, Estoup A, Facon B *et al.* (2014) Rapid increase in dispersal during range expansion in the invasive ladybird Harmonia axyridis. *Journal of Evolutionary Biology*, **27**, 508–517.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FSToutlier tests. *Molecular Ecology*, **23**, 2178–2192.

Lowry E, Rollinson EJ, Laybourn AJ *et al.* (2013) Biological invasions: a field synopsis, systematic review, and database of the literature. *Ecology and Evolution*, **3**, 182–196.

Maggs CA, Castilho R, Foltz D *et al.* (2008) Evaluating signatures of glacial refugia for North Atlantic benthic marine taxa. *Ecology*, **89**.

Merz C, Catchen JM, Hanson-Smith V *et al.* (2013) Replicate phylogenies and post-glacial range expansion of the pitcher-plant mosquito, Wyeomyia smithii, in North America (WJ Etges, Ed,). *PLoS ONE*, **8**, e72262–8.

Morris J (2009) The biology and ecology of the invasive Indo-Pacific lionfish. Doctoral Dissertation.

Morris JA Jr, Akins JL (2009) Feeding ecology of invasive lionfish (Pterois volitans) in the Bahamian archipelago. *Environmental Biology of Fishes*, **86**, 389–398.

Newsted JL, Giesy JP (2000) Epidermal growth factor receptor-protein kinase interactions in hepatic membranes of rainbow trout (Oncorhynchus mykiss). *Fish Physiology and Biochemistry*, **22**, 181–189.

Parmesan C, Yohe G (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, **421**, 37–42.

Partridge CG, MacManes MD, Knapp R, Neff BD (2016) Brain transcriptional profiles of male alternative reproductive tactics and females in bluegill sunfish (H Wang, Ed,). *PLoS ONE*, **11**, e0167509–21.

Peischl S, Excoffier L (2015) Expansion load: recessive mutations and the role of standing genetic variation. *Molecular Ecology*, **24**, 2084–2094.

Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L (2013) On the accumulation of deleterious mutations during range expansions. *Molecular Ecology*, **22**, 5972–5982.

Perry AL (2005) Climate change and distribution shifts in marine fishes. *Science*, **308**, 1912–1915.

Peter BM, Slatkin M (2013) Detecting range expansions from genetic data. *Evolution*, **67**, 3274–3289.

Phillips BL, Brown GP, Shine R (2010) Life-history evolution in range-shifting populations. *Ecology*, **91**, 1617–1627.

Pinsky ML, Worm B, Fogarty MJ, Sarmiento JL, Levin SA (2013) Marine taxa track local climate velocities. *Science*, **341**, 1239–1242.

Price AL, Patterson NJ, Plenge RM *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

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

Pujolar JM, Jacobsen MW, Als TD *et al.* (2014a) Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology*, **23**, 2514–2528.

Pujolar JM, Jacobsen MW, Als TD *et al.* (2014b) Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology*, **23**, 2514–2528.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raj A, Stephens M, Pritchard JK (2013) *Variational Inference of Population Structure in Large SNP Datasets*. Cold Spring Harbor Labs Journals.

Ramachandran S, Deshpande O, Roseman CC *et al.* (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences*, **102**, 15942–15947.

Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM (2013) Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology*, **22**, 2953–2970.

Riedel G, Platt B, Micheau J (2003) Glutamate receptor function in learning and memory. *Behavioural brain research*.

Rius M, Bourne S, Hornsby HG, Chapman MA (2015) Applications of next-generation sequencing to the study of biological invasions. *Current Zoology*, **61**, 488–504.

Schofield P (2010) Update on geographic spread of invasive lionfishes (Pterois volitans [Linnaeus, 1758] and P. miles [Bennett, 1828]) in the Western North Atlantic Ocean, Caribbean Sea and Gulf of Mexico. *Aquatic Invasions*, **5**, S117–S122.

Schofield PJ (2009) Geographic extent and chronology of the invasion of non-native lionfish (Pterois volitans [Linnaeus 1758] and P. miles [Bennett 1828]) in the Western North …. *Aquatic Invasions*.

See KE, Feist BE (2009) Reconstructing the range expansion and subsequent invasion of introduced European green crab along the west coast of the United States. *Biological Invasions*, **12**, 1305–1318.

Shum P, Pampoulie C, Kristinsson K, Mariani S (2015) Three-dimensional post-glacial expansion and diversification of an exploited oceanic fish. *Molecular Ecology*, **24**, 3652–3667.

Silva G, Horne JB, Castilho R (2014) Anchovies go north and west without losing diversity: post-glacial range expansions in a small pelagic fish (C Maggs, Ed,). *Journal of Biogeography*, **41**, 1171–1182.

Simpson SD, Harrison HB, Claereboudt MR, Planes S (2014) Long-distance dispersal via ocean currents connects Omani clownfish populations throughout entire species range. *PLoS ONE*, **9**, e107610–7.

Tan W, Aizen J, Thomas P (2014) Membrane progestin receptor alpha mediates progestin-induced sperm hypermotility and increased fertilization success in southern flounder (Paralichthys lethostigma). *General and comparative endocrinology*, **200**, 18–26.

Tepolt CK, Palumbi SR (2015) Transcriptome sequencing reveals both neutral and adaptive genome dynamics in a marine invader. *Molecular Ecology*, **24**, 4145–4158.

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

Thomas CD, Bodsworth EJ, Wilson RJ, Simmons AD (2001) Ecological and evolutionary processes at expanding range margins. *Nature*, **411**, 577–581.

Toledo-Hernández C, Vélez-Zuazo X, Ruiz-Diaz CP *et al.* (2014) Population ecology and genetics of the invasive lionfish in Puerto Rico. *Aquatic Invasions*, **9**, 227–237.

Travis JM, Munkemuller T, Burton OJ *et al.* (2007) Deleterious mutations can surf to high densities on the wave front of an expanding Population. *Molecular Biology and Evolution*, **24**, 2334–2343.

Tubbs C, Pace M, Thomas P (2010) Expression and gonadotropin regulation of membrane progestin receptor alpha in Atlantic croaker (Micropogonias undulatus) gonads: Role in gamete maturation. *General and comparative endocrinology*, **165**, 144–154.

van Etten, J (2015). gdistance: Distances and Routes on Geographical Grids R package version 1.1-9. https://CRAN.R-project.org/package=gdistance

Vivanco I, Sawyers CL (2002) The phosphatidylinositol 3-Kinase–AKT pathway in human cancer. *Nature Reviews Cancer*, **2**, 489–501.

Volis S, Ormanbekova D, Yermekbayev K, Song M, Shulgina I (2014) Introduction beyond a species range: a relationship between population origin, adaptive potential and plant performance. **113**, 268–276.

White C, Selkoe KA, Watson J *et al.* (2010) Ocean currents help explain population genetic structure. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 1685–1694.

White TA, Perkins SE, Heckel G, Searle JB (2013) Adaptive evolution during an ongoing range expansion: the invasive bank vole (*Myodes glareolus*) in Ireland. *Molecular Ecology*, **22**, 2971–2985.

Zhu Y, Rice CD, Pang Y, Pace M, Thomas P (2003) Cloning, expression, and characterization of a membrane progestin receptor and evidence it is an intermediary in meiotic maturation of fish oocytes. *Proceedings of the National Academy of Sciences*, **100**, 2231–2236.

Chapter 4 Appendices

Appendix I. Sample collection information and size distributions

Samples were collected from a variety of locations, habitat types, and depths within each region. Ages of fish used in RAD-sequencing were estimated using a growth curve as described in the methods and the likely year of recruitment was calculated from age (in days) and date of collection (Figure I.1). The locations used to calculate distances are presented in Table I.1. The sample location, length (standard and/or total), sex, weight, latitude and longitude of location collected, habitat type, collection date, and depth if available is presented in Table I.2 for all collected and used individuals.

Figure I.1. (A) Lionfish ages calculated from length measurements, and (B) likely recruitment years for samples included in the RAD-seq portion of the study.

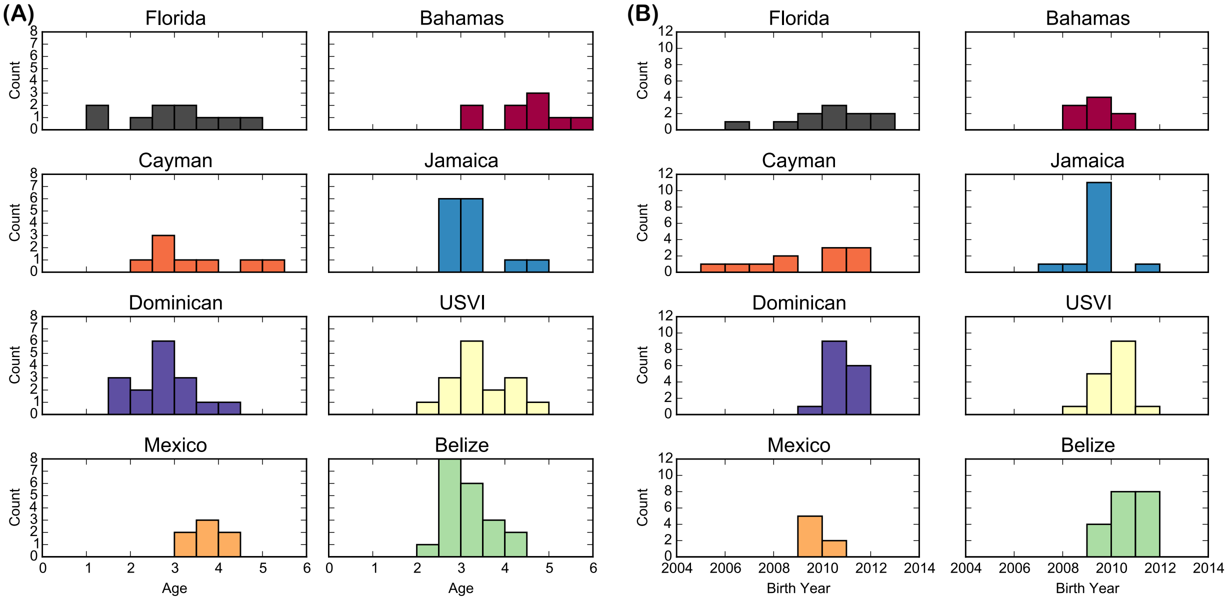


Table I.1. Latitude and longitude used in distance calculations (often the most common location of collection for individual fish).

Location	Dive site name	Latitude (N)	Longitude (W)
Florida, USA	Biscayne Bay	25.5662	-80.0906
The Bahamas	“Ron’s Revenge”	24.5213	-76.2153
Jamaica	“English Reef”	17.8728	-77.7654
Dominican Republic	Bayahibe	18.3431	-68.8338
Grand Cayman	“Pedro’s Castle”	19.2615	-81.2800
US Virgin Islands	Buck Island 1	17.7824	-64.6198
Mexico	Cozumel Marine Park	20.4547	-86.9922
Honduras	Roatan Island	16.2106	-86.3241
Belize	Ambergris Caye	18.1197	-87.8221

Table I.2. Sample information for each individual used in the RAD-sequencing portion of this study.

