2019

Simulations and Modelling for Biological Invasions

Ryan D. Scott
University of Windsor

Simulations and Modelling for Biological Invasions

By

Ryan D. Scott

A Dissertation

Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
at the University of Windsor

Windsor, Ontario, Canada

2019

Simulations and Modelling for Biological Invasions

by

Ryan D. Scott

APPROVED BY:

_____

J. Travis, External Examiner
University of Aberdeen

_____

D. Heath
Great Lakes Institute for Environmental Research

_____

A. Ngom
School of Computer Science

_____

B. Boufama
School of Computer Science

_____

R. Gras, Advisor
School of Computer Science

_____

H. MacIsaac, Co-Advisor
Great Lakes Institute for Environmental Research

May 13, 2019

DECLARATION OF CO-AUTHORSHIP / PREVIOUS PUBLICATION

I. Co-Authorship

I hereby declare that this dissertation incorporates material that is result of joint research, as follows:

Chapters 1, 2, 4, and 6 were authored under the supervision of Dr. Robin Gras and Dr. Hugh MacIsaac. Chapter 3 was co-authored with Dr. Brian MacPherson under the supervision of Dr. Robin Gras. Chapter 5 was co-authored with Dr. Aibin Zhan, Dr. Emily A. Brown, Dr. Frédéric J. J. Chain, and Dr. Melania E. Cristescu, under the supervision of Dr. Robin Gras and Dr. Hugh MacIsaac. In all cases, the key ideas, primary contributions, experimental designs, data analysis, interpretation, and writing were performed by the author. The contribution of co-authors was primarily through the provision of background biological knowledge, feedback on refinement of ideas and analysis, and editing of the manuscripts.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my dissertation, and have obtained written permission from each of the co-author(s) to include the above material(s) in my dissertation.

I certify that, with the above qualification, this dissertation, and the research to which it refers, is the product of my own work.

II. Previous Publication

This dissertation includes three original papers that have been previously published/submitted for publication in peer reviewed journals, as follows:

Dissertation Chapter	Publication title/full citation	Publication status
Chapter 3	Scott R., MacPherson B., Gras R. (2018) EcoSim, an enhanced artificial ecosystem: addressing deeper behavioral, ecological, and evolutionary questions, Cognitive Architectures, Maria Isabel Aldinhas Ferreira, João Silva Sequeira and Rodrigo Ventura Editors, Springer, 223-278.	Published
Chapter 4	Genetic diversity impacts establishment success of digital invaders in heterogeneous environments	Submitted

Chapter 5	Scott R., Zhan A., Brown E.A., Chain F.J.J., Cristescu M.E., Gras R., MacIsaac H.J., Optimization and performance testing of a sequence processing pipeline applied to detection of nonindigenous species, Evolutionary Applications, 11, 891-905, 2018.	Published

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my dissertation. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor.

III. General

I declare that, to the best of my knowledge, my dissertation does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my dissertation, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my dissertation.

I declare that this is a true copy of my dissertation, including any final revisions, as approved by my dissertation committee and the Graduate Studies office, and that this dissertation has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

Biological invasions are characterized by the movement of organisms from their native geographic region to new, distinct regions in which they may have significant impacts. Biological invasions pose one of the most serious threats to global biodiversity, and hence significant resources are invested in predicting, preventing, and managing them. Biological systems and processes are typically large, complex, and inherently difficult to study naturally because of their immense scale and complexity. Hence, computational modelling and simulation approaches can be taken to study them. In this dissertation, I applied computer simulations to address two important problems in invasion biology.

First, in invasion biology, the impact of genetic diversity of introduced populations on their establishment success is unknown. We took an individual-based modelling approach to explore this, leveraging an ecosystem simulation called EcoSim to simulate biological invasions. We conducted reciprocal transplants of prey individuals across two simulated environments, over a gradient of genetic diversity. Our simulation results demonstrated that a harsh environment with low and spatially-varying resource abundance mediated a relationship between genetic diversity and short-term establishment success of introduced populations rather than the degree of difference between native and introduced ranges. We also found that reducing Allee effects by maintaining compactness, a measure of spatial density, was key to the establishment success of prey individuals in EcoSim, which were sexually reproducing. Further, we found evidence of a more complex relationship between genetic diversity and long-term establishment success, assuming multiple introductions were occurring. Low-diversity populations seemed to benefit more strongly from multiple introductions than high-diversity populations. Our results also corroborated the evolutionary imbalance hypothesis: the environment that yielded greater diversity produced better invaders and itself was less invasible. Finally, our study corroborated a mechanical explanation for the evolutionary imbalance hypothesis – the populations evolved in a more intense competitive environment produced better invaders.

Secondly, an important advancement in invasion biology is the use of genetic barcoding or metabarcoding, in conjunction with next-generation sequencing, as a potential means of early detection of aquatic introduced species. Barcoding and metabarcoding invariably requires some amount of computational DNA sequence processing. Unfortunately, optimal processing parameters are not known in advance and the consequences of suboptimal parameter selection are

poorly understood. We aimed to determine the optimal parameterization of a common sequence processing pipeline for both early detection of aquatic nonindigenous species and conducting species richness assessments. We then aimed to determine the performance of optimized pipelines in a simulated inoculation of sequences into community samples. We found that early detection requires relatively lenient processing parameters. Further, optimality depended on the research goal – what was optimal for early detection was suboptimal for estimating species richness and vice-versa. Finally, with optimal parameter selection, fewer than 11 target sequences were required in order to detect 90% of nonindigenous species.

DEDICATION

To all my friends and family who support me and believe in me. To my parents, who taught me and my siblings to always prioritize education, and who continue to demonstrate to us the importance of hard work. To my extended families, the Dixons and the Antoniws – you all treat and support me as one of your own, and I am incredibly grateful. Your love and support through this entire process means the world to me, and I could not have done this without you. To Shawna, who has blazed a trail of achievement and taught me what it takes to succeed. To Stephanie, who has shown me how to be strong and fearless. To Jamie, who has shown me that barriers are to be broken and that anything is possible. To Chris, whose honesty, determination, and character inspires me to always be my best. To Tyler, whose positivity and support has kept me optimistic throughout my life. To Tanya, the love of my life, who has been beside me through all of this, who encourages me to keep pushing, whose patience and support I could not have done this without, and who reminds me that 42 is just 32 and 10.

ACKNOWLEDGEMENTS

I would like to express my thanks to my committee members, whose input helped me immensely. With your input, I was able to look at my research with a different perspective and a more critical eye. I would like to thank Dr. Hugh MacIsaac and Dr. Robin Gras – your honesty, feedback, guidance, and constructive criticism has helped me learn of the standards required to conduct and disseminate research. Throughout this experience I have also learned a lot about myself. It has been an honour working with you both, and it has shaped me into the person I am today.

TABLE OF CONTENTS

CHAPTER 1
Introduction

1.1 Motivation

Biological invasions occur when species are transported outside of their native range and establish in a novel, distinct range (Colautti and MacIsaac 2004). They can inflict ecological, economical, and societal harm (Colautti *et al.* 2006; Vilà *et al.* 2011; Simberloff *et al.* 2013). The field of invasion biology is enormous, important, and interdisciplinary; with ever-increasing globalization, its relevance now and in the future is and will remain unquestionable (Simberloff *et al.* 2013). However, despite the research that has been conducted in this domain, many questions remain unanswered and many theories remain to be tested. Many of these theories are computational in nature or lend themselves to simulations or modelling approaches (Hargreaves and Eckert 2014; Bock *et al.* 2015; Dlugosch *et al.* 2015; Xiong, Li, and Zhan 2016). This dissertation was motivated by the importance of biological invasions, and by open questions remaining in invasion biology that can be answered using computational modelling or simulations approaches.

1.2 Objectives

It has been theorized that introduced populations should be subject to a genetic bottleneck; that is, they should suffer from a reduction in genetic and thus phenotypic diversity relative to the population in their native range, and this is theorized to negatively impact their adaptability and establishment success (Sakai *et al.* 2001; Rius and Darling 2014; Bock *et al.* 2015; Dlugosch *et al.* 2015). Studies of the effects of genetic diversity of introduced populations on their establishment success have been largely inconclusive, so this remains an interesting area of research (Roman and Darling 2007; Bock *et al.* 2015; Estoup *et al.* 2016). There are related questions as well. What is the relationship between genetic diversity of introduced populations and their establishment success (Sakai *et al.* 2001; Bock *et al.* 2015)? Does this relationship vary based on the similarity or harshness of native and introduced ranges (Hufbauer *et al.* 2012; Hufbauer *et al.* 2013; Bock *et al.* 2015; Estoup *et al.* 2016)? Outside of propagule pressure (the total number of individuals introduced), what other factors affect establishment success (Sakai *et al.* 2001; Simberloff 2009; Colautti, Grigorovich, and MacIsaac 2006)? Is there a relationship between fitness of a genotype in the native range and fitness of a genotype in the introduced range (Sakai *et al.* 2001; Bock *et al.* 2015)? Do genetically diverse introduced populations succeed under the same conditions as those that are not (Sakai *et al.* 2001; Roman and Darling 2007; Bock *et al.* 2015)?

Answering these questions is difficult with studies of real biological invasions due to several important limitations. First, to quantify establishment success it is also necessary to quantify establishment failures, which are extremely difficult to observe and study (Roman and Darling 2007). Moreover, few studies, if any, have been able test the diversity of introduced populations independent of other factors such as propagule size (the number of individuals in a single introduction; Colautti, Grigorovich, and MacIsaac 2006). Finally, biological invasions can occur over enormous temporal and spatial scales, often involving hundreds to millions of propagules, rendering the testing of such theories prohibitive due to resource requirements alone. Because of these important limitations, testing such theories lends itself to computational modelling and simulation approaches. With a simulation approach, these limitations can be alleviated; establishment failures can be completely accounted for, diversity of introduced populations can be tested independent of other factors, and minimal resources are required to study thousands or millions of digital organisms over extensive time periods.

Answering theoretical evolutionary and ecological questions was the main purpose of EcoSim, an individual-based, predator-prey ecosystem simulation which came about in 2009 (Gras *et al.* 2009). The simulation has undergone extensive validation and has been instrumental in shedding light on numerous eco-evolutionary theories (Golestani, Gras, and Cristescu 2012; Khater, Murariu, and Gras 2015; Karim Pour *et al.* 2017; MacPherson *et al.* 2017; Scott, MacPherson, and Gras 2018). However, as individuals in its original inception were sexless, there were limitations to the types of theoretical questions it could answer. Also, there were some important potential submodels that were heavily simplified or entirely left out of the simulation due to computational constraints at the time of its inception. As computational power has increased, with the current ubiquity of high-performance computational networks, and to increase the range of questions that EcoSim could answer, advanced versions of EcoSim were needed. Further, some new EcoSim variants would be required if EcoSim were to be used to answer theoretical questions in invasion biology. Thus, a major objective of this dissertation was to develop EcoSim and some new variants of it, and to use it to answer some of the questions posed earlier in this chapter.

From a practical standpoint, translocation of species, accidental or purposeful, is almost inevitable in our current world (Simberloff *et al.* 2013). When introductions do occur, early detection of the introduced individuals is imperative; a rapid response can be enacted, capitalizing on the presumed low abundance of the introduced population (Genovesi 2005; Simberloff *et al.* 2013). Traditional detection methods have recently been displaced in favour of molecular genetic detection methods, particularly metabarcoding of environmental DNA or bulk samples which is showing promise in rapidly detecting a wide range of species with incredibly high sensitivity (Smart *et al.*

2015). Metabarcoding involves the usage of small barcode regions on partial genomes to identify species in samples (Ratnasingham and Hebert 2013). One issue with a metabarcoding approach is that there are many potential sources of error throughout the process, and so computational processing of the sequence data is required (Xiong, Li, and Zhan 2016). The computational processing step itself can be a source of both false-positive (species identified as present when it is not) and false-negative (species not identified as present when it is) error, and optimal parameterization of the computational sequence processing pipeline is completely unknown throughout the process (Flynn *et al.* 2015; Xiong, Li, and Zhan 2016). Thus, we aimed to elucidate how users should select parameters for computational sequence processing, particularly for metabarcoding of zooplankton bulk samples, and the consequences of poor parameter selection. We also aimed to determine if research goals influenced parameter selection – is optimal parameterization when the researcher intends to estimate species richness the same as when metabarcoding is being used for early detection of invasive species? We aimed to discover what level of performance can be expected from a metabarcoding approach, assuming optimal parameterization. Finally, a key question in early detection of invasive species is if metabarcoding is sufficient, or if a targeted approach is necessary for minimization of false positive and false negative errors. Answering this question was another objective of this dissertation.

1.3 Contributions

- Implemented significant improvements to EcoSim, broadening its potential applications and adding numerous features
- Developed a new variant of EcoSim called EcoSim Niches, with a comparatively harsher environment in which resource abundance is lower and more spatially-variable than that of standard EcoSim, which we showed develops greater diversity in prey and predator genotypes, species, and adaptive strategies
- Developed a new variant of EcoSim called EcoSim Invasions, with standard and Niches subvariants, allowing users to study biological invasions from an eco-evolutionary perspective using EcoSim
- Showed that in EcoSim, in which individuals are sexual, combatting Allee effects are of prime importance for introduced populations
- Found that the harshness of the recipient environment, not the degree of similarity between native and introduced environments, mediates a relationship between genetic diversity and short-term establishment success of introduced populations
- Corroborated the evolutionary imbalance hypothesis, which states that environments exhibiting high diversity should produce better invaders and simultaneously be less invasible

- Corroborated a potential explanation for the evolutionary imbalance hypothesis, which involved the evolution of competitive advantages by subjection to a competitively intense environment
- Optimized a sequence processing pipeline for bulk zooplankton metabarcoding, for estimating species richness or for early detection of aquatic invasive species
- Demonstrated the performance of optimal sequence processing of metabarcoded bulk zooplankton samples in estimating species richness and early detection of aquatic invasive species
- Established that metabarcoding alone was insufficient for effective early detection of aquatic invasive species – it should be used in conjunction with a targeted approach to confirm hits discovered with metabarcoding

1.4 Outline

Chapter 2 introduces artificial life, an interdisciplinary research field that involves the creation of artificial systems that simulate life. It then introduces biological invasions, discussing their importance and efforts to predict and manage them. It introduces individual-based models and reviews usage of pragmatic and paradigmatic individual-based models for studying biological invasions. It then discusses the importance of early detection and methods of detection of introduced species.