Sample ID	General Site	Collection Site Name	Standard Length (mm)	Total Length (mm)	Sex (M/F)	Weight (g)	Habitat type	Collection date	Latitude	Longitude	Depth (ft)
EKB_026	Biscayne Bay, FL	Bluefire	315	NA	M		Wreck	8/29/2013	25°33'58.56"N	80° 5'26.28"W	110-120
EKB_030	Biscayne Bay, FL	Bluefire	265	NA	M		Wreck	8/29/2013	25°33'58.56"N	80° 5'26.28"W	110-120
EKB_031	Biscayne Bay, FL	Bluefire	229	NA	M		Wreck	8/29/2013	25°33'58.56"N	80° 5'26.28"W	110-120
EKB_032	Biscayne Bay, FL	Bluefire	235	NA	M		Wreck	8/29/2013	25°33'58.56"N	80° 5'26.28"W	110-120
EKB_033	Biscayne Bay, FL	Bluefire	222	NA	F		Wreck	8/29/2013	25°33'58.56"N	80° 5'26.28"W	110-120
EKB_034	Biscayne Bay, FL	Bluefire	199	NA	M		Wreck	8/29/2013	25°33'58.56"N	80° 5'26.28"W	110-120
EKB_035	Biscayne Bay, FL	Bluefire	184	NA	M		Wreck	8/29/2013	25°33'58.56"N	80° 5'26.28"W	110-120
EKB_036	Biscayne Bay, FL	Bluefire	103	NA	U		Wreck	8/29/2013	25°33'58.56"N	80° 5'26.28"W	110-120
EKB_037	Biscayne Bay, FL	Bluefire	99	NA	U		Wreck	8/29/2013	25°33'58.56"N	80° 5'26.28"W	110-120
EKB_038	Biscayne Bay, FL	Bluefire	257	NA	U		Wreck	8/29/2013	25°33'58.56"N	80° 5'26.28"W	110-120
EKB_040	Biscayne Bay, FL	Long Reef (Drift Dive)	285	NA	M		Deep Reef	8/29/2013	25°26.620' N	80°06.912' W	70-80
EKB-361	USVI (2013)	P3-8	195	NA	M		Patch Reef	2/17/2013	17.78228	-64.60800	
EKB-362	USVI (2013)	LF5-4	264	NA	F		Continuous Reef	5/12/2013	17.80367	-64.63655	
EKB-364	USVI (2013)	LF4-10	245	NA	M		Continuous Reef	2/17/2013	17.79705	-64.63998	
EKB-365	USVI (2013)	I2-5	239	NA	F		Fringing Reef	5/11/2013	17.78340	-64.61470	
EKB-367	USVI (2013)	I2-6	271	NA	F		Fringing Reef	5/11/2013	17.78340	-64.61470	
EKB-368	USVI (2013)	I1-1	219	NA	M		Fringing Reef	5/12/2013	17.78241	-64.61976	
EKB-369	USVI (2013)	P3-1	267	NA	F		Patch Reef	5/11/2013	17.78228	-64.60800	
EKB-373	USVI (2013)	I1-1	203	NA	F		Fringing Reef	10/26/2013	17.78241	-64.61976	
EKB-374	USVI (2013)	LF5-3	286	NA	M		Continuous Reef	8/25/2013	17.80367	-64.63655	
EKB-375	USVI (2013)	P5-1	244	NA	M		Patch Reef	8/24/2013	17.77763	-64.59573	
EKB-378	USVI (2013)	LF5-1	198	NA	U		Continuous Reef	6/22/2013	17.80367	-64.63655	
EKB-379	USVI (2013)	P3-3	242	NA	M		Patch Reef	10/26/2013	17.78228	-64.60800	
EKB-380	USVI (2013)	I2-1	241	NA	M		Fringing Reef	8/24/2013	17.78340	-64.61470	
EKB-381	USVI (2013)	LF5-1	229	NA	M		Continuous Reef	8/25/2013	17.80367	-64.63655	
EKB-382	USVI (2013)	P3-1	246	NA	M		Patch Reef	10/26/2013	17.78228	-64.60800	
EKB-383	USVI (2013)	P1-1	250	NA	I		Patch Reef	6/22/2013	17.77933	-64.61403	
EKB-406	Mexico (2013)	Cozumel Marine Park	220.0	302.0	m		Hard Bottom	8/25/2013	20.4546540	-86.99218	100
EKB-407	Mexico (2013)	Cozumel Marine Park	255.0	350.0	m		Hard Bottom	8/25/2013	20.4546540	-86.99218	100
EKB-410	Mexico (2013)	Cozumel Marine Park	258.0	346.0	m		Hard Bottom	8/25/2013	20.4546540	-86.99218	100
EKB-412	Mexico (2013)	Cozumel Marine Park	255.0	340.0	m		Hard Bottom	8/25/2013	20.4546540	-86.99218	100
EKB-415	Mexico (2013)	Cozumel Marine Park	260.0		m		Hard Bottom	8/25/2013	20.4546540	-86.99218	100
EKB-416	Mexico (2013)	Cozumel Marine Park	242.0	312.0	f		Hard Bottom	8/25/2013	20.4546540	-86.99218	100
EKB-419	Mexico (2013)	Cozumel Marine Park	250.0	335.0	m		Hard Bottom	8/25/2013	20.4546540	-86.99218	100
EKB-421	Honduras (2013)	Roatan Island, Odessey Wreck	NA	NA	U		Wreck / Artificial Reef	10/31/2013	16.2105690	-86.3240890	140
EKB-422	Honduras (2013)	Roatan Island, Odessey Wreck	NA	NA	U		Wreck / Artificial Reef	10/31/2013	16.2105690	-86.3240890	140
EKB-423	Honduras (2013)	Roatan Island, Odessey Wreck	NA	NA	U		Wreck / Artificial Reef	10/31/2013	16.2105690	-86.3240890	140
EKB-424	Honduras (2013)	Roatan Island, Odessey Wreck	NA	NA	U		Wreck / Artificial Reef	10/31/2013	16.2105690	-86.3240890	140

EKB-426	Honduras (2013)	Roatan Island, Odessey Wreck	NA	NA	U		Wreck / Artificial Reef	10/31/2013	16.2105690	-86.3240890	140
EKB-427	Honduras (2013)	Roatan Island, Odessey Wreck	NA	NA	U		Wreck / Artificial Reef	10/31/2013	16.2105690	-86.3240890	140
EKB-429	Honduras (2013)	Roatan Island, Odessey Wreck	NA	NA	U		Wreck / Artificial Reef	10/31/2013	16.2105690	-86.3240890	140
EKB-431	Honduras (2013)	Roatan Island, Odessey Wreck	NA	NA	U		Wreck / Artificial Reef	10/31/2013	16.2105690	-86.3240890	140
EKB-433	Honduras (2013)	Roatan Island, Odessey Wreck	NA	NA	U		Wreck / Artificial Reef	10/31/2013	16.2105690	-86.3240890	140
EKB-434	Honduras (2013)	Roatan Island, Odessey Wreck	NA	NA	U		Wreck / Artificial Reef	10/31/2013	16.2105690	-86.3240890	140
EKB-436	Honduras (2013)	Mangrove Bight, Guanaja Island	NA	NA	U		Coral Reef	10/23/2013	16.313200	-85.525100	40
EKB-438	Honduras (2013)	Mangrove Bight, Guanaja Island	NA	NA	U		Coral Reef	10/23/2013	16.313200	-85.525100	40
EKB-439	Honduras (2013)	Mangrove Bight, Guanaja Island	NA	NA	U		Coral Reef	10/23/2013	16.313200	-85.525100	40
EKB-441	Honduras (2013)	Mangrove Bight, Guanaja Island	NA	NA	U		Coral Reef	10/23/2013	16.313200	-85.525100	40
EKB-442	Honduras (2013)	Mangrove Bight, Guanaja Island	NA	NA	U		Coral Reef	10/23/2013	16.313200	-85.525100	40
EKB-443	Honduras (2013)	Mangrove Bight, Guanaja Island	NA	NA	U		Coral Reef	10/23/2013	16.313200	-85.525100	40
EKB-445	Bhamas (2013)	barge	320.0	386.0	M	610.00	Coral Reef	10/23/2013	NA	NA	50
EKB-446	Bhamas (2013)	Ron's Revenge	278.0	346.0	M	627.00	Coral Reef	10/10/2013	24.521313	-76.215259	30
EKB-447	Bhamas (2013)	barge	280.0	370.0	M	581.00	Coral Reef	10/23/2013	NA	NA	50
EKB-451	Bhamas (2013)	Tunnel Rock	226	310.0	F	370.00	Coral Reef	10/2/2013	NA	NA	35
EKB-452	Bhamas (2013)	Ron's Revenge	281	362.0	M	479.00	Coral Reef	10/10/2013	24.521313	-76.215259	30
EKB-457	Bhamas (2013)	High rock	288.0	375.0	M	603.00	Coral Reef	10/16/2013	NA	NA	35
EKB-459	Bhamas (2013)	barge	289.0	358.0	M	548.00	Coral Reef	10/23/2013	NA	NA	50
EKB-460	Bhamas (2013)	barge	279.0	364.0	M	550.00	Coral Reef	10/23/2013	NA	NA	50
EKB-461	Bhamas (2013)	Random Reef	225.0	295.0	M	315.00	Coral Reef	10/10/2013	24.533505	-76.241298	15
EKB-464	Jamaica (2012)	Dairy Bull	222	296	M	372	Coral Reef	7/17/12	18.473376	-77.387518	NA
EKB-468	Jamaica (2012)	East Fore Reef	267	353	M	464	Coral Reef	7/13/12	NA	NA	NA
EKB-471	Jamaica (2012)	English Reef	221	294	M	290	Coral Reef	7/29/12	NA	NA	88
EKB-472	Jamaica (2012)	Fish Pot (Bluefields)	194	265	F	500	Coral Reef	7/28/12	NA	NA	NA
EKB-473	Jamaica (2012)	Fish Pot (Bluefields)	205	272	F	500	Coral Reef	7/28/12	NA	NA	NA
EKB-474	Jamaica (2012)	Fish Pot (Bluefields)	202	281	M	550	Coral Reef	7/26/12	NA	NA	NA
EKB-475	Jamaica (2012)	English Reef	206	269	F	250	Coral Reef	7/29/12	NA	NA	88
EKB-477	Jamaica (2012)	English Reef	228	307	M	450	Coral Reef	7/29/12	NA	NA	88
EKB-478	Jamaica (2012)	English Reef	226	306	M	310	Coral Reef	7/29/12	NA	NA	88
EKB-479	Jamaica (2012)	East Fore Reef	227	305	M	350	Coral Reef	7/13/14	NA	NA	NA
EKB-480	Jamaica (2012)	Fish Pot (Bluefields)	267	364	M	850	Coral Reef	7/26/12	NA	NA	NA
EKB-483	Jamaica (2012)	Dairy Bull	231	311	M	390	Coral Reef	7/17/12	NA	NA	80
EKB-484	Jamaica (2012)	English Reef	212	284	M	300	Coral Reef	7/29/12	NA	NA	88

EKB-485	Jamaica (2012)	Fish Pot (Bluefields)	195	265	F	500	Coral Reef	7/26/12	NA	NA	NA
EKB-489	Belize (2013)	East Snake, Port Honduras Marine Reserve	248.0	324.0	F	450.00	Coral Reef	8/9/13	16.207774	-88.508053	45
EKB-491	Belize (2013)	Chub Hole, Bacalar Chico Marine Reserve	240.0	310.0	M	442.00	Fringing Reef with many crevices	10/30/13	18.119658	-87.822099	60
EKB-492	Belize (2013)	Remora's Revenge, Caye Chapel	237.0	309.0	M	389.00	spur and groove	10/31/13	17.668380	-88.013850	66
EKB-493	Belize (2013)	Chub Hole, Bacalar Chico Marine Reserve	265.0	345.0	M	1000.00	spur and groove	10/26/13	18.155800	-87.821240	70
EKB-497	Belize (2013)	Remora's Revenge, Caye Chapel	189.0	246.0	F	173.00	spur and groove	10/31/13	17.668380	-88.013850	66
EKB-498	Belize (2013)	Remora's Revenge, Caye Chapel	210.0	280.0	F	190.00	spur and groove	10/31/13	17.668380	-88.013850	66
EKB-500	Belize (2013)	Sand Trap, Caye Chapel	205.0	264.0	M	197.00	spur and groove	10/31/13	17.691630	-88.008660	52
EKB-501	Belize (2013)	Sand Trap, Caye Chapel	211.0	278.5	M	232.00	spur and groove	10/31/13	17.691630	-88.008660	52
EKB-502	Belize (2013)	Chub Hole, Bacalar Chico Marine Reserve	270.0	350.0	M	604.00	Fringing Reef with many crevices	10/30/13	18.119658	-87.822099	60
EKB-503	Belize (2013)	Chub Hole, Bacalar Chico Marine Reserve	205.0	280.0	F	226.00	Fringing Reef with many crevices	10/30/13	18.119658	-87.822099	60
EKB-504	Belize (2013)	Chub Hole, Bacalar Chico Marine Reserve	225.0	280.0	F	337.00	Fringing Reef with many crevices	10/30/13	18.119658	-87.822099	60
EKB-506	Cayman (2013)	LC SPAG, Little Cayman	192.0	255.0	NA	250.00	wall patch	10/21/13	19.65190	80.10772	95
EKB-507	Cayman (2013)	LC SPAG, Little Cayman	203.0	263.0	NA	250.00	wall patch	10/21/13	19.65190	80.10772	95
EKB-508	Cayman (2013)	LC SPAG, Little Cayman	195.0	263.0	NA	250.00	wall patch	10/21/13	19.65190	80.10772	95
EKB-509	Cayman (2013)	Spotts Beach Mooring, Grand Cayman	315.0	410.0	NA	850.00	Wall	10/12/13	19.26862	81.31138	75
EKB-510	Cayman (2013)	Pedro Reef, Grand Cayman	303.0	--	NA	850.00	spur & groove	10/12/13	19.26657	81.29439	30
EKB-511	Cayman (2013)	East of Black Forest, Grand Cayman	NA	282.0	NA	700.00	wall	10/12/13	19.27234	81.39530	80
EKB-512	Cayman (2013)	Pedro Castle, Grand Cayman	287.0	370.0	NA	650.00	spur & groove	10/13/13	19.26148	81.28001	50
EKB-513	Cayman (2013)	Pedro Castle, Grand Cayman	290.0	384.0	NA	800.00	spur & groove	10/13/13	19.26148	81.28001	50
EKB-514	Cayman (2013)	Pilot Wrek, Grand Cayman	230.0	300.0	NA	350.00	wreck	10/13/13	19.36749	81.36927	10
EKB-515	Cayman (2013)	NE Coast, Cayman Brac	318.0	412.0	NA	850.00	wall	10/21/13	19.74171	79.78287	95
EKB-516	Cayman (2013)	NE Coast, Cayman Brac	238.0	326.0	NA	400.00	wall	10/21/13	19.74171	79.78287	95
EKB-528	Belize (2013)	Goliath, Bacalar Chico Marine Reserve	227.0	295.0	F	700.00	spur & groove	10/26/13	18.155800	-87.821240	70
EKB-529	Dominican Republic (2013)	La Caleta MPA	270.0	210.0	NA	246.40	Coral Reef	10/5/13	18.441100	69.694067°	80
EKB-530	Dominican Republic (2013)	La Caleta MPA	310.0	240.0	NA	336.90	Coral Reef	10/5/13	18.441100	69.694067°	80
EKB-531	Dominican Republic (2013)	SOSUA	260	350	NA	20onz	Coral Reef	10/20/13	19° 46.109'N	70° 33.688'W	80
EKB-532	Dominican Republic (2013)	La Caleta MPA	260.0	200.0	NA	210.10	Coral Reef	10/5/13	18.441539	69.693642°	80
EKB-533	Dominican Republic (2013)	La Caleta MPA	270.0	200.0	NA	244.70	Coral Reef	10/5/13	18.441539	69.693642°	80
EKB-534	Dominican Republic (2013)	BAYAHIBE	216.0	285.0	NA	275.00	Patch Reef	9/22/13	18.343100	68.833810	50
EKB-535	Dominican Republic (2013)	BAYAHIBE	224.0	288.0	NA	325.00	Patch Reef	9/22/13	18.343100	68.833810	50
EKB-536	Dominican Republic (2013)	Las Galeras Samana	320.0	240.0	NA	480.00	Coral Reef	10/12/13	19.17.474	069.11.902	108

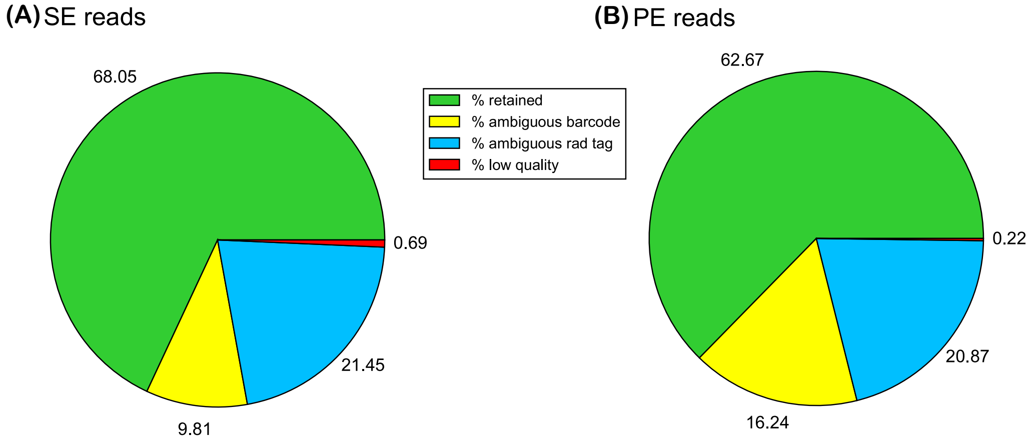
EKB-537	Dominican Republic (2013)	Las Galeras Samana	370.0	290.0	NA	780.00	Coral Reef	10/12/13	19.17.474	069.11.902	108
EKB-538	Dominican Republic (2013)	Las Galeras Samana	330.0	260.0	NA	600.00	Coral Reef	10/12/13	19.17.474	069.11.902	108
EKB-605	Dominican Republic (2013)	SOSUA	230.0	310.0	NA	10onz	Coral Reef	10/20/13	19° 46.109'N	70° 33.688'W	80
EKB-607	Dominican Republic (2013)	SOSUA	260.0	330.0	NA	11onz	Coral Reef	10/20/13	19° 45.519'N	70° 31.230'	100
EKB-608	Dominican Republic (2013)	BAYAHIBE	215.0	282.0	NA	220.00	Patch Reef	9/22/13	18.343100	68.833810	50
EKB-609	Dominican Republic (2013)	La Caleta MPA	400.0	310.0	NA	450.3	Coral Reef	10/5/13	18.441539	69.693642°	80
EKB-610	Dominican Republic (2013)	BAYAHIBE	217.0	285.0	NA	300.00	Patch reef	9/22/13	18.343100	68.833810	50
EKB-611	Dominican Republic (2013)	BAYAHIBE	232.0	297.0	NA	390.00	Patch reef	9/22/13	18.343100	68.833810	50
EKB-612	Belize (2013)	Chub Hole, Bacalar Chico Marine Reserve	220.0	290.0	F	371.00	Fringing Reef with many crevices	10/30/13	18.119658	-87.822099	60
EKB-613	Belize (2013)	East Snake, PHMR, Port Honduras Marine Reserve, Belize	264.0	342.0	M	500.00	Coral Reef	8/9/13	16.207774	-88.508053	45
EKB-616	Belize (2013)	Chub Hole, Bacalar Chico Marine Reserve	210.0	280.0	M	455.00	Fringing Reef with many crevices	10/30/13	18.119658	-87.822099	70
EKB-617	Belize (2013)	Goliath, Bacalar Chico Marine Reserve	240.0	315.0	M	850.00	spur & groove	10/26/13	18.155800	-87.821240	70
EKB-618	Belize (2013)	Goliath, Bacalar Chico Marine Reserve	260.0	340.0	M	800.00	spur & groove	10/26/13	18.155800	-87.821240	70
EKB-619	Belize (2013)	Goliath, Bacalar Chico Marine Reserve	225.0	295.0	F	750.00	spur & groove	10/26/13	18.155800	-87.821240	70
EKB-621	Belize (2013)	Remora's Revenge, Caye Chapel	196.0	261.0	F	215.00	spur & groove	10/31/13	17.668380	-88.013850	66
EKB-624	Belize (2013)	Chub Hole, Bacalar Chico Marine Reserve	220.0	300.0	M	348.00	Fringing Reef with many crevices	10/30/13	18.119658	-87.822099	60

Appendix II: Quality and quantity of RAD-sequencing in the *Stacks* pipeline

Process_radtags program specifics

For the single-end sequence plate (95 samples), Illumina sequencing of the prepared RAD libraries yielded 179,873,518 million reads, 1,238,285 (0.69%) of which were discarded due to low quality, 17,646,010 (9.81%) of which were discarded due to absence of a barcode, and 38,589,757 (21.45%) of which were discarded due to the absence of or ambiguity in the restriction site. After filtering, 122,399,466 (68.05%) were used moving forward in the *Stacks* pipeline (Figure II.1A). For the paired-end RAD-seq plate (25 samples for this study), only the first read of the paired end data are used here as population genomic data. Illumina sequencing of the prepared paired-end RAD libraries yielded 126,382,552 reads, 275,853 (0.22%) of which were discarded due to low quality, 20,527,070 (16.24%) were discarded due to absence of a barcode, and 26,373,050 (20.87%) were discarded due to the absence of or ambiguity in the restriction site. After filtering, 79,206,579 (62.67%) of the total reads were used moving forward in the *Stacks* pipeline (Figure II.1B). For single end Illumina sequencing, for each population, the percentage of reads discarded due to low quality was less than 0.5% and the percentage of reads discarded due to an ambiguous RAD-tag was between 18.2% (for the USVI samples) and 35.5% (for the Mexico samples). For the paired-end Illumina sequencing, for each population, the percentage of reads discarded due to low quality was less than 0.3% and the percentage discarded due to ambiguous barcodes ranged from 18.68% (for The Bahamas samples) and 29.54% (for the Honduras samples).

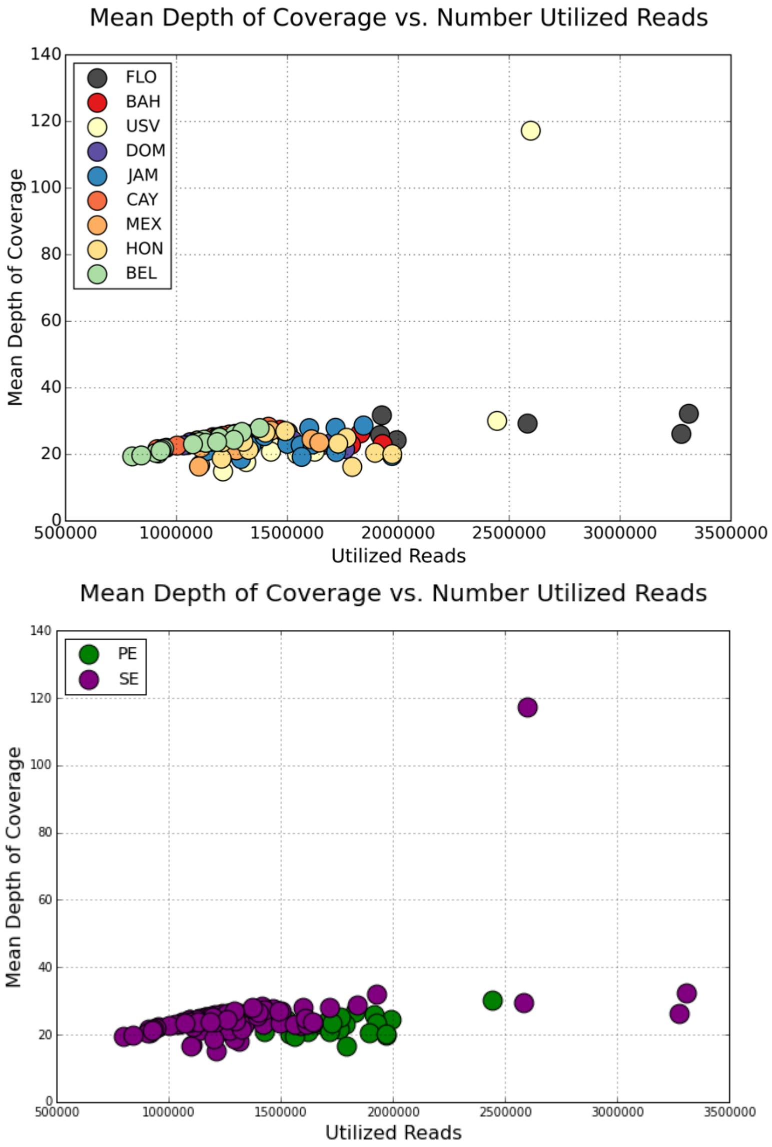
Figure II.1: Total percentages of reads retained and filtered out due to ambiguous barcodes, ambiguous rad tags, or low quality by *process_radtags*. (A) for single end data, (B) for paired end data.



Denovomap.pl and populations programs results

The mean merged depth of coverage remained steady as more reads were added to the analysis, indicating that as reads were added, so were loci (Figure II.2). The sequencing type did not affect the ultimate outcome of depth of coverage, but individuals sequenced with paired-end Illumina sequencing had slightly higher total utilized reads (Figure II.2).

Figure II.2. Mean merged depth of coverage vs. number of utilized reads for each sample shown by population (top panel) and by sequencing type (bottom panel).

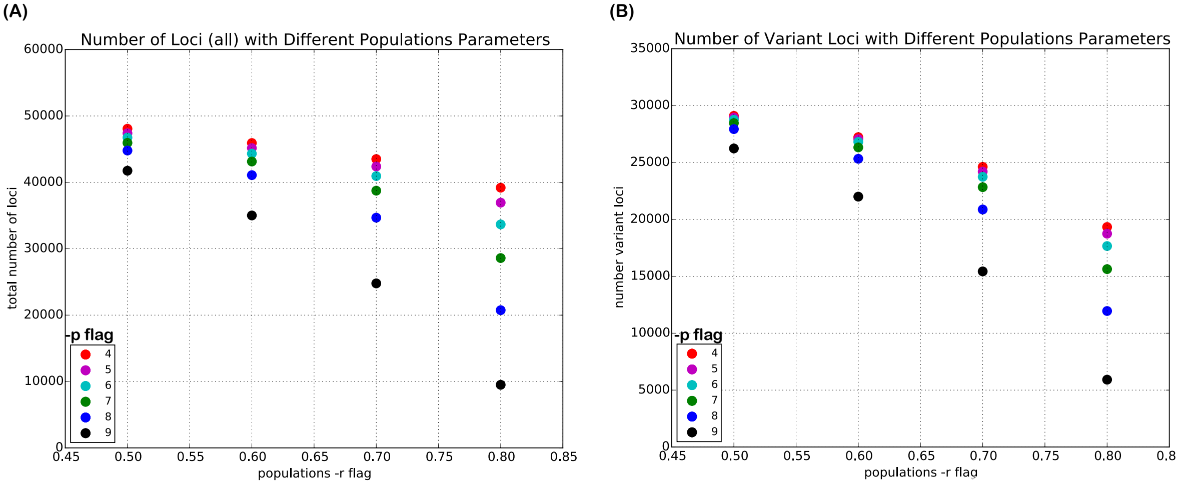


Mean Depth of Coverage vs. Number Utilized Reads

Mean Depth of Coverage vs. Number Utilized Reads

The filtering requirements by *Stacks populations* programs affected the number of loci used in analysis in predictable ways: the more stringent the requirements for loci being shared between populations and individuals, the fewer loci were retained for analysis. The two ways to alter these requirements are with the –p and –r arguments in the program. As the requirements for the number of populations (-p) and percent of individuals within a population (-r) are relaxed, the number of loci used converged on a range between 40,000-50,000 with 25,000 to 30,000 of those being variant loci that informed population genetic statistics (Figure II.3).

Figure II.3. Number of (A) total loci and (B) variant loci analyzed by the *populations* program in *STACKS* for different –p and –r parameters.