Chapter 3 provides a detailed description of the current standard version of EcoSim, an individual-based predator-prey ecosystem simulation. EcoSim was developed as an experimental platform for testing eco-evolutionary theories. In Chapter 4, EcoSim was used to explore the effects of genetic diversity on establishment success of introduced populations in a simulated biological invasion with multiple introductions. In Chapter 5, a simulation experiment using empirical sequence data was conducted to optimize a sequence processing pipeline and subsequently test its performance for early detection of aquatic invasive species and estimation of species richness. Chapter 6 provides a conclusion of this dissertation and discusses potential future work in the realm of simulations of biological invasions.

CHAPTER 2

Background

2.1 Simulations and Artificial Life

Simulations are approximations of real or imagined systems (Banks *et al.* 2001). Development of a simulation involves the development of the models of which it is comprised, and these models may be atomic or composed of submodels themselves. The degree to which a simulated system must reflect a real system depends on the intended purpose of the simulation. For instance, a flight-simulating video game need not be very realistic; its intended purpose is entertainment so there is no harm in simplifying the system to the point where it is far from reality. On the other hand, a flight simulator that is designed for military training necessitates extreme detail in terms of the definition of its models; a user that learned to fly in this simulator is expected to gain skills transferrable to the operation of extremely expensive and dangerous machines. Thus, typical development of a simulation involves describing the entities that must exist within the system, defining the key properties of the entities that must be present based on its intended use, and subsequently testing and validating the simulation for its intended use. Simulations take on many forms; for instance, a simulation may consist of a few mathematical equations (e.g. equation-based modelling) or another may be comprised of hundreds of thousands of lines of C++ code.

Artificial life, often abbreviated ALife, is an interdisciplinary field of research that aims to study and understand natural phenomena by creating artificial systems that simulate natural life. The term "Artificial Life" was coined in the mid-1980s, when Chris Langton introduced it at a workshop "on the synthesis and simulation of living systems" (Aguilar *et al.* 2014). The field has since exploded, yielding two flagship conferences (*Artificial Life* and the *European Conference on Artificial Life* – among others of course) and a journal (*Artificial Life* – again, among others), all of which are coordinated through the International Society for Artificial Life (*ISAL*) which Langton started (Aguilar *et al.* 2014). Langton's idea of ALife expanded upon the concept of the cellular automaton brought forth by von Neumann and Burks (1950). Cellular automata involves the creation of space- and time-discrete models, where space is arranged as a grid of cells which holds a finite number of states. The states, most simply, can be binary-state (i.e. "on" or "off"), and the states of cells change according to a set of predefined rules often based on some predefined neighbourhood about a given cell. Time, also being discrete, can be described in terms of "generations", starting at generation zero. The original inception of the cellular automaton aimed to develop self-replicating machines, which could be studied to help us understand how life adapts to its environment.

Conway's Game of Life, based on cellular automata, is a zero-player game that allows one to set initial states and update rules (Gardner 1970). After setting these parameters, the system proceeds to compute the changing states of each cell according to the predefined rules. The user, after the simulation has begun, is unable to interact directly with it (hence "zero-player"). The user can only observe the system as it proceeds and evolves. What is interesting in the study of such cellular automata is not the initial configurations or the rules (as these are prescribed by the user), but instead the emergent properties of the system as it proceeds. In cellular automata, the rules are defined for individual cells; the emergent properties of the system arise as different individuals of the system interact using these basic rules. Conway's Game of Life can produce some beautiful and intricate patterns that play out through simulated time, and these patterns can remain stable for as little as a few generations, to indefinitely. Indeed, complex and stable systems can arise from extremely simple rules that are mathematically predefined for individual parts of the system (Holland 1998).

John Holland, presumably inspired by the study of such complex, emergent, and adaptive systems, brought forth the use of ALife-inspired algorithms in the study of optimization (Holland 1992). For example, his genetic algorithm takes an evolutionary approach to optimization; a population of solutions is initialized, their fitness is measured, and then the solutions recombine with each other to form subsequent generations based on their level of fitness. In the case of many ALife-inspired optimization tasks, and also in many ALife-inspired simulations, fitness is predefined (Gras *et al.* 2015). For example, consider the optimization of the travelling salesman problem, defined as follows: a salesman must visit n places exactly once, starting and finishing at any one of the n places. The individuals in this system are potential solutions (paths that the salesman could follow), and fitness is defined per solution and might be related to the total distance travelled in the journey, the time it might take, or the total cost of the journey. With a genetic algorithm, a population, initially of randomly-generated individuals, would iteratively replicate according to a set of predefined rules much like what was described in our discussion of cellular automata. As the population of the system evolves, better and better solutions to the optimization task arise in the population. Though optimality is never guaranteed with the use of such an optimization algorithm, such simple, evolving, ALife-inspired algorithms have been shown to produce incredible results in extremely complex tasks; for example, an ALife-designed antenna developed by NASA ultimately made its way into space (Hornby 2006). Now there exist numerous examples of ALife-inspired optimization algorithms, including ant colony and particle swarm optimization algorithms.

There are, indeed, fundamental differences between the synthetic ALife simulations stemming from the earliest works by the likes of von Neumann and Langton

and the ALife optimizations stemming from the works of Holland. ALife simulations aim to synthesize artificial life by describing rules and interactions in a bottom-up manner. On the other hand, the optimization approaches allow us to construct top-down systems as a means of conducting analysis on a complex system. However, when fitness is predefined in an ALife simulation, the evolution of the system is no longer open-ended and the behavior of the system is theoretically no different from an ALife inspired optimization in which the user defined the characteristics of the fitness landscape. For an ALife simulation to feature true open-ended evolution, even the fitness landscape itself must evolve and emerge as a property of the system, due to evolving interactions between its atomic parts. For instance EcoSim, a predator-prey individual-based ecosystem model (see Section 2.3) is one of the few such ALife systems that feature true open-ended evolution. Synthetic ALife systems were initially extremely basic, almost coming off as a mathematics-based form of art or entertainment. Nowadays, ALife systems can be extremely complex and used to describe, analyze, and make predictions of a wide range of physical, chemical, and biological systems. A major part of this dissertation involves the use of individual-based modelling, a major field of ALife, to provide theoretical insights regarding biological invasions.

2.2 Biological Invasions

Biological invasions involve the transportation of individuals from their native geographic range to a new and distinct range, in which they can establish and have negative impacts (Colautti and MacIsaac 2004). The invasion process can be seen as individuals of some species passing through a series of filters or barriers (Kolar and Lodge 2001; Sakai *et al.* 2001; Colautti and MacIsaac 2004); if they pass through all of the filters and cause impact in their new range, they are considered invasive. Potential invaders start as propagules in their native range. The first step for a biological invasion is for the propagules to be taken up into a transport vector. The propagules must then survive their journey to the new location, and subsequently be released into it. The individuals must be able to survive and reproduce in their new environment in order to establish. Further, they must also be able to successfully interact with the local community in the new range, and this involves exploiting an open ecological niche or exploiting some niche better than the natives do, via competition or consumption of resources such as native flora or fauna. Finally, if the species has strong dispersal capabilities, the species can typically become widespread and reach high abundances (for example, one of the most successful invaders of all time, the zebra mussel). On the other hand, the species may not disperse well but can still cause impacts local to its destination, even at relatively low abundances.

As noted, invasions require individuals to be taken up into a transport vector and for the individuals to survive their journey to a new location. These transport vectors are typically provided directly by humans; invasions are indeed initiated and facilitated by human action. Individuals can be introduced, for instance, via ballast water (Ricciardi and MacIsaac 2000), animal and plant trade (Duggan *et al.* 2006; Pyšek and Richardson 2010), cargo, luggage, other animals (Leighton *et al.* 2012), vehicles (De Ventura *et al.* 2017), and even biocontrol attempts (Shanmuganathan *et al.* 2010). Propagule pressure is a term used to describe the total number of individuals introduced to a novel location (Colautti, Grigorovich, and MacIsaac 2006). Propagule pressure consists of two main factors – propagule size and propagule number (Simberloff 2009). Propagule size is the number of individuals introduced in a single introduction event, while propagule number is the number of introduction events (Simberloff 2009). Some authors include other factors in the definition of propagule pressure, such as the frequency of introduction events and the condition of the individuals when released into the novel environment ("propagule quality"; Simberloff 2009). Of course, propagule pressure is widely recognized as a key determinant in the establishment success of introduced populations (Colautti, Grigorovich, and MacIsaac 2006). Though the exact amount of propagule pressure for a given introduction is typically unknown, many have attempted to estimate it via historical data or even genetic data from an existing introduced population (Simberloff 2009). Though propagule pressure is surely a key determinant of establishment success, there are some extreme counterexamples. For example, the North American population of *Lasioglossum leucozonium*, a bee species originating in Europe, is theorized to have originated from a single female (Zayed, Constantin, and Packer 2007). As "invasion biology is largely a probabilistic science" (Simberloff 2009), such examples form the exceptions, not the rule.

Below, we elaborate on why biological invasions are important to study. We then discuss the prediction and management of biological invasions. We then elaborate on individual-based models, which are one of the techniques used in predictions of invasions and review a range of studies using individual-based models with respect to invasions from both practical and theoretical perspectives. Finally, we discuss the importance of early detection of invasive species and techniques used in early detection.

2.2.1 Importance of Biological Invasions

Biological invasions can inflict difficult-to-predict and difficult-to-assess negative ecological, economical, and even societal impacts (Colautti *et al.* 2006; Vilà *et al.* 2011; Simberloff *et al.* 2013). Most countries have hundreds or thousands of non-native species, many of which are still being discovered as they were previously thought to be native (Lodge 1993; Cristescu 2015). Biological invasions have long been recognized to

pose one of the most serious threats to worldwide biodiversity (McKinney and Lockwood 1999; Dirzo and Raven 2003; Shochat *et al.* 2010; Simberloff *et al.* 2013); yet despite a wide range of actions taken to reduce human-mediated introductions, invasion rates are not appearing to slow (Seebens *et al.* 2017). Due to human activities, including but not limited to biological invasions, a global "biotic homogenization" is occurring whereby many "loser" species are being replaced by very few "winner" species, and known invaders represent a relatively high proportion of the latter group (McKinney and Lockwood 1999; Shochat *et al.* 2010). Conservative estimates from Dirzo and Raven (2003) are that 30% of threatened birds, 15% of threatened plants, and 10% of threatened mammals are directly affected by introduced populations. Non-exhaustively, invasive species can impose ecological impacts by competing with, eating, or even hybridizing with native species, resulting in the loss of native biodiversity. From the conservation standpoint alone, biological invasions require attention, not only from scientists but also from policymakers and the public.

Unfortunately, humans generally do not like to acknowledge problems, let alone solve them, until they are costing us money. Invasions do plenty of economic damage. Economic impacts can stem from, for instance, damage to manmade structures, resources spent preventing, managing, or eradicating invasions, and damage to aquaculture and agriculture industries. For instance, in a study by Schultz *et al.* (2011), the cost of biofouling of a single class of US Navy ships was estimated to be >$1 billion over approximately 15 years, largely due to heavily reduced fuel economy but also due to maintenance such as the application of antifouling paints and cleaning. For the entire naval fleet, the costs were estimated to be a minimum of $400 million annually, and the US naval fleet represents <0.5% of ships in the entire world. Nonindigenous species also create serious costs for fisheries, livestock, agriculture, and natural resources industries. For example, Colautti *et al.* (2006) characterized the cost of nonindigenous species in Canada in these industries at that time. On the limited data available, they were able to conservatively attribute approximately $187 million annually in costs due to nonindigenous species and estimate that this value only represents about 1% of the true cost. The losses stemmed from, among other factors, reduced efficiency in hydroelectric power plants, fouling of water intakes, reduced production of timber and non-timber forests, damage to food crops, damage to livestock or production therefrom, and research and management of particularly problematic nonindigenous species. A similar study on the impacts of nonindigenous insects in continental USA estimated that tree boring insects, which can often cause mortality of their hosts and seem to be the costliest guild of nonindigenous insects, cost local governments approximately $1.7 billion yearly with $830 million lost by residents in property value and about $130 million lost in timber (Aukema *et al.* 2011). In terms of the cost of management, as of 2002, Australia was

already spending approximately $100 million AU yearly and New Zealand was spending approximately $121 million NZ to control and prevent invasions (Colautti *et al.* 2006). New Zealand and Australia are considered world leaders in control and prevention of nonindigenous species. On a related note, biological invasions can have societal impacts on humans as well; for example, there exists plenty of legislation (informed by invasion biologists and other scientists) concerned with the transportation of plants and animals outside of their native range. Since 2006, intercontinental ships entering the Great Lakes have been subject to Canada's saltwater flushing policies. In Alberta, there exists mandatory boat checking stations for boats being transported between bodies of water, as an example of infrastructure change caused by biological invasions. Clearly, biological invasions demand our attention and intervention, from both conservation and economic standpoints.