Appendix III. Calculations of distance and impact of distance measurements on summary statistics regressions

As described in the Methods section, distance was calculated multiple ways (Table III.1), and summary statistics were regressed against each measurement (Figure III.1). Results for the regressions varied slightly with the "modified-ocean-distance" the best fit regression for the data (Table III.2).

Table III.1. Different calculations of distance (in km) from Florida using (1) Euclidian distance, (2) least-cost distance through the ocean with the only requirement being that the path remain in the water, and (3) least-cost distance through the ocean with the requirement that for sites to the West of Cuba, the path travel around the East side of the island.

Site	Euclidian	Ocean	Modified Ocean
Florida	0	0	0
Bahamas	406	409.02	409.02
Jamaica	885	1303.79	1303.79
Cayman Islands	678	1084.90	1556.56
Dominican Republic	1409	1518.20	1518.20
Mexico	906	889.1	2122.87
Belize	1148	1127.80	2239.69
Honduras	1222	1266.5	2180.68
US Virgin Islands	1816	1816.40	1816.40

Figure III.1. Regressions of observed heterozygosity against Euclidian distance and through-ocean distance measures. The modified ocean distance regression is in Figure 2 of the main chapter text.

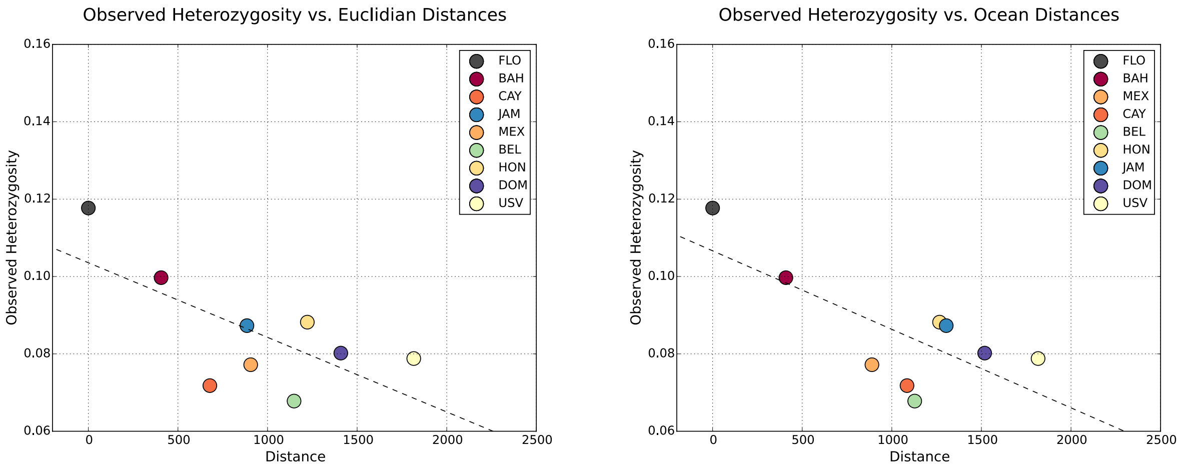


Table III.2. Regression results for observed heterozygosity for multiple distance-regimes.

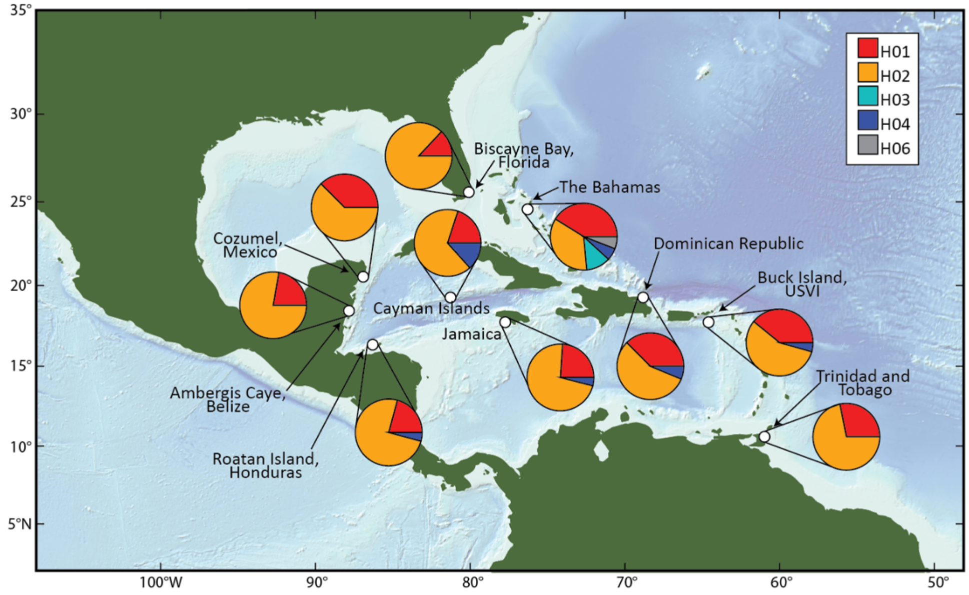
Distance Regime	**R^2**	**p-value**
Euclidian	0.4623	0.044
Ocean	0.5365	0.025
Modified-Ocean	0.7436	0.003

Appendix IV. Results from mitochondrial D-Loop sequencing and analysis

Five haplotypes were sequenced across the nine study sites (Figure IV.1), corresponding to previously reported haplotypes named H01, H02, H03, H04, and H06 (Betancur-R *et al.* 2009). In previous research, nine haplotypes have been described in the entire invaded range, but only four of those have been described in the southern part of the expansion which is the focus of this study (not including The Bahamas, which is often considered to be part of the northern expansion) (Toledo-Hernández *et al*. 2014; Butterfield *et al*. 2015; Johnson *et al*. 2016). H02 is the most common haplotype, both in the present study (n = 146), and in previous studies (*e.g*., Butterfield *et al*. 2015). Haplotype H01 was sequenced in 61 individuals. H03, H04, and H06 were sequenced in 2, 7, and 1 individual(s) respectively.

The four haplotypes that have ever been found south of The Bahamas (in this study and previous studies) are H01, H02, H03, and H04. Contrary to previous findings, haplotype H03 was not found in this study. However, this is not wholly unexpected because in previous studies, H03 was only observed in the Cayman Islands, Puerto Rico, and Panama. Puerto Rico and Panama are not included in this study, and the sample size from the Cayman Islands may be too small to detect a rare haplotype. Therefore, the results of this portion of the study are similar to what has been reported previously. The only exception to expected trends is the low haplotype diversity reported for Florida, where only two haplotypes were sequenced in this study.

Figure IV.1. Distribution of the mtDNA D-Loop haplotypes in the study area.



Appendix V. BayEnv Bayes Factor sorting information

Figure V.1. Number of loci per bin in the BayEnv decimal binning step of identifying the top 5% of Bayes factors. The upper bound of the bin was not included in the bin (*e.g.,* the first bin includes those with a frequency of 0 but not of 0.1)

Bin	Number of loci
0-0.1	4326
0.1-0.2	843
0.2-0.3	486
0.3-0.4	362
0.4-0.5	374
0.5-0.6	363
0.6-0.7	366
0.7-0.8	547
0.8-0.9	841
0.9-1.0	4251

LITERATURE CITED IN THE CHAPTER 4 APPENDICES

Butterfield JSS, Díaz-Ferguson E, Silliman BR *et al.* (2015) Wide-ranging phylogeographic structure of invasive red lionfish in the Western Atlantic and Greater Caribbean. *Marine Biology*, **162**, 773–781.

Johnson J, Bird CE, Johnston MA, Fogg AQ, Hogan JD (2016) Regional genetic structure and genetic founder effects in the invasive lionfish: comparing the Gulf of Mexico, Caribbean and North Atlantic. *Marine Biology*, **163**, 1–7.

Toledo-Hernández C, Vélez-Zuazo X, Ruiz-Diaz CP *et al.* (2014) Population ecology and genetics of the invasive lionfish in Puerto Rico. *Aquatic Invasions*, **9**, 227–237.

CHAPTER 5

Temporal Dynamics of Range Expansion in the Invasive Lionfish, *Pterois volitans*

ABSTRACT

Despite the potential to gain insight through the study of temporal dynamics of range expansions and species invasions, a temporal perspective is largely missing from the empirical range expansion literature. This study describes population genetic patterns at two time points during the invasion of the Indo-Pacific lionfish, *Pterois volitans*, which has invaded both the US Atlantic coast (a northern expansion) and the Caribbean Sea (a southern expansion) over a matter of decades, with the southern expansion commencing as recently as 2004. We utilize individual lionfish samples collected in approximately 2008 and 2013 to represent two time points during this rapid invasion.) We analyzed 1,054 single nucleotide polymorphisms (SNPs) sequenced in 207 individuals across 14 populations from 11 geographic locations at two time points. SNP data were generated using both double-digest restriction enzyme associated DNA sequencing (ddRAD-seq) methods (for the earlier time point) and single enzyme RAD-seq methods (for the later time point). Range expansion signals appeared to be stronger at the later time point, possibly indicating that the forces of oceanographic mixing and dispersal may not disrupt the signals of decreased observed heterozygosity generated by range expansion. While all steps were taken to process the two types of RAD-seq data in the same way, there is still a possibility that different ascertainment biases between the two sequencing methods could be responsible for the variation between observations at the two time points.

INTRODUCTION

The body of theoretical and simulation-based research focused on range expansions has grown dramatically since the early 2000s (Edmonds *et al.* 2004; Travis *et al.* 2007; Hallatschek & Nelson 2008; Excoffier *et al.* 2009; Slatkin & Excoffier 2012). This burgeoning field has developed specific expectations regarding the distribution of genetic diversity during an expansion. Specifically, there is an expectation that genetic diversity will decrease towards a moving range boundary due to genetic drift (Excoffier *et al.* 2009). Patterns of decreased diversity with expansion have been described in several species on different time scales, including in humans during the expansion out of Africa (Ramachandran *et al.* 2005; Sousa *et al.* 2014), although relatively few studies address rapid expansions on decadal time scales. In humans, the genomic signature of expansion is still evident today in the distribution of genetic diversity globally, indicating that the repercussions of expansion can persist long past the original expansion event (Sousa *et al.* 2014).

Patterns of decreased diversity towards a moving range margin are expected in both expansions of species' native ranges due to changing environments, and expansions characteristic of species invasions in which a new region is colonized. Four phases of biological invasions have been described: (1) transport, (2) introduction, (3) establishment, and (4) spread (Blackburn *et al.* 2011). It is the final step, spread, in which range expansion dynamics dominate. In fact, some of the first empirical data pertaining to any form of range expansion have been generated for invasive species (White *et al.* 2013; Tepolt & Palumbi 2015; Bors*,* Chapter 3; Bors*,* Chapter 4). In the invasive bank vole in Ireland, for example, in the wake of a post-establishment range expansion, decreases in observed and expected heterozygosity as well as allelic richness correlate with distance from the point of invasion (White *et al.* 2013). In marine systems, signals of range expansion may be less pronounced, as was reported in Chapter 4 in the lionfish, *Pterois volitans*. Lionfish populations were shown to have decreased observed heterozygosity with increased distance from Florida, but unlike the bank vole, demonstrated steady levels of expected heterozygosity and allelic richness throughout the invaded range (Bors, Chapter 4).

A temporal perspective is largely missing from the nascent empirical range expansion literature. Temporal dynamics could be critical to understanding the ways that range expansions affect population genomics and the duration of resulting genomic signatures. Many of the theoretical models of expansion assume limited dispersal (Hallatschek & Nelson 2008), which may be appropriate in some systems but does not apply to many marine species. Laboratory experiments demonstrating the most striking results of range-expansion related allele surfing occur in systems without any dispersal (Hallatschek *et al.* 2007). Assumptions about limited or negligible dispersal that are made in many of the simulations and models of range expansion, make it difficult to use those models to predict how dispersal and post invasion expansion interact to shape genetic diversity in a natural system. What is the strength and duration of genomic signals of expansion in an invading marine species with high dispersal capabilities? Will signals persist long enough to affect the future evolution of the species or will dispersal act to homogenize genetic signals? Intrinsic to these questions is the tension between the spatial forces of range expansion and the processes responsible for developing and maintaining genetic connectivity between populations, namely migration and larval dispersal. There is a balance between these processes that varies based on species and ecosystem.

Comparisons of genetic data from multiple time points during an invasion are essential for identifying changes in the composition and distribution of genetic diversity through time. For the invasive shrimp *Palaemon macrodactylus*, for example, mitochondrial data from multiple time points allowed for the detection of a second introduction along the US Atlantic coast (Bors, Chapter 3). Genetic studies that have taken a historical or temporal perspective have used amplified fragment length polymorphic markers (Fennell *et al.* 2014), mitochondrial DNA sequencing (Fonseca *et al.* 2010), and microsatellite markers (Chen *et al.* 2010). It is only recently that next generation sequencing technologies paired with reduced representation library preparation techniques for non-model species have started to be used for temporal studies in any marine species. For example, single nucleotide polymorphism (SNP) data have recently been used to track genomic shifts in commercially targeted cod species over the course of approximately 80 years, showing a correlation in genomic changes with temperature and also fishing pressures (Therkildsen *et al.* 2013).

Here, we focus on temporal genetic signals in a highly successful marine invasive species, the Indo-Pacific lionfish, *Pterois volitans*. The lionfish invasion in the Western Atlantic and Caribbean Sea is unprecedented in both rate of geographic spread and ecological damage (Hixon *et al.* 2016). For a complete description of the invasion pathway, please see Chapter 4. The rapid expansion of lionfish represents an opportunity to explore the temporal population dynamics of a highly dispersive invasive marine fish undergoing range expansion. Luckily, because of the ecological and economic impacts of the invasion, removal efforts have been ongoing in many parts of the invaded range. Non-governmental organizations, researchers, and recreational divers have been collecting fish as part of removal efforts, improving the availability of samples throughout the invaded range from multiple time points, thus creating an opportunity to use temporal samples in genomic research. This study presents the first temporal population genomic data using restriction enzyme associated DNA sequencing (RAD-seq) in the lionfish.

The present study uses genomic data from 1,054 single nucleotide polymorphisms (SNPs) generated for lionfish from a total of 14 populations from 11 geographic locations at two time points during the invasion. We use samples from 2007-2009 that represent a time point earlier in the invasion and samples from 2013-2014 that represent a time point approximately 5 years later in the invasion. The two datasets generated from these two time points in the invasion will be called TP1 ("time point 1") and TP2 ("time point 2"). Data are presented for 207 individuals across the 14 populations (North Carolina, Bermuda, The Bahamas, the Cayman Islands, and Mexico from TP1; Florida, The Bahamas, the Dominican Republic, Jamaica, the US Virgin Islands, the Cayman Islands, Mexico, Honduras, and Belize from TP2). Yet another level of complexity exists in the sampling effort: North Carolina and Bermuda were invaded in 2000 as part of the northern expansion of lionfish (before the Caribbean expansion). The expansion up the US east coast will be referred to as the "northern expansion" while the expansion into the Caribbean will be referred to as the "southern expansion." Data presented here represent a re-analysis of data from Chapter 4 (TP2) augmented by additional RAD-seq data generated for TP1.

This study tests three main hypotheses regarding the temporal dynamics of range expansion genomics in the lionfish, *Pterois volitans*. The first hypothesis stipulates that in TP1 data (earlier time point), the northern expansion should have weaker genomic signatures of decreased diversity along the expansion axis than the southern expansion at either TP1 or TP2 due to the strength of the Gulf Stream as a source of oceanographic mixing and the fact that the northern invasion is older, thus giving genomic signatures of range expansion more time to dissipate in the face of other forces driving population dynamics. The second hypothesis stipulates that with time, the genomic signature of range expansion—specifically decreased diversity towards the edge of the invaded range—should weaken. Therefore, we predicted that relationships of decreased observed heterozygosity with distance from Florida like those observed for the later time point (TP2) in Chapter 3 would be stronger earlier in the invasion (TP1). The third hypothesis stipulates that overall diversity in lionfish populations in the Caribbean region should be lower earlier in the invasion (TP1) than later in the invasion (TP2) due to the initial colonization event of the Caribbean followed by an influx of more diversity through time.

METHODS

Sample collection and DNA preparation

Lionfish individuals were used from a total of 14 populations from 11 geographic locations across two time points (Figure 1, Table 1). For the first time point (TP1), lionfish individuals were collected from North Carolina, Bermuda, The Bahamas, and the Cayman Islands by teams of SCUBA divers in 2007 and 2008 and tissue was subsampled and stored at the University of North Carolina, Wilmington. Samples from Mexico were collected in 2009 by teams of SCUBA divers. Subsamples of fin clip tissue from those samples were stored on Whatman Paper (GE Life Sciences, Pittsburgh, PA, USA) and archived in the U.S. National Oceanic and Atmospheric Administration (NOAA) Beaufort Laboratory in Beaufort, North Carolina. Extractions of genomic DNA for TP1 samples were carried out at UNC, Wilmington, using methods described in (Freshwater *et al.* 2009). Samples from TP2 were collected as described in the Chapter 4 Methods.

For TP2 samples as well as the lionfish from Mexico collected in 2009, genomic DNA (gDNA) was extracted from muscle or fin clip tissue using a CTAB and proteinase K digest, a phenol-chloroform purification, and ethanol precipitation method as described in Herrera *et al*. (2015). gDNA was stored in AE buffer from a QIAGEN DNeasy Blood and Tissue extraction kit (Qiagen GmbH, Germany) at 4° C or -20° C until gene amplification and sequencing.

Figure 1. Map of the study region showing sampling locations. Locations with a red star are included in both the TP1 and TP2. Colored contours on the map show the extent of the invasion in the years from 2004-2009, by which point all of the nine sites had been invaded.

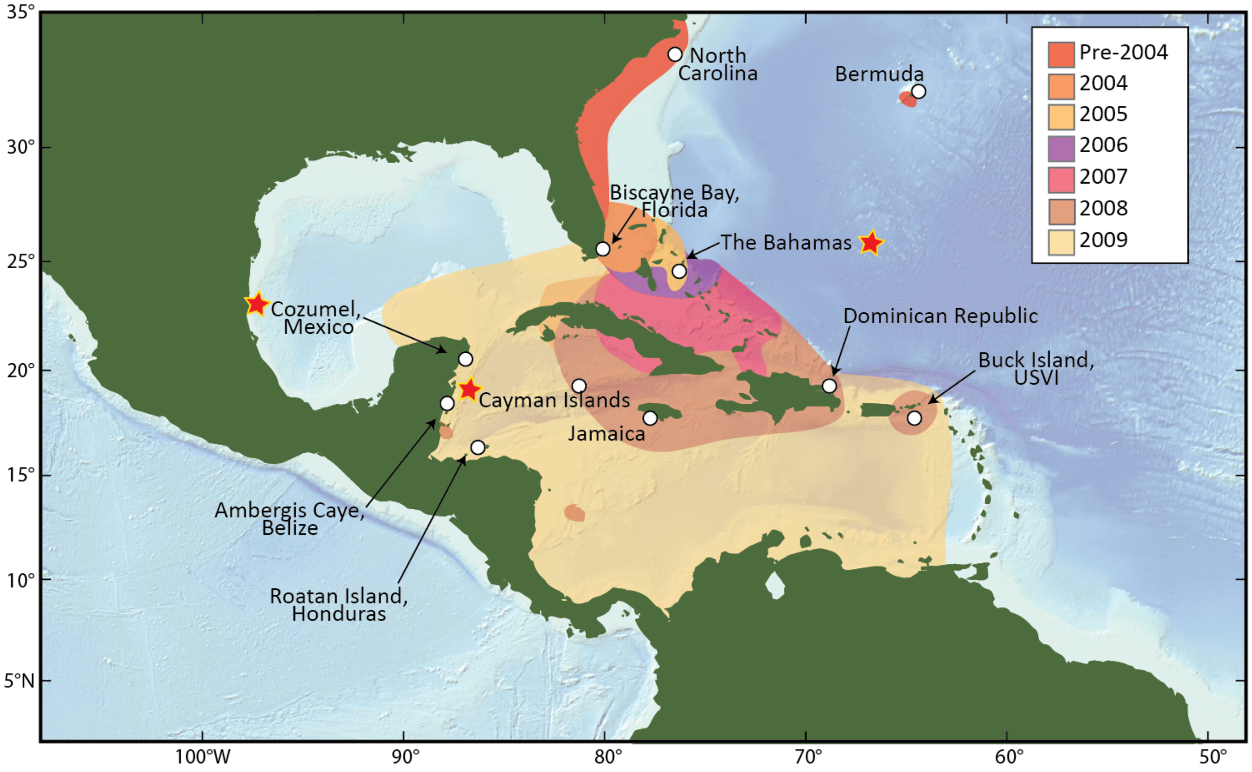


Table 1. Collection locations, years, number of samples collected, latitude, longitude, and distance from Florida (the point of invasion). Sample sites are specified as belonging to either TP1 or TP2.

Location	**Collection Year**	**n**	**Latitude**	**Longitude**	**Distance from FL (km)**
TP1					
Bermuda	2008	16	32.2983	-64.6811	1737.1
North Carolina	2008	15	33.6389	-76.9417	941.71
The Bahamas	2007	22	24.5213	-76.2153	409.02
Cayman Islands	2008	20	19.2615	-81.2800	1556.56
Mexico	2009	13	20.4547	-86.9922	2122.87
TP2					
Florida	2013	11	25.5663	-80.0906	0
The Bahamas	2013	9	24.5213	-76.2153	409.02
Jamaica	2013	14	17.8728	-77.7654	1303.79
Dominican Republic	2013	16	18.3431	-68.8338	1518.20
Cayman Islands	2013	11	19.2615	-81.2800	1556.56
US Virgin Islands	2013, 2014	15	17.7824	-64.6198	1816.40
Mexico	2013	7	20.4547	-86.9922	2122.87
Honduras	2013	16	16.2106	-86.3241	2180.68
Belize	2013	20	18.1197	-87.8221	2239.69

RAD-sequencing and data processing

In order to take advantage of the opportunity to combine datasets to gain a more in depth temporal perspective, genomic data generated by two different types of RAD-sequencing is analyzed together in this study. Both double digest RAD-sequencing (ddRAD-seq) (Peterson *et al.* 2012) and single digest RAD-sequencing (RAD-seq) (Miller *et al.* 2007; Baird *et al.* 2008) were undertaken at two different institutions. For samples undergoing RAD digest with a single enzyme, two different sequencing rounds resulted in some individuals being sequenced with paired-end Illumina sequencing and some with single-end Illumina sequencing (Illumina Inc., San Diego, CA, USA).

Double digest RADseq libraries were prepared using a modified protocol described in Peterson *et al.* (2012). High quality genomic DNA was isolated using a Wizard® *Plus* SV Miniprep DNA Purification kit (Promega, Madison, WI, USA) with suction manifold, and DNA concentrations were quantified using a Qubit 2.0 (Life Technologies) fluorometer. To prepare RADseq libraries, 500 ng of DNA from each sample was double digested with 10 U each of SBfI-HF® (8-cutter) and MspI (4-cutter) (New England Biolabs, Ipswich, MA) at 37 °C for 3 hours. Digested DNA was purified using AMPure (Agencourt) magnetic beads. Thirty-two custom P1-SBfI oligo adapters with unique 6 basepair barcodes were ligated to DNA fragments from each of the respective 32 samples from each site. Indexed custom P2-MspI oligo adapters were annealed to the other end of each fragment. Ligation reactions consisted of 6.25 nM P1 adapter, 0.625 μM P2 adapter, 1.25 mM rATP and 200 U T4 DNA ligase (New England Biolabs), incubated at 4^{o}C over night. Ligation reactions were heat-killed at 65 °C for 10 min, followed by slow (2%) ramp to room temperature.

Samples were divided into three libraries, each with 32 labeled individuals (total of 96) and each library was indexed again before sequencing. Libraries were purified using AMPure (Agencourt) magnetic beads and 2 μl of each DNA pool was amplified using 1x Phusion® High-Fidelity PCR Master Mix (New England Biolabs) and 1 μM of a modified Illumina amplification primer mix (P1-forward: 5’-AATGATACGGC GACCACCG*A-3’ ; P2-reverse: 5’-CAAGCAGAAGACG GCATACG*A-3’). PCR was run as follows: 30 s 98 °C, 30 s 58 °C, 1 min 72 °C, 18 cycles. Following magnetic bead purification of the amplified product, RAD libraries were quantified and sequenced on a single Illumina Hi-Seq 2500 lane by the Duke University Center for Genomic and Computational Biology sequencing center (50 basepair single-end run). For samples in the TP2 dataset and those samples from the ddRAD that were sequenced using a paired end Illumina run, RAD-seq was carried out as described in Chapter 4.

Double digest RAD-seq data were filtered using the *process_radtags* program in *Stacks v. 1.19* (Catchen *et al.* 2013). Reads were filtered for quality with a minimum phred score of 10 in a sliding window of 15% read length (default settings) and sorted by individual-specific barcode. Reads were truncated to 42 basepairs including the 6 basepair restriction site.

To proceed with parallel data for further analyses, specific filtering measures were taken for those samples that were not sequenced with ddRAD methods. In order to directly compare previously sequenced and analyzed RAD-seq data for TP2 presented in Chapter 4, with the ddRAD seq data produced for TP1, the sequencing data from individuals were truncated to 42 nucleotide basepairs in length. Samples from the TP2 that were sequenced using a single enzyme for RAD digest and sequenced with single end Illumina sequencing were initially processed with the *Stacks v. 1.19* RAD-seq analysis software (Catchen *et al.* 2013) *process_radtags* program with default settings as described in Chapter 4. Individual de-multiplexed fastq files were then truncated to 42 basepairs in another run of the *process_radtags* program in *Stacks v. 1.35* (*-t* 42 and *-e* sbfI). The first read of paired-end Illumina sequences were filtered for quality with a minimum phred score of 10 in a sliding window of 15% read length (default settings) and sorted by individual-specific barcode. Reads were truncated to 42 nucleotides. Fastq files for modern samples that were sequenced using paired end Illumina sequencing were concatenated (making one fastq file moving forward to the next processing steps). Processed sequence data for the TP1 that were sequenced with using paired-end Illumina sequencing were concatenated with the 42 basepairs reads for the same individual fish from the ddRAD digest.