2.2.2 Prediction and Management

Bock *et al.* (2015) describe the relationship of ecologists and evolutionary biologists to invasive species as "love-hate". Invasive species indeed cause significant problems, such as those described above, so scientists rightfully tend to dislike them for that. But they are certainly interesting from a practical standpoint, in terms of trying to predict the nature of their occurrences, impacts, and even interactions with humans (e.g. management, public). Because of the importance of invasions, we need to be able to predict where invasions might take place, what species might be involved, how they might get to their destinations, and the types and magnitudes of impacts they may cause. To these ends, the literature is rife with, for instance, transport vector maps and analyses (Muirhead and MacIsaac 2005; Herborg, O'Hara, Therriault 2009; Hulme 2009), analyses and predictions of animal and plant trade (Bertolino 2009; Humair *et al.* 2015), niche-based models (Thuiller *et al.* 2005; Broennimann and Guisan 2008; Rodda, Jarnevich, and Reed 2009; Goldsmit *et al.* 2018), and analyses of efficacy of prevention or management strategies (Briski *et al.* 2010; Chivers, Drake, and Leung 2017). Comparisons of functional response (how consumption rate of a predator changes with prey density) between invasive and non-invasive sister species have yielded insights on what makes species invasive, on invasive and native predator-prey interactions, and on the importance of context in impact prediction (Dodd *et al.* 2014; Bovy *et al.* 2015; Laverty *et al.* 2015). From these examples and others, there are many different intriguing ways scientists have made practical predictions about invasions. One method that is of relevance to this dissertation is the use of individual-based models (Section 2.3). Studies using pragmatic individual-based models which are designed to make practical predictions are discussed in Section 2.3.1.

Invasions provide interesting situations to try to understand or predict from theoretical ecological and evolutionary standpoints as well. Individuals are subject to a wide range of difficult challenges when they are introduced to a novel range. We can study the selection pressures on introduced populations, which can give us insights as to how species naturally expand their range or colonize new geographic regions. Phillips *et al.* (2006) showed that invasive cane toads in Australia were evolving longer legs, which they theorized at the time should aid in dispersal of the species. Subsequently, Phillips, Brown, and Shine (2010), showed that cane toads farther from the invasion origin (Cairns, Australia) also had a genetic predisposition for increased dispersal, helping to confirm their earlier hypothesis. We can also study what, from an evolutionary standpoint, contributes to invasiveness (i.e. the capacity for a species to become invasive). For instance, a review by Richards *et al.* (2006) discussed the roles of phenotypic plasticity in the success of invasive plants, citing evidence of and highlighting the differences between invasive and non-invasive species of three flavours – jack-of-all-trades (i.e., generalist), master-of-some (i.e., mostly specialist), and the jack-and-master which the authors introduced in that work (i.e. mostly generalist, but can exploit some favourable niche). Lodge (1993) proposed the possibility that invaders are better competitors in their invaded ranges because they evolved amongst stiff competition in their native ranges. On the other hand, there is evidence of the evolution of greater competitive advantage in invaders; comparisons between native and introduced conspecifics in some cases have shown that the introduced individuals are larger and more fecund (Siemann and Rogers 2001).

In a given introduction, typically the number of introduced individuals is low. This alone should set an enormous challenge, because this increases stochasticity and reduces the success of cooperative and density-dependent interactions (e.g. mating in sexual species, group feeding, niche modification). Allee effects are when fitness of individuals positively correlates with density at low population sizes. A critical question in invasion biology is, in the face of Allee effects and all these other challenges, what allows introduced populations to successfully establish (Sakai *et al.* 2001; Bock *et al.* 2015)? Kanarek and Webb (2010) showed theoretically that if populations can rapidly evolve to reduce Allee effects (via, for instance, increasing mate detection distance), evolutionary rescue could be possible when otherwise the population would go extinct. That study still begs the question of the possibility of this evolutionary rescue happening in the time frame during which invasions may take place. This question itself is difficult to answer as well, because even successful invaders may undergo a lag phase where population size is incredibly low (perhaps undetectably so) for an extended period (Sakai *et al.* 2001). This means that we often have no concept of the time it takes for an invasion to take place. Effects of genetic drift, the random shift in allele frequency potentially

resulting in the loss of alleles, are known to more strongly impact small than large populations (Bock *et al.* 2015). Interesting questions in this domain remain. What impact does genetic drift have on introduced populations? Do introduced populations undergo rapid shifts in allele frequency? Does this affect their establishment success? On a related note, as introduced populations are typically small, they should usually be subject to a genetic bottleneck (reduction in genetic diversity relative to their native population) as well. Reduced genetic diversity is expected to reduce the capacity for populations to adapt to novel or changing environments (Bock *et al.* 2015). A key question to ask is how then can introduced populations survive when their adaptive potential is so severely reduced? These types of theoretical questions are extremely difficult to answer studying real invasions for a multitude of reasons. Thus, as was possible with practical studies for biological invasions, some scientists have turned to individual-based models to make theoretical predications as well. Section 2.3.2 reviews the use of paradigmatic individual-based models to answer such theoretical questions related to ecology and evolution in invasions. Chapter 4 aims to elucidate the effects of genetic diversity of introduced populations on their establishment success using EcoSim.

A large body of research has also accumulated from attempts to manage biological invasions, both successful and failed. In a review of eradications in Europe, Genovesi (2005) noted that the easiest taxon to eradicate was mammals, while insects and plants were typically only successfully eradicated on islands but not on mainland. Genovesi also noted the importance of early detection of the invasions followed by subsequent rapid response, and that Europe was facing issues in achieving both of these. Pluess *et al.* (2012) conducted a meta-analysis of invasion eradication programs worldwide and over the last century and framed their analysis as a classification problem. They developed several decision trees to explain the factors affecting eradication success, finding that temporal and spatial extent of invasion were key predictive factors, along with whether sanitary measures (defined as banning transport of potentially infested materials) were taken. Such studies yield important insights of what factors affect success of eradications, but it is also important to know what strategies have worked for particular taxa – and there are plenty of such studies. In addition to looking back into historical eradication attempts, a proactive approach can also be taken to plan or test management efforts. Individual-based models can also help researchers plan and preliminarily test management strategies (see Section 2.3.1). Management clearly stands to strongly benefit from rapid response, but rapid response is predicated on early detection. Section 2.4 discusses the importance and means of early detection of nonindigenous species.

2.3 Individual-based Models

Among biological disciplines, behavioral ecology has a strong tradition of accounting for the role of organism-environment interactions in behavior (Krebs and Davies 1997). Behavioral ecology and the related field of optimal foraging theory (Stephens and Krebs 1986) model animal behavior in terms of optimal adaptation to environmental niches. The goal is not to test whether organisms actually behave optimally, but to use normative expectations to interpret behavioral data and/or generate testable hypotheses. One approach for understanding the behavior of complex ecosystems is through individual-based models (IBMs), which provide a bottom-up approach allowing for the consideration of the traits and behavior of individual organisms. Ecological modelling is still a growing field, at the crossroads between theoretical ecology, mathematics, and computer science (Ricotta 2000). Since natural ecosystems are very complex (in terms of number of species and of ecological interactions), ecosystem models aim to characterize the major dynamics of ecosystems in order to synthesize the understanding of such systems and allow for predictions of their behavior. Ecosystem simulations can also help scientists to answer theoretical questions regarding evolutionary process, the emergence of species, and the emergence of learning capacities. One of the most interesting aspects of such ecosystem simulations is that they offer a global view of the evolution of the system, which is difficult to obtain in nature. However, the scope of ecosystem simulations has always been limited by the computational possibilities of their time. Today, it is possible to run simulations that are more complex than ever due to the availability of high-performance computing resources.

Several individual-based ecosystem simulation platforms with various features exist. For example, Echo, one of the first such models, is a basic ecosystem simulation in which resources are limited and agents evolve (Hraber *et al.* 1997). In Echo, each agent, upon obtaining the required resources to copy its genome, replicates itself with some mutations. The agents, through interaction with other agents (combat, trade, or mating) or the environment, can acquire resources. Polyworld is another such IBM software (Yaeger 1994) to evolve artificial intelligence through natural selection and evolutionary algorithms. It displays a graphical environment in which trapezoidal agents search for food, mate, and create offspring. The number of agents is typically only in the hundreds, as each agent is rather complex and the environment consumes considerable computational resources. In this model, each individual makes decisions based on a neural network, which is derived from each individual's genome. Recently, Polyworld has been used to study the effects of different neuromodulation models on the adaptability of its individuals (Yoder and Yaeger 2014), finding that neuronal plasticity modulation (decreasing or increasing the rate at which neuron weights change) tends to produce individuals that adapt more effectively. It has also been used to study the way in

which network topologies influence the evolved complexity of the networks (Yaeger 2013) and, most recently, the level of chaos as the individuals in the system evolve (Williams and Yaeger 2017). Avida is another artificial life software platform for studying the evolutionary biology of self-replicating and evolving computer programs (Ofria and Wilke 2004), inspired by the Tierra system (Thearling and Ray 1994). Unlike Tierra, Avida assigns every digital organism its own protected region of memory and executes its program with a separate virtual CPU. A second major difference is that the virtual CPUs of different organisms can run at different speeds. The speed at which a virtual CPU runs is determined by several factors, but most importantly, by the tasks that the organism performs: logical computations that the organisms can carry out to reap extra CPU speed as a bonus. With increasing computational power, individual-based ecosystem simulation platforms such as Tierra, Avida, Polyworld, and EcoSim (Ray 1991; Lenski *et al.* 1999; Yaeger 1994; Gras *et al.* 2009) can be used to address increasingly difficult questions in biology (Lenski *et al.* 2003; Clune *et al.* 2008; Clune *et al.* 2011; Golestani *et al.* 2012). EcoSim (Gras *et al.* 2009), in particular, has been designed to model large-scale virtual ecosystems.

Recently, much has been done in the field of ecological IBMs on three main fronts: formalization and development practices of IBMs, pragmatic modelling, and paradigmatic modelling. With respect to formalization and development practices, some insist that there is an increasing need for developers of IBMs to be transparent about the process used to develop a model (Schmolke *et al.* 2010; Augusiak *et al.* 2014; Grimm *et al.* 2014). They argue that potential clients need to have a thorough understanding of the model so that they can know whether the model is applicable to whatever they would like to test. Clients need formal statements of the question(s) the model is designed to answer, descriptions of the submodels and their organization within the model, information on the degree of testing performed on the model, and the rationale behind making any modifications throughout the long and iterative process that is the "modelling cycle". So, several researchers have proposed and subsequently revised (Grimm *et al.* 2014) a new standard format for the description of an IBM, TRAnsparent and Comprehensive Ecological modelling documentation (TRACE) (Schmolke *et al.* 2010), which differs from the previously-proposed ODD protocol (Grimm *et al.* 2006) in that TRACE is more comprehensive and more concerned with describing the development cycle and practical ability of a model. Furthermore, the ODD protocol can be used within TRACE as a means of describing the model's implementation. TRACE complements the principle of "evaludation" (Augusiak *et al.* 2014), representing an urged evaluation and validation of a model throughout the development, application, and analysis of it. The current revision of TRACE intends to focus the developer on documenting the modelling process for the

sake of ensuring quality and credibility throughout said process, as the originally proposed TRACE was less efficient and less specific regarding its goals.

MacPherson and Gras (2016) argue that there is too much of a focus on "evaludation" and that not all IBMs are "merely adjunctive tools". More specifically, pragmatic models, focusing on a particular species or system, usually with the intent of making predictions in applied ecology, *should* undergo a more rigorous parameterization process using empirical data, be subject to evaludation, and be more stringently documented. Pragmatic models are often tied to conservation efforts or the management of delicate ecosystems, and so a model must be realistic enough to effectively predict how a specific (very complex) ecological system will behave. On the other hand, MacPherson and Gras (2016) argue that paradigmatic models are, in fact, experimental platforms. Though they must be realistic enough, in the general sense, there should be less of a focus on incorporating empirical data into the calibration or parameterization of them, as they are typically designed to answer rather general theoretical questions, the results of which we often have no means of historically validating due to the scale of interactions being emulated in the simulation. Furthermore, they argue that paradigmatic models can lose generalizability by over-calibrating the model empirically. They propose a relaxed notion of model evaluation by removing the constraint of empirical calibration; they instead insist that the calibration be "reasonable", that is, consistent with general observations in nature.

Pragmatic models are those that aim to model a specific system or population, and most IBMs are pragmatic in nature (DeAngelis and Grimm 2014). De los Santos *et al.* (2015), for instance, designed an IBM of a marine amphipod, *Gammarus locusta*, to assess the effect of long-term exposure to a chemical pollutant, aniline. They used real life-history traits of *G. locusta* to parameterize the model and observed significant negative impacts in individual survivorship and production of offspring with exposure to aniline. Other works in pragmatic modelling include a toxicological model for zebrafish (Hazlerigg *et al.* 2014), a model mediating effects of climate change on population dynamics in European anchovies (Pethybridge *et al.* 2013), a model for conservation and management of brown trout in Europe (Frank and Baret 2013), and a model for motion of the blue mussel, *Mytilus edulis* (de Jager *et al.* 2013). With respect to IBM usage in the field of biological invasions, pragmatic models are more common (see Section 2.3.1).