Once the sequences for all samples were filtered using the same stringency requirements and trimmed to 42 basepairs, and all duplicate runs of individuals were concatenated, the *Stacks* program *denovomap.pl* was run on the entire dataset (n = 207) using a *stack-depth parameter* (-*m*) of 3, such that three reads were required to generate a stack (*i.e.,* a locus); a *within-individual distance parameter* (-*M*) of 3, allowing for three SNP differences in a read; and a *between-individual distance parameter* (-*n*) of 3, allowing for three fixed differences between individuals to build a locus in the catalog.

Population genomic analyses

Population summary statistics (allele frequencies, observed and expected heterozygosities, π, and F_{IS}) were calculated by the *populations* program in *Stacks* v.1.35, using loci found in all 14 populations and in at least 80% of individuals per population (*-p* 14, *-r* 0.8). For each RAD-tag, only one SNP was used through filtering with the program flag –*write_single_snp* (if there were two or more SNPs in the sequence, *Stacks* would use the first one in analysis). Heterozygosity (observed and expected) values were also calculated in the R Package PopGenKit (https://cran.r-project.org/web/packages/PopGenKit/index.html) to provide secondary validations of reported values. Allelic richness was calculated using PopGenKit. For generating a regression of population genomic summary statistics against distance from Florida, distances were calculated in R with the package 'gdistance' as described in Chapter 4. Distances to Bermuda and North Carolina, however, were calculated as the shortest path through water in the program Google Earth using the ruler tool between two geo-locations.

Three methods were used to describe the genetic structure of lionfish populations in the study area: principal component analysis, a Bayesian structure analysis, and F_{ST} calculations. The *smartpca* program in *EIGENSOFT* (Price *et al*. 2006) was used to perform a principal component analysis (PCA) of genetic diversity. Custom iPython notebooks used to convert *Stacks* PLINK output files into *EIGENSOFT* input files, and for the visualization of the PCA are available at the author's GitHub (https://github.com/ekbors/thesis_scripts). *Smartpca* in *EIGENSOFT* was run with no iterations of outlier removal ('*numoutlieriter'* = 0) with otherwise default parameters. In addition to the PCA analysis, *fastSTRUCTURE* (Pritchard *et al*. 2000; Hubisz *et al*. 2009; Raj *et al*. 2013) was run with the number of genetic lineages (the value of *k*) set to values between one and fourteen to assess genetic structure through a hierarchical analysis, and the program *chooseK.py* was run to select the value of K most consistent with the program's spatial structure model. F_{ST} values were calculated by the *populations* program in *Stacks* using a p-value cutoff of 0.05 and a Bonferroni correction (using the *'bonferroni_gen'* flag in the *populations* program).

RESULTS

RAD-seq and single nucleotide polymorphism calling

After sequence processing by the programs *process_radtags* and *denovomap.pl*, a catalog of 1,611,368 RAD-tags was generated by *Stacks* v. 1.35 (Catchen *et al.* 2013). For population genomic analyses, 1,054 loci were identified that were shared across all 14 populations with a requirement that 80% of the individuals in a population have the locus. This number of loci is an order of magnitude lower than what was used for analysis in Chapter 4 in which longer sequencing reads and only single-digest RAD-seq were used. The double digest RAD-seq generated fewer loci and likely represent a small bottleneck in these analyses. Still, the number of loci is also greatly restricted by the requirement that all loci used be present in 14 of the populations. Using a less stringent filtering requirement would result in more loci for anlaysis.

Spatial and temporal population genomics

Three expansions were characterized in this study: the “northern expansion” in the TP1 dataset, the “southern expansion” in the TP1 dataset, and the “southern expansion” in the TP2 dataset. Observed heterozygosity ranged from 0.054 in the TP2 Cayman Island population to 0.0918 in the TP1 Bahamas population, with consistently lower values for observed heterozygosity in the modern data than the historic data (Table 2). Observed heterozygosity plots for the northern expansion in the TP1 dataset have no clear linear trend with distance from Florida with a slope almost equal to zero (Figure 2). Heterozygosity regressions for both the TP1 and the TP2 for the southern invasion have negative slopes but only the TP2 regression is significant (Figure 2). Both the northern and southern regressions for TP1 include The Bahamas but still, in both cases, only 3 populations were used in the regression analysis. From these limited data, it is difficult to determine exactly what the relationship or regression may be. Therefore, conclusions must be treated as preliminary in nature and final assessment of trends in each dataset will depend on the addition of more data. In summary, there were apparent although not always significant negative relationships between observed heterozygosity and distance from Florida in both theTP1 and the TP2 southern expansions. Allelic richness and expected heterozygosity were both almost constant throughout the range and across the sampled years (Figures 3 and 4).

Table 2. Summary statistics including the number of private alleles, the average number of individuals (averaged across all the loci), the value of the major allele (*p*), observed heterozygosity (H_{OBS}), expected heterozygosity (H_{EXP}), nucleotide diversity (pi), and F_{IS}.

Site (year)	*Private alleles*	*N (avg)*	*p*	*H_{OBS}*	*H_{EXP}*	*Pi*	*F_{IS}*
FLO (2013)	30	10.7412	0.9384	0.0849	0.0869	0.0912	0.0256
BAH (2013)	36	8.7194	0.9342	0.0757	0.0927	0.0984	0.0647
USV (2013)	88	14.0578	0.9368	0.0629	0.0902	0.0936	0.1074
DOM (2013)	53	14.9175	0.9389	0.0629	0.0869	0.0899	0.0949
JAM (2013)	50	13.3128	0.9395	0.0686	0.0859	0.0893	0.0688
CAY (2013)	37	10.2919	0.9411	0.0537	0.0829	0.0872	0.1037
MEX (2013)	40	6.5991	0.9404	0.0586	0.0827	0.0895	0.0828
HON (2013)	49	15.1137	0.9373	0.0672	0.0881	0.0912	0.0763
BEL (2013	61	18.2237	0.9378	0.0549	0.0884	0.0909	0.1266
CAY (2008)	7	19.4047	0.9427	0.0835	0.081	0.0832	0.0054
MEX (2009)	93	11.9374	0.941	0.0782	0.0854	0.0892	0.0348
BAH (2007)	27	21.6474	0.9375	0.0912	0.0918	0.094	0.0136
BER (08)	10	17.8218	0.9395	0.0907	0.0874	0.0899	0.0024
NC (2008)	23	16.6427	0.9361	0.0916	0.0927	0.0956	0.0168

Figure 2. Observed heterozygosity plotted against distance (km) from Florida for both TP1 and TP2 samples. Sites that are part of the TP1 dataset are represented by squares (northern) and diamonds (southern); sites that are part of the TP2 dataset are represented by circles. Heavy dashed line is the regression for TP2 ($R^2 = 0.71$, p-value = 0.004), other regression lines are not shown on this figure because they were not statistically significant.

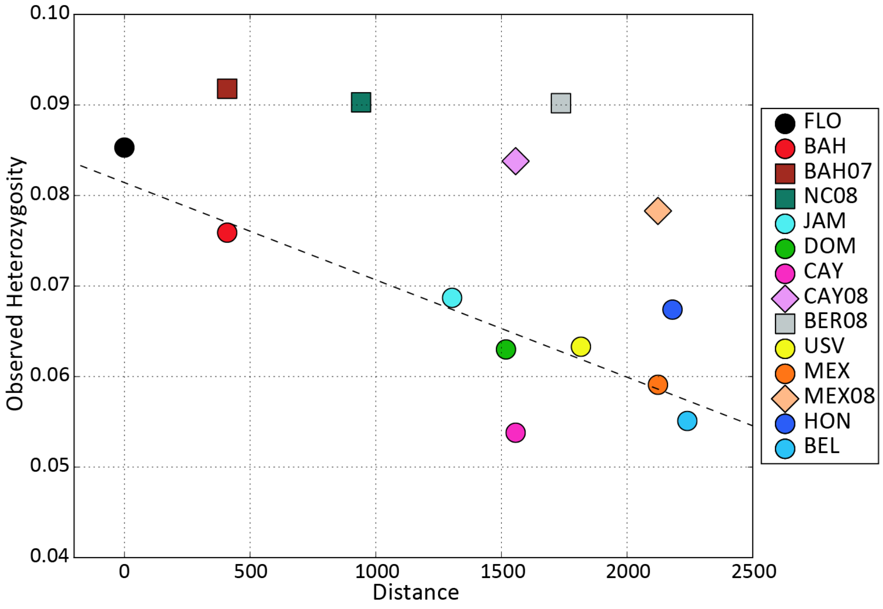


Figure 3. Expected heterozygosity plotted against distance (km) from Florida for both TP1 and TP2 samples. Sites that are part of the TP1 dataset are represented by squares (northern) and diamonds (southern); sites that are part of the TP2 dataset are represented by circles.

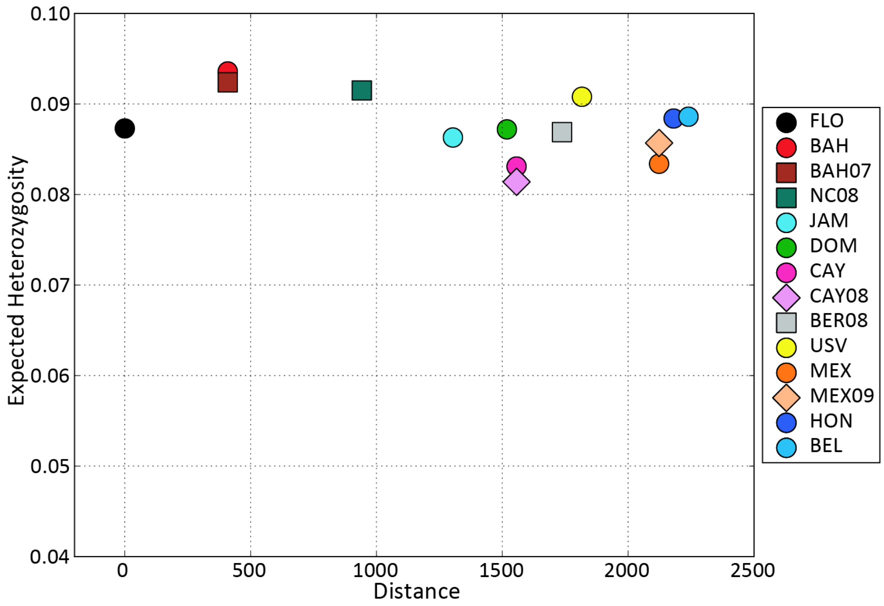


Figure 4. Allelic richness plotted against distance (km) from Florida for both TP1 and TP2 samples. Sites that are part of the TP1 dataset are represented by squares (northern) and diamonds (southern); sites that are part of the TP2 dataset are represented by circles.

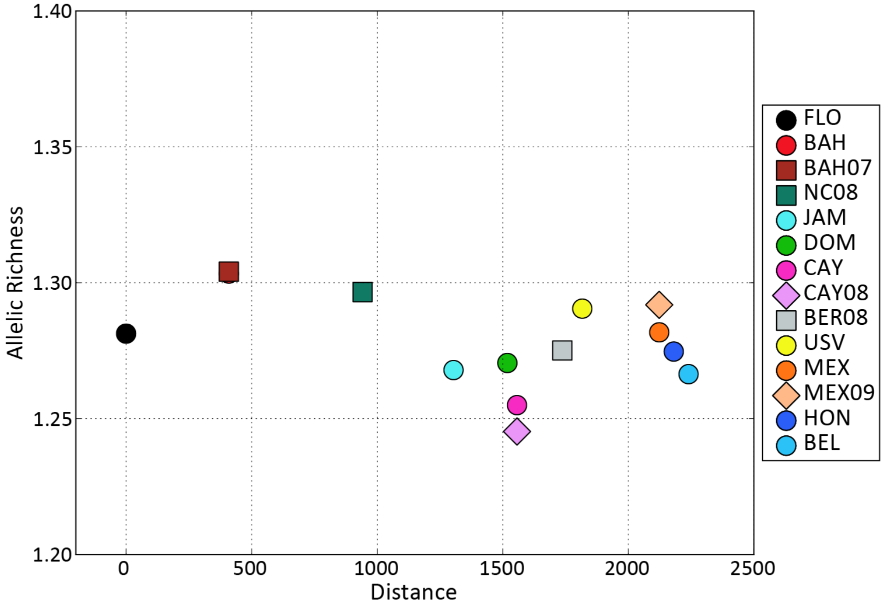
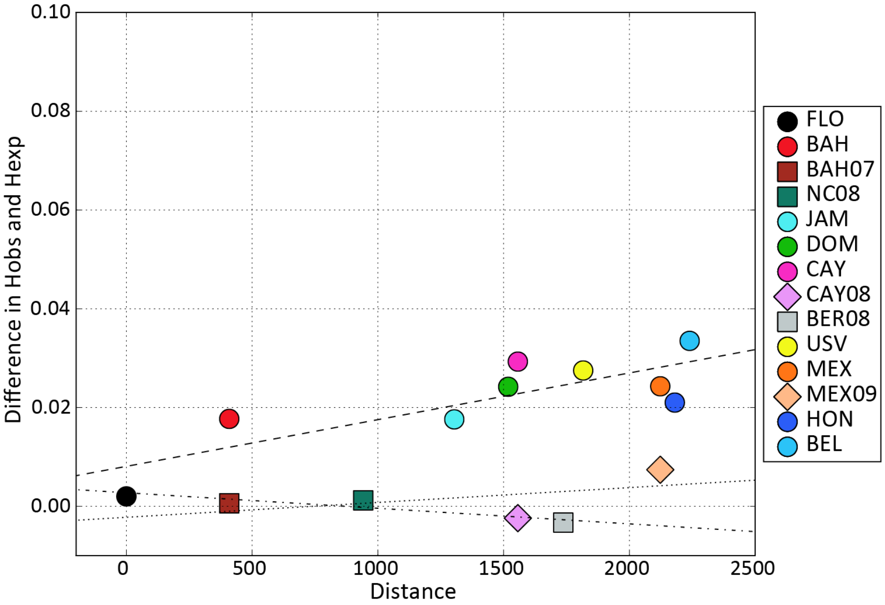


Figure 5. The difference between expected heterozygosity and observed heterozygosity plotted against distance for both TP1 and TP2 samples. Sites that are part of the TP1 dataset are represented by squares (northern) and diamonds (southern); sites that are part of the TP2 dataset are represented by circles. Heavy dashed line is the regression for TP2 ($R^2 = 0.67$, p-value = 0.007), lighter dashed line is the TP1 southern data (not a significant regression), and the dotted line is the regression for the TP1 northern data (not a significant regression).