As the naming convention suggests, paradigmatic modelling moves away from answering questions about specific species or ecosystems and instead aims to uncover the underlying causes of more generalized ecological or evolutionary phenomena (DeAngelis and Grimm 2014). Zaman *et al.* (2014), for example, used Avida to show that parasite-host interactions increase the complexity and evolvability of digital organisms over a long time-frame. Avida has been used in several other recent works (Fortuna *et al.* 2013;

Goldsby *et al.* 2014; Ostrowski *et al.* 2015, LaBar *et al.* 2016). Similar to Zaman *et al.*, Kvam *et al.* (2015) also studied the complexity of the brain of a population of digital organisms, in this case Markov Brain agents. In contrast, they studied complexity in light of the problem-solving environment the agents were subject to. Olson *et al.* (2013) used Markov Brain Agents as well, but instead they placed the agents into a toroidal world and observed changes in physical cluster tightness when subject to different types of predator attacks. Botta-Dukát and Czúcz (2016) generated a spatially-implicit IBM to simulate community compositions and tested the ability of five functional diversity indices. Functional diversity indices aim to determine the number of functionally different species in a community. Their simulation accounted for habitat filtering (suitability of an individual to a habitat – a means of local trait convergence) and trait-similarity-based competition for resources (a means of local trait divergence) in composing the simulated communities. With mechanisms causing individual trait divergence and convergence, they could effectively test the functional diversity indices for their ability to detect these two key assembly processes. They found trait divergence was difficult to detect for all the indices tested, whereas trait convergence was detectable by some indices. Uchmański (2016) found, using an IBM, that dispersal mechanisms of individuals affect the persistence of metapopulations. In different runs of the simulation, individuals would disperse from their current habitat to another unoccupied neighboring habitat for different reasons (when one gains no resources due to competition, when competition yields insufficient resources to produce an offspring, random chance, or when no individuals in a habitat could reproduce). If individuals dispersed due to total loss of resources resulting from competition, the metapopulations persisted longest. Similarly, when individuals dispersed due to insufficient resources for reproduction, the metapopulations persisted longer than by chance. If individuals waited until none in a habitat could reproduce, the metapopulations failed to persist longer than cases in which dispersal was random. Another recent paradigmatic IBM tested the effects of patch size and refuge abundance on the strength of predator-prey interactions and population dynamics (Li *et al.* 2017). They found that refuge availability decreased the interaction strength between prey and predators, which consequently improved the stability of populations. CDPOP (Landguth and Cushman 2010) and its descendant CDMetaPOP (Landguth *et al.* 2017) are both IBMs that use Mendelian inheritance with any number of alleles and loci to study the effects of a varying landscape of (nearly) any complexity on the genetic structure and composition of populations or metapopulations. Though natural selection does occur, individual fitness is also influenced by user-specified spatially-explicit fitness values for each genotype that is selected upon. Paradigmatic models have been used in the study of biological invasions, but they are less common than pragmatic models (see Section 2.3.2 for examples).

We will now review the use of individual-based modelling in the study of biological invasions. We divide our discussion into two parts. We first discuss pragmatic models for biological invasions – which are quite common – and paradigmatic models for biological invasions – which are relatively rare. In this dissertation, EcoSim, introduced in Chapter 3, is used in Chapter 4 to determine how establishment success is impacted by genetic diversity of introduced populations.

2.3.1 Pragmatic Models for Biological Invasions

There are numerous examples of pragmatic models for biological invasions. Many of the early ecological IBMs were models of plants (e.g. JABOWA and its descendants; Botkin, Janak, and Wallis 1972) – and in invasion biology applications this was true as well (Higgins, Richardson, and Cowling 1996; Higgins and Richardson 1998; Higgins and Richardson 1999; Higgins, Richardson, and Cowling 2000; Buckley, Briese, and Rees 2003; Goslee, Peters, and Beck 2006). An early example was a model of invasive pine trees dating back to 1996 (Higgins, Richardson, and Cowling 1996) which was subsequently expanded upon (Higgins and Richardson 1998; Higgins and Richardson 1999). Their model aimed to elucidate the roles of a variety of factors in the spread of three invasive pine species in South Africa. Their model was spatially-explicit and time-discrete, with a two-dimensional 150x400 grid of cells (100x200 in the follow-up implementation) each representing approximately $100m^2$, and a time step of the simulation representing a year. Using their IBM was advantageous over reaction-diffusion models because they were able to study impacts other than the rate of spread, such as the mean density of pines and the mean perimeter of the invasion front. Also, IBMs, unlike reaction-diffusion models could account for a wide variety of individual interactions, stochastic effects, and spatial relationships – reaction-diffusion models lacked this. In the 1996 paper, they found that mean seed dispersal distance was a key factor across all impacts, though adult fecundity and age of reproductive maturity were impactful as well. Comparatively, though there was empirical evidence that wildfires aided in the invasion of these pines, there was little impact of fire-related factors (frequency and survival rate of adult pines) on the impact measures they considered. In the 1998 model, which also incorporated disturbance regimes and multiple environment types, showed the dominance of invasive *Pinus radiata* in shrubs and grasslands, while *Pinus strobus* dominated the forests. Their model also predicted that disturbance most strongly affected how easily invaded the shrub and grassland environments were, while forest environments were most resistant to pine invasions overall. Interestingly, high disturbance led to smooth invasion fronts while low disturbance levels generated scattered invasion fronts.

Pragmatic IBMs remain a useful option for modelling plant invasions (Travis *et al.* 2011; Murphy, Johnson, and Viard 2016; Xiao *et al.* 2016). Animal invasion modelling using IBMs became more prevalent later, likely because of the impetus to often incorporate more complex behavioural and biotic interaction models (Grimm and Railsback 2005). An interesting advancement in pragmatic modelling, both of plants and animals, was the incorporation of geographic information system (GIS) data in the models, such that very specific environments and locations could be accurately spatially modelled, for instance in terms of land use, obstruction, climate, or chemistry. An early study showed that an IBM could be useful in the preliminary testing of management strategies for the invasion of the American mink in conservation of the water vole in the UK (Bonesi, Rushton, and Macdonald 2007). American minks prey on water voles, and this interaction has led to the vole's protected status in the UK. Their model was spatially-explicit and time-discrete, with direct mappings to real space and time, using GIS data to directly model the UK's Upper Thames catchment and its surrounding region. The goal of their model was to determine the most effective trapping strategy for managing the invasive mink. In terms of trapping strategies, they varied number of traps and the seasonal temporal distribution of trap use, while also modelling immigration of minks from local regions. Unfortunately, when looking at control strategies aimed at vole conservation, only 24% of strategies led to 20-year conservation of the voles. Interestingly, their model determined that if female minks were held to a density 0.15/km, vole populations would generally persist. With high mink immigration, the strategy that prevailed was to trap in January, October, and November, whereas with low mink immigration, all strategies yielded similar effectiveness. The model showed that, barring unwavering and continuous commitment from conservation authorities, water vole extinction was imminent in the UK. Another model was produced to simulate management efforts in Brisbane, Australia, for the invasive red fire ant (Keith and Spring 2013). They had a probabilistic model for the spread of ant nests and their discovery by management personnel. After calibration of the model against a rich dataset of the invasion history, they used their model to predict the ultimate outcome of the current management regime. Similar to the above study of the American mink, they found that a significant increase of management effort would be necessary to successfully eradicate the ants and that the current strategy was insufficient.

Another team of modelers developed several IBMs for the invasion of the brown plant hopper in the Mekong River Delta of Vietnam (Phan, Huynh, and Drogoul 2010; Nguyen *et al.* 2011). The brown plant hopper is a pest species that destroys rice crops via consumption and the spread of various diseases. Similar to the above models, these were also spatially-explicit and informed by real GIS data. Their model covered a 700 square mile region of the delta, and their models were calibrated using real life-history and

physical characteristics of the brown plant hopper. Their model also featured simple models for rice crops. With their model, they were able to produce a parameterization that accurately replicated the invasion that had occurred to date – the pattern-oriented modeling approach (Grimm *et al.* 2005) – and they subsequently used the model to conduct predictions of future spread (Phan, Huynh, and Drogoul 2010). In following work (Nguyen *et al.* 2011), they added climatological parameters (e.g. wind, temperature, etc.) and land use information to the model, with the aim to determine the impact of the brown plant hopper on rice production. They also aimed to discover rice planting regimes that would lead to maximum rice yield in the face of the ongoing brown plant hopper invasion. They found that with the addition of the new factors, their model produced more accurate estimations of brown plant hopper spread. A key finding was that if farmers temporally staggered their rice planting regimes, they could minimize the impact of the brown plant hoppers on rice yields because the brown plant hoppers would be unable to spread to adjacent mature rice crops.

Interestingly, pragmatic individual-based models, even in early stages of development, can be used to engage and inform stakeholders while also providing future direction to both research, management, and even the general public (Samson *et al.* 2017). In this case, the modelers were producing an individual-based spread model, built on RangeShifter (Bocedi *et al.* 2014), for the round goby in and around the Baltic Sea. As mentioned earlier, biological invasions can affect people from all walks of life – in this case, the stakeholders consisted of management and government personnel, local recreational and professional fishermen, and others from the general public. There is typically a fear of losing credibility when presenting premature models to stakeholders, but in this case the development of the model took place alongside symposiums during which presentations of the model progress were delivered to stakeholders and feedback was received from them. In this case, developing the model alongside stakeholders allowed them to provide useful feedback – there were several qualitative parameters that the scientists were able to acquire from the stakeholders, several knowledge gaps were identified and filled in for all parties involved, and the modelers were able to determine the key focal points for the model in terms of what it should and need not predict. For example, the modelers were informed, through their early engagement with stakeholders, that gobies can disperse to deeper waters than they had originally thought, and particularly that this more often occurred during winter months. This had consequences for the model itself, and the scientists were able to incorporate this information into their model. Similarly, the stakeholders informed the scientists of the impacts of round gobies on the long-tailed duck via competition, which overwinters in certain habitats near the Baltic Sea. This allowed the researchers to identify sites where the long-tailed duck could overwinter, and inform conservationists of the predicted possibility and timing of spread

of round gobies to these sites. They used a pattern-oriented modelling approach to parameterize the model using historical invasion data in the Gulf of Gdansk, in conjunction with the use of empirical data and feedback from the stakeholders. Their model, in such early stages of development, did not produce results that generalized well to the invasion of the Baltic Sea – of course future work was needed to improve the model. Despite the poor generalization of the model, the study highlighted the importance of scientists working alongside stakeholders on equal footing in the context of biological invasions, as both parties typically have useful and different perspectives, and both parties stand to gain from each other's involvement. There are many other recent examples of pragmatic IBMs modelling animal invasions, owing to the usefulness of the approach in modeling and subsequently predicting interactions between invasive individuals, native individuals, their environment, and even management or the general public (e.g. Van Petegem *et al.* 2016; Yoann *et al.* 2016; Anderson and Dragićević 2018; Bonte and Bafort 2018; Day *et al.* 2018).

2.3.2 Paradigmatic Models for Biological Invasions

There are several examples of paradigmatic IBMs that were produced for investigating theoretical aspects of biological invasions. Studies investigated, for example, evolution of dispersal (Travis and Dytham 2002; Travis *et al.* 2009; Fronhofer, Poethke, and Dieckmann 2015; Henriques-Silva *et al.* 2015), tracking of shifting suitable habitats (Santini *et al.* 2016), the role of learning versus evolution in exploring novel environments (Sutter and Kawecki 2009), the role of sex structure of introduced populations in establishment (Shaw, Kokko, and Neubert 2018), and the spatial distribution of alleles during range expansion (Klopfstein, Currat, and Excoffier 2006; Travis *et al.* 2007; Burton and Travis 2008; Peischl and Excoffier 2015). Travis *et al.* (2009) developed a spatially-explicit time-discrete IBM to elucidate how dispersal strategies evolved differently for individuals in the stationary range versus the expansion front. Dispersal is the movement of an organism from the position where it was born to a new position where it reproduces – this has great importance for individuals on the expanding or contracting front of their species' range (e.g. due to novel colonization of a region, tracking a climatically-suitable spatial range, or other environmental changes). Their simulations took place on a 700 x 20 2D world, in which the invasive species was to expand its range along the x axis. The individuals in their model were haploid and asexual, having genotypes with three values that determine their density-dependent dispersal predispositions and fixed density-dependent behaviour for reproduction. Dispersal was stochastic, and the probability of dispersal was controlled by a sigmoid function which the three genes controlled – one value controlled the maximum dispersal probability, another controlled the thresholding location of the function, and the last one controlled the slope of the sigmoid curve. They ran the simulation in several ways. First,

they ran the simulation on a 20 x 20 world for 10000 time steps to observe the evolutionary equilibrium when the population was so physically constrained. They subsequently ran the model with the full 700 cells on the x-axis, starting at time-step 10001, aiming to observe if and how individuals closest to the invasion front (the five columns most recently expanded upon) differed evolutionarily from the others. They also conducted the same experiments, but for comparison, the individuals' dispersal was density-independent – dispersal probability was simply modelled explicitly as a single value. Their simulations in the restricted world showed that the genes for thresholding abundance and maximal dispersal were selected upon most strongly, while the gene controlling the slope of the curve was weakly selected upon. It also showed that a variety of dispersal strategies were viable. With the expansion range open, they found that individuals along the expansion front differed genetically from those in the stationary range, regardless of whether dispersal was density-dependent or independent. Their model also yielded different evolutionary strategies for density-dependent range expansion – typically one of the three genes differed dramatically such that it favoured dispersal more often. Finally, they found that during range expansion, dispersal at low abundance was favoured and typically higher maximum dispersal probability evolved, even despite increased cost of dispersal, which their model allowed them to easily test. With a relatively small-scale IBM, these authors were able to uncover novel evolutionary differences between individuals in the invasion front versus those in the stationary range of an invasion. There have been several other works investigating the evolution of dispersal using IBMs (e.g. Fronhofer, Poethke, and Dieckmann 2015; Henriques-Silva *et al.* 2015). In fact, a review by Hargreaves and Eckert (2014) noted that the vast majority of models of dispersal are paradigmatic, spatially-explicit IBMs in which demographics and patch occupancy are considered. Interestingly, though many models considered range expansion, few models considered the dynamics of population genetics along a contracting range. They also noted that most models have haploid, asexual individuals, limiting their applications in studying how mating systems might evolve alongside dispersal strategies.

Another example of a paradigmatic IBM studying invasions is that of Sutter and Kawecki (2009). Few theoretical works previously aimed to establish links between learning capacity and evolution aiding in the ability of populations to expand to novel environments. There was limited support of the idea that learning could be facilitated by evolution, and equally limited support of the theory that learning could suppress evolution. Individuals in their model lived on a 1D ring world, consisting of two environments (one in which the individuals began and another in which the individuals could expand to) and two transition regions. The environments were similar in that they each had limited resources of the same quantity, denoted A and B, but the resources were

of opposite and inversely-related quality in each environment. The transition zones created a gradient such that there was no spatial bias in their world. Their model had diploid individuals with eight binary-encoded loci, and the sum of the alleles determined preference for one of the two resources. Individuals in their model underwent several rounds of feeding, during which their learning rate allowed them to modify their preference as they fed in their location. After feeding, survival was computed, being density-dependent and related to the amount of energy accumulated from the feeding phase. Lastly, those that survived were able to reproduce, during which all individuals were considered hermaphroditic. Upon initialization, all individuals were placed in the starting environment with perfect preference for the high-quality resource in that environment. Over the course of simulations, they experimented with the learning rate and environmental parameters to determine how learning influenced the evolution of the individuals as they expanded their range from the starting environment. They ran a number of simulations for a wide range of parameter combinations, stopping the simulations at 5000 time steps or when an invasion across the novel environment was considered successful. Their simulations showed that rapid expansion was favoured in environments with a gradual transition between habitats, with individuals having high learning ability, and with weak selection on resource preference. They also found that learning had little effect on expansion with a gradual transition between habitats, whereas with a sharp transition between environments learning was extremely important (with the converse being true for evolution). They also determined that genetic diversity was greatest in local populations along transition zones, while genetic diversity was reduced in local populations in non-transitional regions. Finally, their research showed that learning reduced the degree to which populations evolved and also increased the amount of genetic diversity within the populations.

Several related studies have been conducted investigating the spatial distribution of alleles through populations undergoing range expansions. Klopfstein, Currat, and Excoffier (2006) built on a pre-existing paradigmatic IBM (SPLATCHE; Currat, Ray, and Excoffier 2004), to construct an IBM in which an initial population of haploid individuals expanded horizontally along a lattice subject to cell carrying capacity, inter-cell migration rate, and population growth rate. Their modification allowed users to specify the location and timing of the rise of a mutant neutral allele. This allowed the researchers to track the spread of this allele as the population experienced range expansion. Interestingly, over all parameter sets they tested, the spatial distribution of the neutral allele had a common characteristic – its density was highest at the location parallel to its initialization and nearest to the expansion front (a phenomenon they called expansion wave surfing). They were also able to determine that the probability of neutral allele survival and surfing was dependent on the characteristics of the population – low

carrying capacity, high migration, and high population growth rate led to increased rates of survival and surfing of the mutations. Lastly, their study showed that the neutral mutation propagated more effectively if it occurred shortly after the arrival of the expansion front to the target location – if the development of the mutation was delayed, it was unable to surf the expansion wave. A follow-up by Travis *et al.* (2007) found that deleterious mutation can also surf the expansion wave, but equally as intriguing, beneficial mutations cannot. The findings of Travis *et al.* (2007) with respect to the neutral and deleterious mutations were very similar to findings of Klopfstein, Currat, and Excoffier (2006) regarding the neutral mutation – the same relationships with respect to carrying capacity, migration rate, and intrinsic growth rate were discovered. However, they found that although the beneficial mutation did reach high abundance in the expanding populations, it lacked the propensity or necessity to surf the expansion wave. On the other hand, neutral and especially deleterious mutations needed to surf the expansion wave in order to persist. The authors also tested two types of mutations – those affecting survival and those affecting the number of offspring an individual could produce. Though both types of mutants experienced the same fate, the mutations affecting reproduction rate directly impacted the rate of expansion. The deleterious mutations, in this case, further relied on surfing as a means of propagation – but as the mutation itself impacted the rate of propagation of the population (i.e. via range expansion), the likelihood of survival of the mutation decreased drastically. The findings of both studies highlight an extremely interesting phenomenon that likely has importance in invasion biology – keeping such a spatial distribution of neutral and even deleterious mutations could aid in maintaining higher genetic diversity nearer to the invasion front, which they theorized could be instrumental in helping individuals reach fitness peaks. Another follow-up by Burton and Travis (2008) explored this theory with a similar experiment; they found that not only can the surfing of neutral and deleterious alleles maintain genetic diversity on the invasion front, but it could also help subpopulations on the invasion front to cross adaptive valleys and reach new fitness peaks in genomes exhibiting sign epistasis (where presence of some mutant allele in some genetic context is deleterious, but in other contexts is beneficial).

2.4 Early Detection of Invasive Species

As noted by Simberloff *et al.* (2013) and many others cited above, the most cost-effective and successful form of dealing with invasions is to prevent them outright; both efficiency and effectiveness decrease as an invasion progresses. Unfortunately, it is not always possible to prevent introductions, but it is still clearly important to detect them early and to respond to them as quickly as possible. So important, in fact, that the four research themes of the Canadian Aquatic Invasive Species Network (CAISN) – selected in collaboration with federal government agencies – included a theme of early detection and

another of rapid response. Surveillance, and particularly early detection, is imperative in successfully thwarting an invasion; however, when the number of introduced individuals is low it is difficult to detect their presence using traditional methods such as trawling, recruitment plates, netting, trapping, fishing, or electrofishing (Harvey, Qureshi, and MacIsaac 2009; Smart *et al.* 2015). Thus, a major advancement in surveillance of introduced species is the use of molecular methods for early detection (Ficetola *et al.* 2008; Jerde *et al.* 2011). Molecular methods could include the detection of metabolites or RNA (Pochon *et al.* 2017), but DNA-based detection approaches have the advantage of the persistence of DNA in the environment and throughout processing (Barnes *et al.* 2015; Pochon *et al.* 2017), and currently the taxonomic resolution DNA provides due to large databases cataloguing specific markers for a wide variety of species (Hebert *et al.* 2002; Ratnasingham and Hebert 2013).

DNA-based methods have been proposed for some time, but the advent of barcoding and metabarcoding (Fonseca *et al.* 2010; Ratnasingham and Hebert 2013) made these methods a reality. Barcoding involves the use of small genetic 'barcode' regions (Hebert *et al.* 2003) to assign taxonomy to genetic data. Metabarcoding extends the use of this barcoding approach to samples of mixed sequences, aiming to assign taxonomy usually to particular subsets of taxa in the sample. In conjunction with the arrival of next-generation (high-throughput) sequencing technology, a vast reduction in the resource requirements (i.e. time and money) to conduct these methods have rendered these genetic methods viable options for early detection (Zhan *et al.* 2013). Barcoding and metabarcoding approaches could be used in the context of detection of environmental DNA (eDNA), i.e. the detection of genetic matter from for instance shed skin, secretions, gametes, and carcasses in an environment (Ficetola *et al.* 2008). They could also be used alongside traditional methods such as netting or trawling (e.g. Chain *et al.* 2016); such traditional methods often rely on morphological taxonomic assignment, and many species can be morphologically variable throughout their lifespan or just difficult to identify. It is also possible that partial organisms are contained in a given sample, rendering morphological assignment impossible. Genetic methods could be used in place of morphological assignment in these traditional methods to greatly increase their sensitivity, accuracy, and efficiency. Metabarcoding, particularly, is an interesting prospect in early detection of invasive species because it can be used on huge, diverse eDNA or bulk samples to efficiently determine the presence of a wide range of taxa (Ficetola *et al.* 2008).

Metabarcoding is not without its issues; despite improvements in the associated technology (e.g. sequencers) there are a huge number of potential sources of error in the barcoding or metabarcoding process for early detection of nonindigenous species (Xiong, Li, and Zhan 2016). These sources include sample contamination, primer selection,

marker selection, sequencing, and incomplete reference databases. To reduce the frequency and impact of these errors, computational processing of the sequence data is typically employed, for example to filter out likely erroneous sequences, group spurious reads with likely correct representative reads, and cluster groups of reads together to form operational taxonomic units (OTUs). Unfortunately, even this computational processing can be a source of error. A false positive error, for instance, might stem from falsely considering an erroneous read correct and assigning it to a species that does not exist in the sample. A false negative error, for example, might occur when a correct sequence is wrongly filtered out with no other representative sequences from the species in the sample. Clearly, either type of error (false positive or false negative) could be detrimental to early detection of nonindigenous species. The effects of this computational processing, and the best way to conduct it, are poorly understood but imperative to understand if this technology is to be used in practice. Chapter 5 consists of a study we conducted under the CAISN to aid in the computational processing of sequences, applied to conducting species richness assessments and early detection of nonindigenous species using the metabarcoding approach.

2.5 Conclusion

Biological invasions are extremely important but also difficult to study, often necessitating the use of computer simulations to generate data that are too difficult (sometimes impossible) to obtain naturally. These data can be used to make predictions regarding spread, interactions of native and introduced species involved, impacts caused by the introduced populations, and the effectiveness of control strategies. In this dissertation, a paradigmatic predator-prey ecosystem individual-based model called EcoSim was produced (Chapter 3). Subsequently, several novel variants of EcoSim were developed to determine how genetic diversity affects establishment success of introduced populations (Chapter 4). Finally, a simulation using real genetic data was employed to determine optimal parameterization of a sequence processing pipeline for usage in detecting aquatic invasive species and estimating species richness, and then test its performance (Chapter 5). This dissertation highlights the importance, effectiveness, necessity, and viability of usage of computer simulations for the study of biological invasions, ecology, and evolution.

CHAPTER 3

EcoSim

3.1 Introduction

EcoSim is a large-scale evolving predator-prey paradigmatic ecosystem simulation that can be used to perform studies in theoretical biology and ecology (Golestani *et al.* 2012; Mashayekhi *et al.* 2014). It has been shown that EcoSim generates patterns of complexity similar to those observed in real ecosystems (Golestani and Gras 2010). Several studies have been done using EcoSim. Devaurs and Gras (2010) have shown that the behavior of this model is realistic by comparing the species-abundance patterns observed in the simulation with real communities of species. Furthermore, chaotic behavior (Golestani and Gras 2010) and multi-fractal properties (Golestani and Gras 2011) of the system have been demonstrated to be similar to those in real ecosystems (Seuront *et al.* 1996), and Golestani, Gras, and Cristescu (2012) measured the effect of small geographic barriers on speciation in EcoSim. The effect of the spatial distribution of individuals on speciation was investigated by Mashayekhi and Gras (2012). Khater *et al.* (2014) demonstrated that introduction or removal of predators in an ecosystem can have widespread effects on the survival and evolution of prey by altering their genomes and behavior. Mashayekhi *et al.* (2014) showed that the extinction mechanisms in EcoSim are similar to those of real communities. Lastly, a study by Gras *et al.* (2015) used EcoSim to explore the roles of natural selection and spatial isolation in the speciation process. They were able to unequivocally demonstrate that in order to observe genetic clusters (species), natural selection must be present. The number of individuals per species was much greater, species abundance distributions were far more even, the compactness and separation of genetic clusters were far greater, and hybrid production was far lower (after sufficient time had passed in the simulation) in runs where natural selection was present.

Real ecosystems are extremely complex systems with numerous interacting components and feedback loops. No paradigmatic model has all of the features of real ecosystems; consequently, these artificial systems are restricted to a small spectrum of possible questions they could answer. EcoSim was already quite complex and diverse in the types of questions it could answer, but we have added specific features to further improve its realism and applicability. Our objective is to propose to the community an improved simulation platform that models as many of the important features of real ecosystems as possible. Of course, not every significant feature of real ecosystems could be integrated into such a simulation platform. However, we have chosen a set of features that seem most important in modelling a stable, long-term evolutionary ecosystem and providing the mechanisms needed to answer the largest possible spectrum of important theoretical questions. The three most important features we have added to EcoSim are

fertilization of soil via animal excretion, the ability of prey to defend against attacking predators (individually or cooperatively), and a female/male binary sex system with sexual reproduction. In previous iterations of EcoSim, individuals were of uniform sex and any two individuals of the same type (prey or predator) could attempt to reproduce.

There is a vast array of indirect impacts of herbivores on plant community features (Augustine and McNaughton 1998; Olff and Ritchie 1998). Most importantly, herbivores affect the quantity and quality of organic matter returning to the soil (Hobbs 1996; Bardgett *et al.* 1998; Bardgett *et al.* 2003; Wardle 2002). Generally, animal excreta facilitates decomposition through increasing soil microbial biomass (Frank and Evans 1997; Bardgett *et al.* 1998) and net Carbon (C) and Nitrogen (N) mineralization (Molvar *et al.* 1993; Frank and Groffman 1998). Feces and urine also make it easier for plants to absorb, thereby increasing their growth rates (Hamilton and Frank 2001). Thus, herbivores are able to influence their own food supply (Hik and Jefferies 1990; Drent and Van der Wal 1999; Van der Wal *et al.* 2004) by producing negative feedback against the reduction of resources they consume. In order to include this complex feedback mechanism, we introduced a new concept to our simulated ecosystem called "fertilizer", which models the effect of prey fertilizing their environment.