The PCA generated by the *smartpca* program in *EIGENSOFT* showed one large cluster of individuals with some slight variation of which populations were dominant in different parts of the larger cluster. The first eigenvector explains 13.1% of the variation in the data. The second eigenvector explains 11.8% of the variation in the data. The third eigenvector explains 11.4% of the variation in the data. The PCA resulted in some subtle clustering patterns mostly between the northern TP1 data and other sites (Figure 6). Notably, North Carolina, The Bahamas, and Bermuda from TP2 cluster to an edge of the PCA (*i.e.*, on the bottom right in Figure 6A and to the bottom left in Figure 6B). Structure analysis yielded similar patterns. The ChooseK.py program from *fastSTRUCTURE* reported a value of k (number of genetic lineages) that maximizes likelihood of one, and a value of k that accounts for genetic structuring of three (Figure 7). With a value of $k = 3$, a clear difference in samples from the TP1 northern range and the other populations (dark blue and light blue in Figure 7). Interestingly, individuals from The Bahamas in 2013 have more similarity to samples from 2008 than the other modern locations. The two red samples highlighted in Figure 7 are likely an artefact of the inappropriate number of *k,* meaning that while the program postulated 1-3 as the correct value, here just $k = 3$ is shown; however, $k = 2$ may be closer to the true accurate value for these data. Furthermore, F_{ST} values were only statistically significantly different from zero in some of the pairwise comparisons. Most of the significant values occurred between historic populations and modern populations (Table 3).

Figure 6. Principal Component Analysis of SNP data for all samples. Sites that are part of the TP1 dataset are represented by squares (northern) and diamonds (southern); sites that are part of the TP2 dataset are represented by circles. Axes have been zoomed in to show the main cluster of sites.

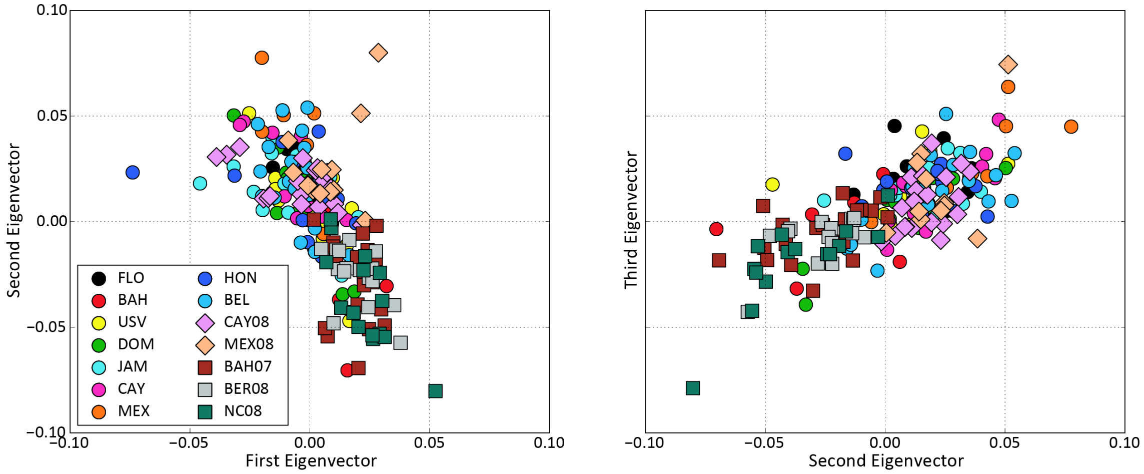


Figure 7. fastSTRUCTURE plot for *k* = 3 in which each genetic lineage is a different color (see key).

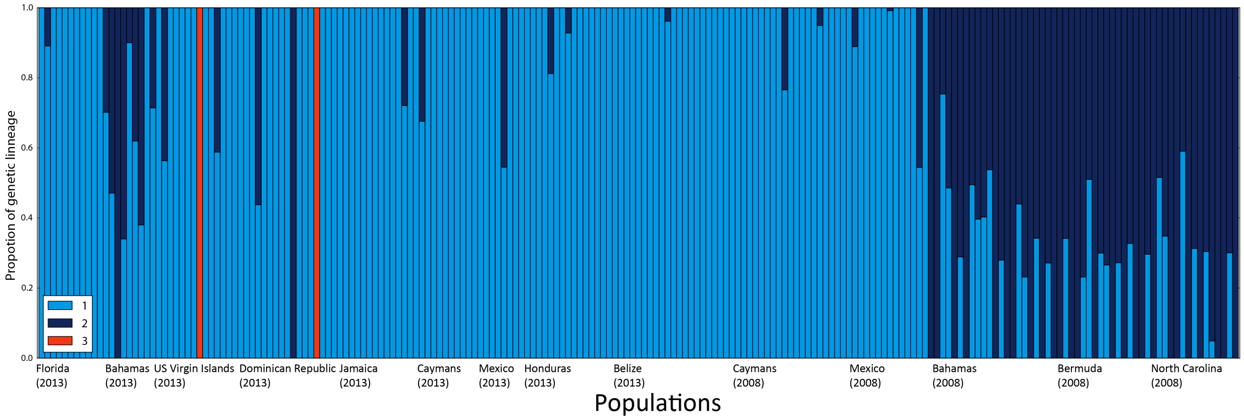


Table 3. F_{ST} values for comparisons between TP1 sites and all sites. Values were corrected in *Stacks* with a Bonferroni correction and all those reported are statistically significantly different from zero, albeit very small.

	CAY08	MEX08	BAH07	BER08	NC08
FLO	0.000995025	0	0.00155837	0.000851471	0
BAH	0	0	0	0	0
USV	0.000647401	0	0.000606709	0.00270324	0
DOM	0.000752807	0	0.000639074	0.0021891	0
JAM	0	0	0.0012228	0.0030433	0
CAY	0	0	0.00211288	0.00121501	0.00170131
MEX	0	0	0	0	0
HON	0	0	0.00165499	0.002827	0
BEL	0	0	0.00100577	0.0028938	0.000593235
CAY08		0	0.000523948	0	0
MEX08			0	0	0
BAH07				0	0
BER08					0

DISCUSSION

This study used 1,054 SNPs sequenced in the invasive lionfish, *Pterois volitans*, across 14 populations at 11 geographic locations from two time points in the invasion to describe how patterns of genetic diversity during the invasion changed through time and to test three hypotheses. Of the three hypotheses tested in this work, only the first was supported by the results. The first hypothesis predicted a weaker signal of expansion in the northern expansion pathway for TP1 than in the southern expansion pathway (in both TP1 and TP2). This was supported, albeit by limited data, for the northern expansion in which all measures of genetic diversity—H_{OBS}, H_{EXP}, and A_{RICH}—were essentially constant while H_{OBS} showed a slight negative trend in the southern expansion (not statistically significant). The second hypothesis generated the prediction that genomic signatures of range expansion would be weaker in TP2 than in TP1, because with time, population processes would lead to dispersal and mixing which would act to diminish the genomic signals of expansion. This was not observed. In fact, the decrease in observed heterozygosity was more pronounced in TP2 data than in TP1 data, although, as discussed in the Results, conclusions regarding the regressions of diversity are tenuous due to limited numbers of population sampled. Additionally, the slopes of the two lines may not be comparable because the data included in the TP1 dataset did not cover the same geographic range as that for the TP2 dataset, potentially affecting the slope of the line and therefore the results. Interestingly, the difference between expected and observed heterozygosity only had a significant relationship related to distance from Florida in the modern dataset. For TP2 populations, with increasing distance from Florida, the difference between expected and observed heterozygosity became greater, indicating a greater tendency towards equilibrium (as discussed in Chapter 4). However, this tendency towards disequilibrium was predicted to be stronger earlier in the invasion (at TP1), before mixing and population growth might offset disequilibrium signals from expansion. The final hypothesis was that genetic diversity overall would be lower earlier in the invasion due to founder effects. This hypothesis was not supported by the data. In fact, genetic diversity was, overall, higher earlier in the invasion. While all efforts were made to make data for TP1 and TP2 equivalent, it is still possible that the differences in ddRAD-seq and a single enzyme RAD-seq resulted in artefacts in heterozygosity data.

In addition to investigating patterns of genetic diversity for the different invasion pathways at both time points, population genetic structure was compared across all data sampled. For TP1 data, there were two groups of populations—one to the north and one to the south, representing the two invasion pathways (the initial expansion from Florida in 2000 and the second to the south in 2004). The Bahamas from TP1 clustered in the PCA with northern populations despite the timing of the first observation there. These data supported our first hypothesis that the two invasion pathways in the historic dataset would be different. The exact mechanism (oceanographic dispersal or other dynamics) is not possible to determine from these data.

Results presented here are counterintuitive to what was expected for two time points in an invasion: heterozygosity appears to be decreasing with time throughout the invaded range. This could indicate that the process of expansion is more complex than predicted. Indeed, additional theory that expands on the original prediction of range expansion theory used to generate hypotheses presented here, have sometimes shown counterintuitive patterns of genetic diversity in range expansions depending on the specific mechanisms driving spread (*i.e.,* pushed waves vs. pulled waves) (Bonnefon *et al.* 2014). Also, while all steps were taken to process the two types of RAD-seq data in the same way, there is still a possibility that ascertainment bias introduced by the second kind of RAD-seq data is responsible for the variation between observations at TP1 and TP2. Therefore, additional data from TP1 generated using single-enzyme RAD-seq will be critical to determine the degree to which this potential bias might affect observed patterns.

Future Directions of This Research

Results presented here generate rather than resolve a series of questions pertaining to the dynamics of range expansion over time in a marine invasive species. Intriguing and counterintuitive results for genetic comparison of lionfish between (approximately) 2008 and 2013 showed lower observed heterozygosity in the second time point (TP2) than in the first (TP1). These results highlight the need for further study of this system because the dynamics driving population genomic summary statistics are may be highly complex (Bonnefon *et al.* 2014). Continued temporal sampling and analysis may be able to distinguish between a persistent trend and random change through time.

To reiterate, the differences in RAD-sequencing techniques used for different portions of this dataset also lead to questions about the ways in which methodology could bias the heterozygosity measurements. While all samples were filtered, trimmed, analyzed the same way, the genomic library and the sequencing reads were not generated the same way. RAD-seq methods have only been developed recently and further research is needed to ascertain if there were biases to sequencing of diversity using one method over another.

Ultimately, the results from this study indicate that the temporal and spatial dynamics of range expansion are more complex than predictions based on a simple model of decreased diversity with distance from the point of invasion. While expectations for this pattern were met in terms of absolute values of observed heterozygosity for some portions of the dataset, the temporal signals of diversity did not meet expectations of expansion signals become weakened with time.

ACKNOWLEDGEMENTS

My gratitude goes to Dr. Wilson Freshwater for providing us with gDNA for many of the historic sampling sites. I also thank the many participants of the Gulf and Caribbean Fisheries Institute for providing lionfish samples from around the Caribbean region, as well as Dr. Bernard Castillo at the University of the Virgin Islands and Kristian Rogers at the Biscayne Bay National Park. I would like to acknowledge Alex Bogdanoff at NOAA, Beaufort NC, for assistance with sample acquisition, and Camrin Braun at WHOI, for assistance with the calculation of oceanic distances between sites. An additional thanks to Jack Cook of the WHOI Graphics department for his assistance in generating the maps of the study area. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374. Sequencing funding was provided in part by the PADI Foundation Grant No. 14904. Additional support for this research was provided by the National Science Foundation (OCE-1131620 to TMS) and the James Education Fund for Ocean Exploration within the Ocean Exploration Institute of the Woods Hole Oceanographic Institution.

LITERATURE CITED

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers (JC Fay, Ed,). *PLoS ONE*, **3**, e3376.

Blackburn TM, Pyšek P, Bacher S *et al.* (2011) A proposed unified framework for biological invasions. *Trends in Ecology & Evolution*, **26**, 333–339.