There is limited experimental evidence in the ecological literature regarding mobbing behavior as a kind of reciprocal altruism between heterospecifics. Krams *et al.* (2006) and Krams *et al.* (2008) report that breeding *Fecedula hypoleuca* (pied flycatchers) engage in mobbing behavior primarily with heterospecifics as a form of defense against predation. As Krams *et al.* (2006) note, there is little empirical evidence for the existence of mobbing behavior as a form of reciprocal altruism. EcoSim could thus be used to test for mobbing behavior as a form of reciprocal defense in the presence of predation. In a related vein, an important unresolved debate in the biological literature is whether eusociality evolved via kin selection or group selection; Nowak *et al.* (2010) claim that group selection rather than kin selection (inclusive fitness) combined with haplodiploidy theory is the best way to explain eusociality. They suggest that there may be no real relation between haplodiploidy and eusociality, and they argue that inclusive fitness theory is not sufficiently general since it is a simple mathematical theory that has great limitations (Nowak *et al.* 2010). Furthermore, Nowak *et al.* (2010) argue that there is no empirical confirmation of inclusive fitness theory. On the other hand, Marshall (2016) and Abbott *et al.* (2010) argue that recent evidence helps to support inclusive fitness theory. Since there is apparently an argumentative stalemate regarding whether kin selection or group selection drives evolution, EcoSim could help to resolve this debate by testing the hypothesis that kin selection explains the evolution of eusociality and altruism. Finally, another important issue in evolutionary theory is whether predation selects for morphological defenses in prey. Bollache *et al.* (2006) argued that the main

reason that the invasive amphipod, *Gammarus roeseli* was eaten less than the native amphipod species *Gammarus pulex* was due to the presence of a spine on *G. roeseli*, as opposed to behavioral differences. EcoSim could be used to help resolve the debate regarding whether morphology or behavior is a key inducible defense against predators.

Typically, in sexually reproductive species in which sexual dimorphism exists, females are generally choosier than males when selecting mates. Compared to males, females typically invest far more resources (time and energy) into offspring. For instance, females typically provide more parental care than males. Females also invest more in gametes for sexual reproduction; males produce the microgamete sperm, whereas females produce large, nutritious eggs. Moreover, unlike males, females only produce a limited number of eggs as long as they are reproductively active; therefore, there is more risk associated with mate choice (Andersson 1994). To broaden the applicability and increase the realism of EcoSim, we introduced a model for sexual reproduction into the simulation. Previously, there was no categorization of individuals by sex; any individual could attempt reproduction with any other of the same type (prey with prey, predators with predators). Now, prey and predator individuals are divided into two groups, males and females. Furthermore, we have made significant modifications to reproduction mechanisms such as selection of mates, energy dynamics, and genetic recombination; these changes reflect the information-gathering and decision-making process that is mate choice (Bateson 1983). These new improvements were aimed at unravelling some of the most complex issues in behavioral ecology, such as the evolution of female preference.

In addition to presenting the new version of EcoSim following the updated 7-points Overview, Design concepts, and Details (ODD) standard protocol (Grimm *et al.* 2006, Grimm *et al.* 2010), we present and discuss data from EcoSim in its default configuration. We also analyze the divergence of two sister species in EcoSim. We then present a sensitivity analysis on three parameters of EcoSim: the amount of energy spent per time step for prey and predators, the maximum amount of grass held in cells, and the initialization of newly added social concepts related to defense. The purpose of this sensitivity analysis was to show how sensitive or robust EcoSim is to these parameters. Finally, we present a case study of EcoSim's application; we determined the behavior and evolution of individuals under two conditions: reduced primary production (thereby increasing competition) and reduced energy expenditure. This study serves as an example of the types of study that are made possible by the EcoSim platform.

3.2 ODD Description

EcoSim is an individual-based ecosystem simulation (Gras *et al.* 2009; Mashayekhi *et al.* 2014.b) for simulating animals' behaviors in a dynamic, evolving ecosystem. The individuals of EcoSim are prey and predators acting in a simulated environment. A

description of the older version of EcoSim can be found in Mashayekhi *et al.* (2014.a, 2014.b). In addition to the main features outlined above, EcoSim has been expanded by adding several smaller features, such as: new individuals' perceptions of their environment, new actions, new physical traits (governed by what we call the physical genome), sex-linked genes, various modes of reproduction, modified acting priority for individuals, new ways to control the dynamics of the environment, and new crossover and mutation operations that consider an individual's sex. Below, we describe the new version of EcoSim following the updated 7-points Overview, Design concepts, and Details (ODD) standard protocol (Grimm *et al.* 2006; Grimm *et al.* 2010). EcoSim source code (in C++) can be obtained from the repositories at https://github.com/EcoSimIBM, and more information can be found at https://sites.google.com/site/ecosimgroup/home.

3.2.1 Purpose

EcoSim was designed to simulate animal behavior in a dynamic and evolving ecosystem. The main purpose of EcoSim is to study biological, ecological, and evolutionary theories by constructing a complex adaptive system that leads to a generic virtual ecosystem with behaviors like those found in nature. Due to the complexity, scale, and resource requirement of studying these theories in real biological systems, simulations of this nature are necessary. EcoSim uses a fuzzy cognitive map (FCM; Kosko 1986) to model an individual's behavior. Since the FCM is coded in the genome and heritable, behavior can evolve during the simulation. Importantly, the fitness of a given set of behaviours and physical traits is not pre-defined. Instead, fitness emerges from interactions between the model organisms and their biotic and abiotic environment.

3.2.2 Entities, State Variables, and Scales

Individuals

EcoSim has two types of individuals: prey and predators. Each individual possesses two types of traits: acquired and inherited traits (Table 3.1). The former varies depending on the environmental conditions and the latter is encoded in an individual's genome and is fixed during its lifetime. The *Age* (number of time steps that the individual has been alive in the simulation) and *Speed* (number of cells the individual moved in a given time step) are initialized to zero for newborn individuals, while energy, a crucial property of the individual, is initialized based on the amount of energy invested into a newborn by its parents at reproduction time (*State of Birth* or *SOB* – see *Reproducing* under *Submodels*). Afterward, energy is provided to the individuals by resources (food) they find in their environment. Prey consume grass, which is dynamic in quantity and location (see *Submodels* for grass diffusion model), whereas predators hunt for prey individuals or

scavenge their remains when they die. *Strength* of an individual is calculated based on its current energy (*Energy*), maximum energy (*MaxEnergy*), age (*Age*), maximum age (*MaxAge*) and reproductive age (*RepAge*). Young (*Age* is less than *RepAge*) and old individuals (*Age* is greater than or equal to *MaxAge* minus *RepAge*) have less *Strength*. *Strength* can range from 25% of an individual's *MaxEnergy* (if the individual is too young or old and has energy approaching zero) to 100% of the individual's *MaxEnergy* (if the individual has energy greater than or equal to 1/3 of its *MaxEnergy* and the individual is not too young or old).

Table 3.1. Several physical and life history characteristics of individuals from five independent runs. The values for the inherited features are the values at initialization, and for the acquired features they are the average values over 20,000 time steps.

Type	Characteristic	Male Predator	Female Predator	Male Prey	Female Prey
Inherited	Maximum Energy	3000	3000	2500	2500
	Maximum Age	50	50	46	46
	Vision	20	20	8	8
	Maximum Speed	20	20	6	6
	Minimum Age of Reproduction	5	5	6	6
	State of Birth	14	18	12	16
	Defense	N/A	N/A	0.05	0.05
	Cooperative Defense	N/A	N/A	0.05	0.05
Acquired	Average Energy	2312.2	2211.4	1664.9	1678.3
	Average Age	16.5	13.7	14.3	12.3
	Average Speed	3.4	2.9	6.5	6.0
	Average Strength	3306.3	3107.9	2478.9	2439.7

Each individual performs one unique action during a time step, based on its perception of the environment and state (see *Emergence* under *Design Concepts*). At each time step, each individual spends energy depending on its selected action (*e.g.*, reproduction, eating, moving), the complexity of its behavioral model (number of existing edges in its FCM; see *Adaptation* under *Design Concepts* for details), and its physical characteristics (encoded in its physical genome; see *Adaptation* under *Design Concepts* for details). To achieve a realistic rate of energy expenditure we involved as many of its contributory factors as possible and used empirically-determined physiological scaling rates (see Eq. 3.1, per time step energy penalty for prey, and Eq. 3.2, per time step energy penalty for predators). In general, any action performed by a living organism is involved in spending some amount of energy (Butler *et al.* 2004), dependent on what the action is (Blaxter 1989). Thus, the action performed was included as a contributing factor in energy expenditure (Eqs. 3.1 and 3.2). Moreover, the size of a living organism plays a fundamental role in its metabolic rate (Chapman and Reiss 1999).

In EcoSim, the size of each individual is modelled through its *MaxEnergy* and *Strength*. *MaxEnergy* is a heritable limit on an individual's capacity to store energy, whereas *Strength* is a slightly more complex proxy of size, being derived from an individual's *MaxEnergy*, *Energy*, and *Age*. Experimental and empirical investigations have demonstrated that there is a nonlinear relationship between an adult animal's body mass and their metabolic rate, which is best described by a ¾ scaling exponent (Kleiber 1932; Hemmingsen 1960; Kleiber 1961; Stahl 1965; Stahl 1967; Pedley 1977; Prothero 1979; Schmidt-Nielsen 1984; Peters 1986; Niklas and Enquist 2001). Consequently, the metabolic rate of an individual in EcoSim is quantified through a power function of coefficient ¾ on its *MaxEnergy* (Eqs. 3.1 and 3.2). Energy expenditure associated with movement is also modelled in EcoSim using the kinetic energy equation (KE), and here we use *Strength* as a proxy of mass (KE = mass × speed^2, Eqs. 3.1 and 3.2). The complexity of an organism's behavioral model increases an individual's energy expenditure, because it has been accepted that species belonging to a higher-level taxonomic affiliation require more energy to survive (Mueller and Diamond 2001; Nagy 2005). Individuals with a larger brain also require more energy, as the brain is an expensive organ in terms of specific chemical and thermoregulatory needs (Wheeler 1984; Falk 1990). Consequently, possessing a large brain leads to a heavier metabolic requirement (Safi *et al.* 2005). The complexity and the size of the brain vary in different species; while some species possess a very simple and small brain, many higher vertebrates have a brain so large and complex that it is considered the most complex organ in these species (Shepherd 1994). Therefore, we also include this parameter in calculating the energy spent by an individual. Taking these points into consideration, the energy spent by prey (1) and predators (2) at any time step is given by the following equations:

(3.1)

$$\begin{aligned} &Energy\ Spent\ by\ Prey \\ &= \left(0.8 \times max((NbArcs - 100)^{0.75}, 1)\right) + \frac{(Strength \times Speed^2)}{10000} \\ &+ \left(\frac{MaxEnergy}{5.5}\right)^{0.75} + (Vision \times 5.0)^{0.75} + (MaxSpeed \times 5)^{0.75} \\ &+ (Defense \times 100)^{0.75} + (CoopDefense \times 75)^{0.75} + (max(0, 8 - RepAge))^{2.3}, \end{aligned}$$

(3.2)

$$\begin{aligned} &Energy\ Spent\ by\ Predator \\ &= \left(0.8 \times max((NbArcs - 130)^{0.75}, 1)\right) + \frac{(Strength \times Speed^2)}{11000} \\ &+ \left(\frac{MaxEnergy}{5.5}\right)^{0.75} + (Vision \times 5.0)^{0.75} + (MaxSpeed \times 5)^{0.75} \\ &+ (max(0, 7 - RepAge))^{2.3}, \end{aligned}$$

where *NbArcs* is a measure of the complexity of the individual's brain based on the number of edges in its FCM (see *Adaptation* under *Design Concepts* for details), *Vision* refers to the distance up to which the individuals can see (which is initially 8 cells for prey and 25 cells for predator), *Defense* quantifies the ability of the prey individuals to protect themselves when they are attacked by predators, *CoopDefense* quantifies the ability of a prey individual to protect other prey in its cell, and *RepAge* is the *Age* at which the individuals can start reproducing.

All individuals first perceive their environment (all the surrounding cells in their vision range) before using their behavioral model to choose a single action (see *Emergence* under *Design Concepts* for details of how individuals choose actions). After perceiving its environment (including grass resources, prey, predators, *etc.*), the possible actions for a prey individual are: evade (escape from predator), search for food (if there is not enough grass available in its cell, move to another cell to find grass), socialize (move to the closest prey in the vicinity, move to the cell with strongest prey, move to the cell with the greatest total prey *Strength*, or move to a cell with the least total prey *Strength*), explore, rest (to save energy), eat, or reproduce. Predators also perceive their environment to gather information used to choose an action among: hunt (to catch and eat a prey), move to the cell with strongest prey, move to the cell with the least total prey *Strength*, move to the cell with the weakest prey, search for food, socialize (move to the closest predator in the vicinity or move to the cell with strongest predator), explore, rest, eat, or reproduce. See the *Submodels* section for a full description of actions. Every individual takes one action per time step, after which its energy level and *Strength* are adjusted. The *Age* of all individuals is also increased by one unit at each time step. In addition to the acquired physical traits mentioned above, each individual has many state variables that, together, represent its state of mind. These variables are the values held in the nodes of each individual's FCM. Each FCM node has a single value that is its activation level (degree of stimulation) of its represented concept. Concepts can either be sensory, such as the individual's perception of local food, internal, such as the individual's hunger, or action, such as the individual's willingness to perform the eat action (see *Emergence*, *Adaptation*, and *Submodels* for more information).

Time Step

Each time step involves each individual perceiving its environment, making a decision, and performing one action. In addition, species memberships are updated and all relevant variables (*e.g.*, quantity of available grass) are recorded (see *Process Overview and Scheduling* for algorithm).

Cells and Virtual World

The smallest units of the environment are cells. Each cell represents a large space, which may contain an unlimited number of individuals, some limited amount of food, and some limited amount of fertilizer. The number of individuals a cell can host, therefore, is indirectly limited by the amount of food a cell contains. There are two types of food: grass, which only prey can eat, and meat, which only predators can eat. Grass amounts are controlled by a grass diffusion and growth model, and meat is generated when predators kill prey (see *Submodels* for grass diffusion model and meat generation). Fertilizer is produced by individuals residing in a cell (see *Submodels* for fertilizer dynamics). The virtual world consists of a matrix of 1000×1000 cells. The world is large enough that an individual moving in the same direction over the course of its entire life could not even cross half of it, and thus high-level movement patterns can be observed. The virtual world wraps around to remove any spatial bias. In addition, the dimensions of the world are adjustable, but expanding the dimensions increases the computational requirements (time and memory) of the simulation.

Species

By default, numerous prey and predators coexist in the simulation at any time step. Alternatively, the simulation can be run without predators. For each type, there is some number of species determined by the genetic makeup of the sets of individuals. There is at least one prey species and one predator species unless an extinction occurs, and at most there can be one species per individual. A species is a set of individuals with sufficiently similar genomes (see *Collectives* under *Design Concepts* for more details about speciation).

3.2.3 Process Overview and Scheduling

At each time step, the value of the state variables of individuals and cells are updated. The overview and scheduling of every time step is as follows:

1. For prey individuals:
   1.1. Perceive environment
   1.2. Compute next action
   1.3. Increase *Age*
   1.4. Females that chose to *Reproduce* act in order of decreasing *Strength* (to simulate female choice in mate selection)
   1.5. Remaining prey act in order of decreasing *Strength*
   1.6. Update list of prey (as some may have died due to depletion of *Energy* or maximum *Age*)

2. For predator individuals:
    2.1. Perceive environment
    2.2. Compute next action
    2.3. Increase *Age*
    2.4. Females that chose to *Reproduce* act in order of decreasing *Strength* (to simulate female choice in mate selection)
    2.5. Remaining predators act in order of decreasing *Strength*
    2.6. Update list of predators and prey (for predators, some may have died due to depletion of *Energy*, maximum *Age*, or combat with prey; for prey, some may have died due to predation)
3. Sort prey in order of decreasing *Strength*
4. Sort predators in order of decreasing *Strength*
5. Update prey species
6. Update predator species
7. For every cell in the world
    7.1. Update *Fertilizer* level
    7.2. Update *Grass* level
    7.3. Update *Meat* level

The complexity of the simulation algorithm is mostly linear with respect to the number of individuals. If we consider that there are N_1 prey and N_2 predators, then the complexity of parts 1 and 2 of the above algorithm, including the clustering algorithm used for speciation, will be $O(N_1)$ and $O(N_2)$, respectively (Aspinall and Gras 2010). The sorting parts (parts 3 and 4) have a complexity of $O(N_1 log(N_1))$ and $O(N_2 log(N_2))$ but are negligible in computational time so we will exclude them from the complexity computation. The complexity of parts 5 and 6 will be $O(N_1 + N_2)$. The default virtual world of the simulation has 1000 x 1000 cells, therefore the complexity of part 7 will be $O(k = 1000 \times 1000)$. As a result, the overall complexity of the algorithm is $O(2N_1 + 2N_2 + k)$, which is $O(N = 2N_1 + 2N_2)$. In terms of computational time, the speed of simulation per time step is related to the number of individuals. Recent executions of the simulation produced approximately 20,000 time steps in 60 days.

3.2.4 Design Concepts

Basic Principles

The genome of each individual consists of two parts: a physical genome and a behavioral genome. An individual's genome is fixed at birth. When a new offspring is created, it receives a genome that combines the genomes of its parents with some possible mutations. An individual's physical genome determines its physical characteristics and its behavioral genome determines its behavioral characteristics. An individual's physical genome comprises values that represent its physical attributes (see Table 3.1, inherited traits).

The behavioral model of each individual is encoded as an FCM (Gras *et al.* 2009) (Figure 3.1). Formally, an FCM is a directed graph that contains a set of nodes C and a set of edges I (Figure 3.1; Kosko 1986). Each node C_i represents a concept and each edge I_{ij} represents the influence of the concept C_i on the concept C_j. A positive weight associated with the edge I_{ij} corresponds to an excitation of the concept C_j from the concept C_i, whereas a negative weight represents inhibition. A zero value indicates that there is no influence of C_i on C_j. The edges of an FCM can be represented by an $n \times n$ matrix, L, in which n is the number of concepts and L_{ij} is the influence of the concept C_i on the concept C_j. If $L_{ij} = 0$, there is no edge between C_i and C_j. An individual's behavioral genome is its set of FCM edges (its matrix L). Since the edges of the FCM are encoded in the genome, the behavioral model is heritable, mutable, and subject to evolution. Individuals act at each time step by using their FCM to compute their action (see *Emergence*). The activation level (degree of stimulation) of each concept, represented as the value held in its corresponding node, is dynamic in each individual. Collectively, the activation levels of every one of an individual's nodes represent the individual's behavioral state. In each FCM, three kinds of concepts are defined: sensory (such as distance to foe or food, amount of energy, *etc.*), internal (fear, hunger, curiosity, satisfaction, *etc.*), and action (evade, socialize, explore, reproduce, *etc.*). At each time step, the activation level of a sensory concept is computed by performing a fuzzification of the information that the individual perceives in the environment (changing its real scalar value into a fuzzy value, *i.e.*, transforming the input value by a potentially nonlinear function). Subsequently, for an internal or action concept C, the activation level is computed from the weighted sum of the current activation level of all input nodes by applying a defuzzification function (another nonlinear function transforming the fuzzy input value into the final 'real' value).

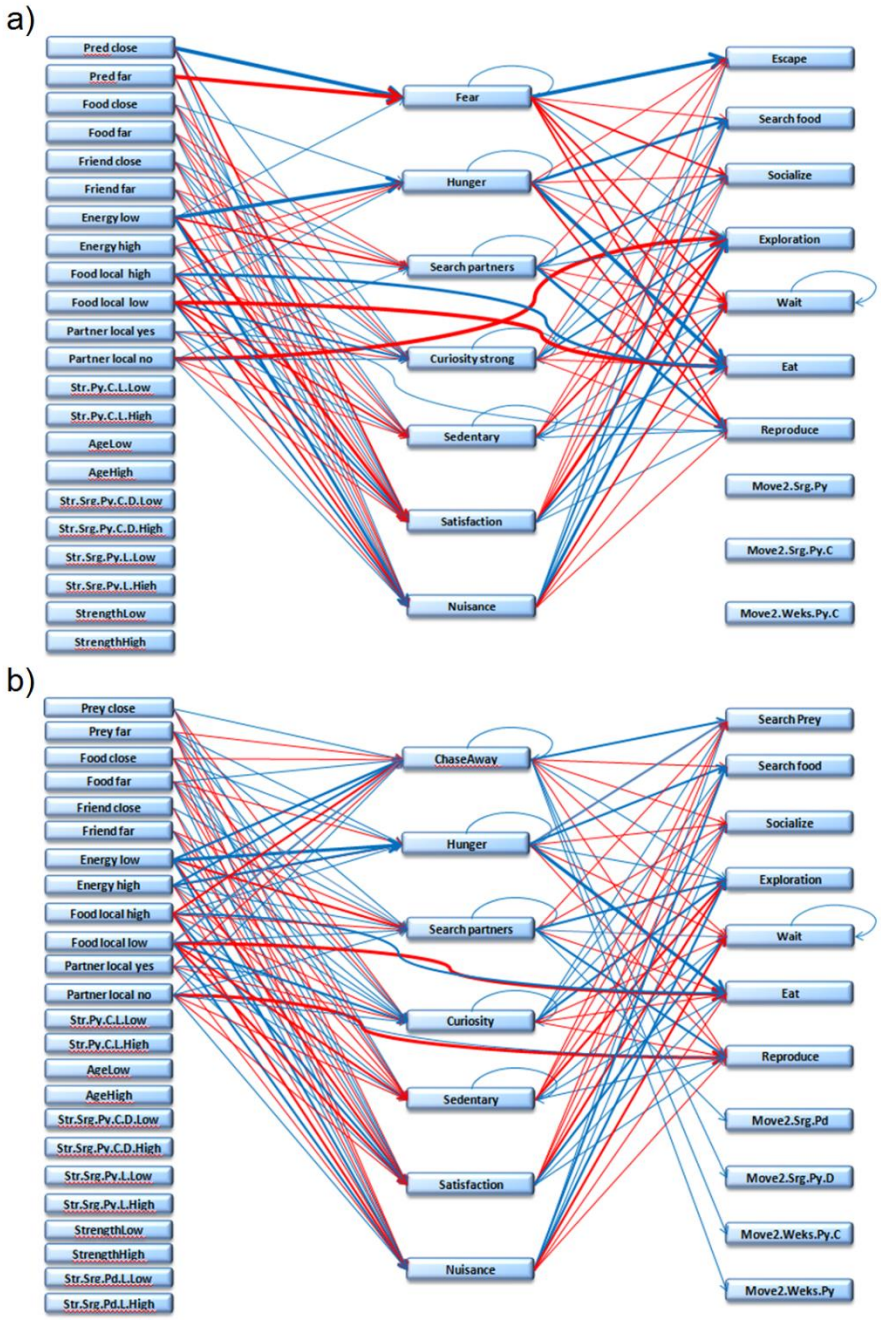


Fig. 3.1. An example FCM of a predator (a) and prey (b). Red edges between nodes indicate negative association (inhibition) of a concept (where the edge begins) with another (where the edge points to), and blue edges indicate positive association (excitation). The thickness of the edges represents the magnitude of the gene. The leftmost column of nodes is sensory concepts, the middle is internal concepts, and the rightmost is action concepts. There are many unconnected nodes because we aim to observe evolution in action; over time, new edges may form and others may disappear.

We will illustrate the operation of the FCM with a simplified example prey FCM (Figure 3.2) consisting of only four nodes (*EnemyClose*, *EnemyFar*, *Fear*, and *Evade*). *EnemyClose* and *EnemyFar* are sensory concepts, whereas *Fear* is internal and *Evade* is an action. All sensory nodes appear in pairs, like *EnemyClose* and *EnemyFar*; the

activation level of one of these nodes is always equal to 1 - *a*, where *a* is the activation level of the other. The individual perceives its environment to get a raw value for the distance to the nearest predator; this raw value is fuzzified to compute values between 0 and 1 for the activation levels of *EnemyClose* and *EnemyFar* by nonlinearly transforming it. To compute the activation level of *Fear*, a weighted sum of the activation levels of all nodes with incident edges to *Fear* is computed and the weights are the edge values from the behavioral genome. From our example, *Fear* has incident edges from *EnemyClose* and *EnemyFar*, thus we use edge weights from the behavioral genome for *EnemyClose*→*Fear* and *EnemyFar*→Fear to compute the weighted sum. The same computation is performed for the activation level of *Evade*. Finally, if *Evade* is the action selected by the individual (if, of all action concepts, it has the highest activation level), the speed of evasion is computed by defuzzifying the activation level of *Evade*. In the behavioral genome where no edge exists between two nodes (for instance, *EnemyClose*→*Evade*), the corresponding genes have values of zero. However, as individuals evolve, new edges can be added and pre-existing edges could be removed.

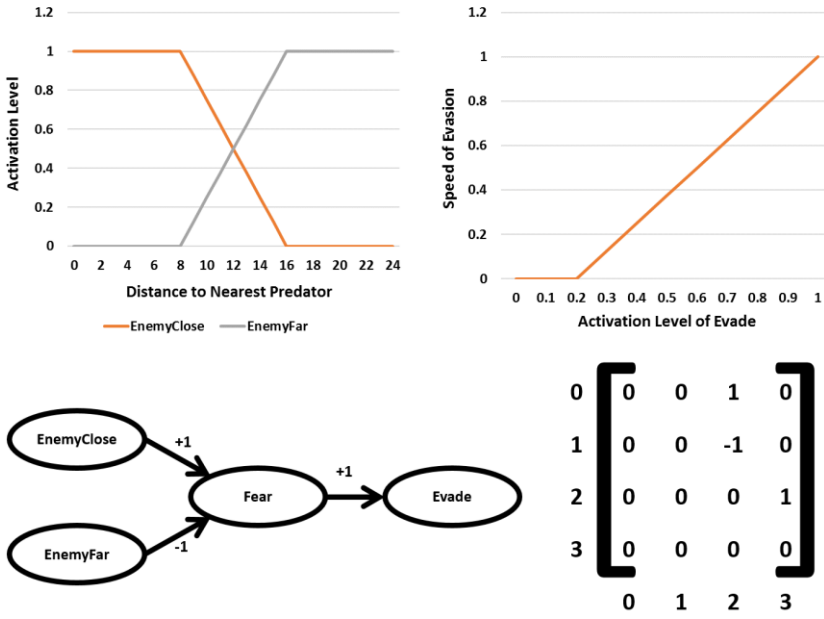


Fig. 3.2. A simplified example prey FCM for detection of predators (bottom left), with fuzzification (top left) and defuzzification (top right) functions, and its matrix (bottom right), which is the behavioral genome of the individual. *EnemyClose* and *EnemyFar* are sensory concepts, *Fear* is an internal concept, and *Evade* is an action concept. The edges of the FCM show influence of the activation level of a node on another. In the matrix, rows represent influencing concepts and columns represent those that are influenced. Row and column indices of 0 represent *EnemyClose*, 1 represent *EnemyFar*, 2 represent *Fear*, and 3 represent *Evade*.