Bonnefon O, Coville J, Garnier J, Hamel F (2014) The spatio-temporal dynamics of neutral genetic diversity. *Ecological Complexity*, **20**, 282–292.

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

Chen YH, Berlocher SH, Opp SB, Roderick GK (2010) Post-colonization temporal genetic variation of an introduced fly, Rhagoletis completa. *Genetica*, **138**, 1059–1075.

Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences*, **101**, 975–979.

Excoffier L, Foll M, Petit RJ (2009) Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 481–501.

Fennell M, Gallagher T, Vintro LL, Osborne B (2014) Using soil seed banks to assess temporal patterns of genetic variation in invasive plant populations. *Ecology and Evolution*, **4**, 1648–1658.

Fonseca DM, Widdel AK, Hutchinson M, Spichiger SE, Kramer LD (2010) Fine-scale spatial and temporal population genetics of Aedes japonicus, a new US mosquito, reveal multiple introductions. *Molecular Ecology*, **19**, 1559–1572.

Freshwater DW, Hines A, Parham S *et al.* (2009) Mitochondrial control region sequence analyses indicate dispersal from the US East Coast as the source of the invasive Indo-Pacific lionfish Pterois volitans in the Bahamas. *Marine Biology*, **156**, 1213–1221.

Hallatschek O, Nelson DR (2008) Gene surfing in expanding populations. *Theoretical Population Biology*, **73**, 158–170.

Hallatschek O, Hersen P, Ramanathan S, Nelson DR (2007) Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences*, **104**, 19926–19930.

Herrera S, Watanabe H, Shank TM (2015) Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Molecular Ecology*, **24**, 673–689.

Hixon MA, Green SJ, Albins MA, Akins JL, Morris JA Jr (2016) Lionfish: a major marine invasion. *Marine Ecology Progress Series*, **558**, 161–165.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species (L Orlando, Ed,). *PLoS ONE*, **7**, e37135.

Price AL, Patterson NJ, Plenge RM *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

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

Raj A, Stephens M, Pritchard JK (2013) *Variational Inference of Population Structure in Large SNP Datasets*. Cold Spring Harbor Labs Journals.

Ramachandran S, Deshpande O, Roseman CC *et al.* (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences*, **102**, 15942–15947.

Slatkin M, Excoffier L (2012) Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics*, **191**, 171–181.

Sousa V, Peischl S, Excoffier L (2014) Impact of range expansions on current human genomic diversity. *Current Opinion in Genetics & Development*, **29**, 22–30.

Tepolt CK, Palumbi SR (2015) Transcriptome sequencing reveals both neutral and adaptive genome dynamics in a marine invader. *Molecular Ecology*, **24**, 4145–4158.

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

Travis JM, Munkemuller T, Burton OJ *et al.* (2007) Deleterious mutations can surf to high densities on the wave front of an expanding Population. *Molecular Biology and Evolution*, **24**, 2334–2343.

White TA, Perkins SE, Heckel G, Searle JB (2013) Adaptive evolution during an ongoing range expansion: the invasive bank vole (Myodes glareolus) in Ireland. *Molecular Ecology*, **22**, 2971–2985.

CHAPTER 6

Conclusion

I. The Big Picture

We live in the Anthropocene, a geological age defined by human influence (Waters *et al.* 2016). Anthropogenic pressures on the marine environment, from the coasts to the deep sea, are unprecedented (Miles 2009; Barange *et al.* 2010; Ramirez-Llodra *et al.* 2011; Van Dover *et al.* 2012). Global change is affecting species distributions in both terrestrial (Parmesan & Yohe 2003; Thomas *et al.* 2004; Parmesan 2006; Sunday *et al.* 2012) and marine systems (Parmesan & Yohe 2003; Perry 2005; Sabatés *et al.* 2006; Sorte *et al.* 2010; Booth *et al.* 2011; Jones & Southward 2012; Sunday *et al.* 2012; Poloczanska *et al.* 2013). In addition to changing species ranges where they already exist, global climate change, habitat alteration, and increased international trade and travel are likely to amplify the number, rate, and consequences of species invasions (Mainka & Howard 2010). In the world's oceans, increases over the last two centuries in marine transport and shipping as well as expanding reliance on marine resources have led to significant increases in the number of non-native species introduced to marine ecosystems (Ruiz *et al.* 2000).

In the face of such a dramatic reshuffling of life on earth, there is an urgency now to describe repercussions of rapid change. This dissertation contributes to that effort by producing population genetic and genomic information for multiple marine species undergoing dynamic population changes.

II. Dissertation Review

The four data chapters in this dissertation highlight specific characteristics of the processes of invasion, expansion, and connectivity in marine populations. Chapter 2 reported population genetic patterns of two deep-sea invertebrates in New Zealand with varying life history traits. The study evaluated the placement of benthic protection areas in New Zealand, highlighting the importance of both life history and regional oceanographic patterns in determining population structure of benthic species. The work reports patterns that are an averaged summary of historic population genetic connectivity, identifying substantial physical barriers to dispersal that lead to persistent patterns of genetic diversity. In Chapter 3, the context of a rapid marine invasion along the US Atlantic coast provided a system to refine our concept of important processes driving connectivity patterns.

The benefit of existing genetic resources for *Palaemon macrodactylus* allowed for a comparison in Chapter 3 of newly generated data in one invaded region with published global data. Results supported an invasion scenario of multiple introductions of *P. macrodactylus* in the invaded range. Chapter 3 highlights the importance of genetics for describing marine invasions and uncovering invasion pathways that are sometimes impossible to piece together through observations. The use of genome-wide single nucleotide polymorphisms (SNPs) also allowed for a directionality index analysis of the invasion (Peter & Slatkin 2013), which corroborated mitochondrial DNA evidence for multiple introductions. SNP data also facilitated a more in-depth population structure analysis which led us to conclude that human-mediated transport and introduction of marine invasive species may be at least as important as other factors in shaping population genetic structure and determining invasion dynamics in some marine species. This result represented a dramatic addition to the existing paradigm of genetic connectivity research: one in which anthropogenic activities are explicitly considered as part of marine systems. In this way, Chapter 3 logically elaborated on the conclusions of Chapter 2.

Chapter 4 contains the first population genomic data generated using RAD-seq for the invasive lionfish, *Pterois volitans*. Using 12,759 loci, geographic patterns were observed correlating diversity with distance from the point of invasion—specifically a pattern of decreased observed heterozygosity with increased distance from Florida—despite a lack of spatial metapopluation genetic structure. Patterns in F_{IS} indicated that in addition to the spatial processes of range expansion, population demography (*e.g.,* population density) could play a role in shaping diversity in different portions of the range. Chapter 4 emphasizes the utility of population genomic data for generating hypotheses about invasion pathways. Finally, novel methods were introduced that identify specific loci in the genome with unique patterns of diversity in the range—*e.g.,* those that changed from major to minor allele or those with large differences in frequency throughout the range—and we compared those loci to genomic regions identified with outlier analyses. These types of analyses could become more common as the utility of RAD-seq data is more fully realized.

With a spatial perspective developed in Chapter 4, a temporal approach was then taken in Chapter 5 to extend and deepen our understanding of how the lionfish invasion is unfolding in the Caribbean Sea and the US East Coast. Data generated for Chapter 5 did not support the hypothesis that with time, genetic signatures of range expansion would weaken due to the interaction of forces that promote connectivity and dispersal with the spatial forces of invasion. Further research on the specific temporal dynamics of this system—specifically into what could drive decreases in overall heterozygosity through time—will help clarify what the temporal dynamics of the lionfish have been and what the continuing invasion will mean for lionfish population genomics.

III. Future Directions and Finding the Right Tool for the Job

Undeniably, reduced representation library methods like RAD-seq have enabled the non-model species research community to explore genomic questions in species without a sequenced genome. Still, RAD-seq has limitations. Most notably, not being able to accurately designate all the reads to a specific genomic region—or even to specify the type of region (*e.g.*, gene-coding, regulatory, neutral)—represents a major limitation of the method. Significantly more genetic information can be gathered when RAD-seq can be used to generate genetic maps through crossing of individuals in a laboratory setting (historically the initial purpose of the method). As a population genomic tool, RAD-seq is the state of the art for non-model species but significantly more insight will come from a deeper knowledge of the sites being sequenced. Combining RAD-sequencing with other methods, like transcriptomics, may generate a more nuanced view of the markers being used in population genomic analyses. Future work is underway to utilize the paired-end sequences generated for Chapter 4 to assemble longer contiguous reads to better identify the regions of the genome containing analyzed SNPs. Another potential problem with RAD-sequencing, which was encountered in Chapter 5, is that it is not always possible to accurately identify and quantify the ascertainment bias in the sequencing. It is known that polymorphism in restriction sites can lead to ascertainment biases that underestimates diversity (Arnold *et al.* 2013). However, it is not clear how variable the impact of this ascertainment bias is nor how ascertainment bias will affect the combination of different types of RAD-sequencing (Chapter 5).

One major question unanswered by my work concerns the relative importance of genetic drift and selection during range expansion. In the analyses presented here, disentangling drift and selection remains challenging because of the similarity in the population genomic signals of genetic drift and selection in an expansion. One example of successful disentanglement used approximate Bayesian computation to test invasion scenarios (Antoniazza *et al.* 2014). Another possible method for disentangling the drift-selection balance during expansion could be based on coalescent comparisons of patterns from expansion and selection (Nullmeier & Hallatschek 2013), but this would require RAD-seq data amenable to coalescent analysis, which is still in development.

This dissertation presents empirical research motivated by theory. Empirical and theoretical approaches to scientific questions participate in an iterative, reflexive process in which the two parallel tracks of inquiry periodically inform one another. Some of the questions generated by this research will be best further explored with theoretical or computational tools. For example, understanding the circumstances that break the expectations of range expansion can possibly be observed in empirical work. I described one such situation in the northern expansion of lionfish along the US East Coast in which dispersal and oceanographic processes possibly erase any patterns generated by allele surfing. However, actually determining the point at which range expansion expectations fail is a task perhaps best approached with simulations or laboratory experiments. It is important, one might say, to use the right tool for the job. Thus, the place for this dissertation research is within an iterative process of theory and empiricism in which predictions are made, patterns are observed, paradigms are amended, and new hypotheses are developed.

LITERATURE CITED

Antoniazza S, Kanitz R, Neuenschwander S *et al.* (2014) Natural selection in a postglacial range expansion: the case of the colour cline in the European barn owl. *Molecular Ecology*, **23**, 5508–5523.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

Barange M, Field JG, W S (2010) Marine Ecosystems and Global Change.

Booth DJ, Bond N, Macreadie P (2011) Detecting range shifts among Australian fishes in response to climate change. *Marine and Freshwater Research*, **62**, 1027–1042.

Jones SJ, Southward AJ (2012) Climate change and historical biogeography of the barnacle Semibalanus balanoides. *Global ecology and Biogeography*.

Mainka SA, Howard GW (2010) Climate change and invasive species: double jeopardy. *Integrative Zoology*, **5**, 102–111.

Miles EL (2009) On the increasing vulnerability of the world ocean to multiple stresses. *Annual Review of Environment and Resources*, **34**, 17–41.

Nullmeier J, Hallatschek O (2013) The coalescent in boundary-limited range expansions. *Evolution*, **67**, no–no.

Parmesan C (2006) Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 637–669.

Parmesan C, Yohe G (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, **421**, 37–42.

Perry AL (2005) Climate change and distribution shifts in marine fishes. *Science*, **308**, 1912–1915.

Peter BM, Slatkin M (2013) Detecting range expansions from genetic data. *Evolution*, **67**, 3274–3289.

Poloczanska ES, Brown CJ, Sydeman WJ *et al.* (2013) Global imprint of climate change on marine life. *Nature Climate Change*, **3**, 1–7.

Ramirez-Llodra E, Tyler PA, Baker MC *et al.* (2011) Man and the last great wilderness: human impact on the deep sea (P Roopnarine, Ed,). *PLoS ONE*, **6**, e22588–25.

Ruiz GM, Fofonoff PW, Carlton JT, Wonham MJ (2000) Invasion of coastal marine communities in North America: apparent patterns, processes, and biases. *Annual review of ecological systems*.

Sabatés A, Martín P, Llorer J, Raya V (2006) Sea warming and fish distribution: the case of the small pelagic fish, Sardinella aurita, in the western Mediterranean. *Global Change Biology*, **12**, 2209–2219.

Sorte CJ, Williams SL, Carlton JT (2010) Marine range shifts and species introductions: comparative spread rates and community impacts. *Global Ecology and Biogeography*, **19**, 303–316.

Sunday JM, Bates AE, Dulvy NK (2012) Thermal tolerance and the global redistribution of animals. **2**, 686–690.

Thomas CD, Cameron A, Green RE *et al.* (2004) Extinction risk from climate change. *Nature*, **427**, 145–148.

Van Dover CL, Smith CR, Ardron J *et al.* (2012) Designating networks of chemosynthetic ecosystem reserves in the deep sea. *Marine Policy*, **36**, 378–381.

Waters CN, Zalasiewicz J, Summerhayes C *et al.* (2016) The Anthropocene is functionally and stratigraphically distinct from the Holocene. *Science*, **351**, aad2622–aad2622.