Emergence

The FCM representation of the behavioral model allows for the appearance of positive and negative feedback loops. For instance, an individual may evolve a positive edge between the internal concept *Fear* and itself – this positive feedback loop can allow complex phenomena such as paranoia to emerge. Similarly, negative feedback loops can evolve that stabilize individual behavior. For instance, a negative association between *EnergyHigh* and *Hunger* with a positive association between *Hunger* and *Eat* means that after an individual replenishes its energy by performing the *Eat* action, it is less willing to eat again until its energy levels are lower. The fuzzification and defuzzification mechanisms allow for nonlinear transformations of the perception signal, which permits, for example, the representation of saturation of information. An individual's action is selected based on the action node with the highest activation level. Because of the way in which the behavioral genome determines the behavior of individuals and how the physical genome determines their physical capabilities, the evolution of behavioral and physical properties of individuals is emergent and it also influences other emergent properties of the system such as number of individuals, spatial compactness of individuals (a proxy of competition for resources), and number of species.

At the initiation of the simulation, prey and predators are scattered randomly all around the virtual world (see *Stochasticity* for a description of this process). Through the course of the simulation, the distribution of the individuals in the world changes based on many different factors such as behavior selection (prey escaping from predators, individuals socializing to form groups, and individuals moving to find food resources). In addition, emergent high-level migration phenomena and grouping patterns with spiral waves can be observed because of these complex interactions between the individuals and their environment. The distribution of individuals forming spiral waves is one property of prey-predator models (Golestani and Gras 2012; Figure 3.3).

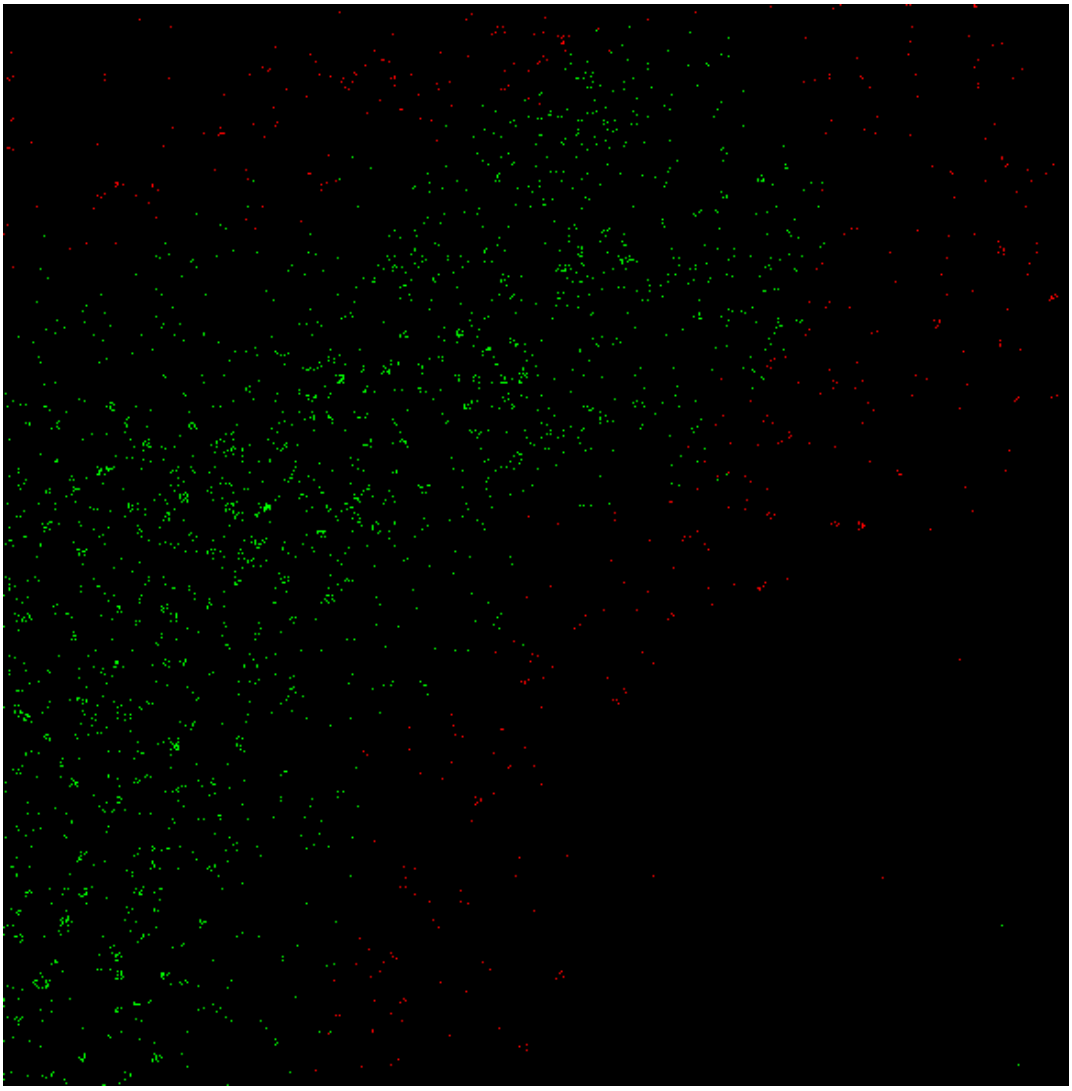


Fig. 3.3. A cropped image of an EcoSim run at time step 20000. Hungry predator individuals (red) chase fleeing prey individuals (green), one of the many contributory factors to the emergent high-level movement patterns we observe.

Adaptation

The behavioral genome's maximal length is fixed (663 genes for prey and 756 for predator), where each site corresponds to an edge between two concepts of the FCM. However, many edges have an initial value of zero; only 117 edges for prey and 131 edges for predators have non-zero values at initialization. Each gene of the behavioural genome follows the continuum-of-alleles model (Bürger 2000) and can take values between -12 and 12. These alleles represent the strength of the positive or negative influence of one concept on another, such as the strength of the association between level of hunger and willingness to eat. In addition to the behavioral genome, every individual has a physical genome that describes its physical characteristics, with each trait coded by one gene. Maximum energy (*MaxEnergy*), maximum age (*MaxAge*), vision (*Vision*),

maximum speed (*MaxSpeed*), minimum reproductive age (*RepAge*), and state of birth (*StateOfBirth*) are physical traits that both prey and predators possess. Prey have two more traits: defense (*Defense*), and cooperative defense (*CoopDefense*), so they can protect themselves from predators. The mechanisms involving the various physical traits are described further below and under *Submodels*.

Both genomes have two representations – a lightweight byte vector representation used for efficient storage in save files and for the computing of evolutionary distances and evolutionary operations, and a floating-point vector representation used for all other computing (activation levels, action selection, physical distances, energy dynamics, *etc.*). The mapping between these representations differs between the genomes. Both representations are fixed at birth for the individual's lifespan. For the behavioral genome, the byte value of zero maps to the floating-point value of zero. Any byte value less than 128 is reduced by 128 and then divided by 10 to get its associated floating-point value. Any byte value greater than or equal to 128 is reduced by 127 and then divided by 10 to get its associated floating-point value. Thus, byte values from zero to 127 take the range of [-12.7, 0] and byte values from 128 to 255 take the range of [0.1, 12.8]. For example, under this representation, a byte value of 76 yields a floating-point value of -5.2 ((76-128)/10) and a byte value of 200 yields 7.3 ((200-127)/10). For the physical genome, the floating-point representation of each gene has a *minimum* and a *step*. For byte value k, its floating-point equivalent is $minimum + (k \times step)$. For instance, *MaxEnergy* has a minimum of 100 and a step of 25. Thus, a byte value of 17 for *MaxEnergy* yields a floating-point value of 525.0.

The genomes of two parent individuals are transmitted to an offspring individual after recombination and potentially some mutations. EcoSim incorporates genetic recombination through crossover, and in the behavioral genome this includes epistasis (*e.g.*, multiple stimuli can influence a given drive) but no pleiotropy (each gene influences only one link between nodes). To model this form of linkage, alleles of the behavioral genome are transmitted by blocks. All incident edges for a given FCM node are transmitted together from a randomly selected parent with equal probability (there is no recombination among genes representing edges to a given node). Sex-linkage occurs for perception nodes, as the selected parent is of the same sex as the offspring. Sex-linkage of *MaxEnergy* occurs as it is a weighted sum of that of its parents. The parent with the same sex as the offspring has five times the influence on the offspring's *MaxEnergy* as the other parent (Eq. 3.3; *MaxEnergy* is abbreviated to *ME*; subscripts o, m, and f represent offspring, mother, and father, respectively). Sex-linkage occurs for *StateOfBirth* as well, as an offspring's *StateOfBirth* is equal to that of its parent of the same sex. All genes in the physical genome are potentially mutated after crossover with some probability ($p = 0.001$). A mutation on a gene in the physical genome is a

modification of its byte value (randomly drawn from a truncated normal distribution between -6 and +6). Mutations in the behavioral genome occur due to the formation of new edges (with a probability of 0.001), removal of existing edges (with a probability of 0.0005), and changes in the weights associated with existing edges (with a probability of 0.005). The effect of a given mutation is modification of the value randomly drawn from a truncated normal distribution between -0.6 and +0.6 on the floating-point value of a gene. The probability of mutation in the behavioral genome is doubled for old individuals ($Age > MaxAge - RepAge$). New genes may emerge from the initial pool of edges with a zero value. This emergence and disappearance of the genes in FCM is due to natural selection and genetic drift, which leads to the adaptability of individuals (Gras *et al.* 2015).

$$ME_o = \begin{cases} \frac{5 \times ME_m + ME_f}{6}, if\ offspring\ is\ female \\ \frac{5 \times ME_f + ME_m}{6}, if\ offspring\ is\ male. \end{cases} \quad (3.3)$$

Fitness

To measure the capacity of an individual to survive and produce offspring that can also survive, the fitness of a species is calculated as the average fitness of its individuals. The fitness of an individual is defined as the age of death of the individual plus the sum of the age of death of its direct offspring. Accordingly, the fitness value represents the individual's ability to survive and produce well-adapted offspring. There is no pre-defined explicit fitness-seeking process in the simulation; rather, fitness is a consequence of natural selection. Individuals who are better adapted to the environment sustain a higher level of energy, live longer, are able to have more offspring, and transfer their efficient genomes to them (Gras *et al.* 2009; Gras *et al.* 2015). The fitness value is only computed for analysis of the results of the simulation and is not used in process during the simulation.

Prediction

So far, there is no learning mechanism for individuals, and they cannot predict the consequences of their decisions. The only information available to an individual for decision-making comes from its perception at a given time step and the value of the activation level of the internal and action concepts at the previous time steps. The activation levels of the concepts of an individual are never reset during its lifetime. As the previous time step activation level of a concept is involved in the computation of its next activation level, this means that the previous states of an individual participate in the computation of its current state. Therefore, an individual has a basic memory of its own past that will influence its future behaviour. As the action undertaken by an individual at

a given time step depends on the current activation level of the action concepts, the behavior of the individual depends on a complex combination of the individual's perception, the current internal states, the past states it went through during its life, and its genome.

Sensing

Every individual in EcoSim can perceive its local environment inside of its range of vision. Some of these senses are common between prey and predator; both can perceive nearby friends and foes, how close food is, their energy level, the amount of food in their cell, how many potential reproductive partners are in their cell, and their *Age*. Additionally, new to EcoSim, all individuals can perceive their *Strength* and the maximum *Strength* of potential mates in their cell. Also new to EcoSim, prey individuals can sense the sum of *Strength* of prey in their cell and the sum of *Strength* of the cell within vision range that has the highest sum of prey *Strength*. Similarly, predator individuals can sense the sum of *Strength* × (1 + *Defense*) of prey in their cell, the distance to the cell in vision range with the highest sum of prey *Strength* × (1 + *Defense*), and the maximum *Strength* × (1 + *Defense*) in their cell. These new sensory concepts serve several purposes related to the notion of prey defending against predators, new to EcoSim. With these new sensory concepts, prey can use strength-related sensory information to join a cell with other strong prey to bolster cooperative defenses. Similarly, predators can use strength-related information to avoid conflict with stronger prey individuals or groups of strong prey. Alternatively, if the predator is very strong, it may use this information to gain a larger energy reward for killing stronger prey. Individuals can only reproduce with individuals of the same type in their current cell. Having the ability to sense strong individuals and move to them means that (with the right combination of edges) there is potential to improve the chance of reproducing with strong individuals. Thus, these concepts can also lead to some potentially interesting evolutionary phenomena, such as a strength-based evolutionary arms race between prey and predator populations.

Interaction

In EcoSim, there are direct and indirect interactions amongst individuals and between individuals and their environment. These interactions stem from actions that prey and predator individuals can perform. The only direct interaction that requires a coordinated decision by two individuals is *Reproduction*. *Reproduction* occurs between two prey or two predators. For *Reproduction* to be successful, the two parents need to be in the same cell, have sufficient *Energy*, choose the *Reproduction* action, and be genetically similar. The individuals cannot determine their genetic similarity with their potential partner; they

try to mate and if the partner is too dissimilar (the dissimilarity between the two genomes is greater than some percentage of the speciation threshold, by default 62.5%), the reproduction fails. See *Reproducing* under *Submodels* for more details of the *Reproduction* action.

The *Hunting* action of predators is a direct interaction that occurs between a predator and some number of prey existing in a cell. For *Hunting* to succeed, the predator must be able to move to the cell containing its target prey individual and it must have greater *Strength* than its target's *Energy*. Should the *Hunt* succeed, the prey target is killed and the predator receives some amount of *Energy*. The predator also receives an *Energy* penalty if the target prey tries to defend itself, or if other prey in the cell were defending the target. See *Hunting* under *Submodels* for more details of the *Hunting* action.

Lastly, there are several ways that individuals can indirectly interact with each other and their environment. An individual's perception of its local environment causes its actions and movement to be influenced by the distribution of other individuals and food resources. Moreover, individuals that share a cell compete for the limited resources that the cell contains (food and mates), and this yields density dependence. Competition generally comes in two main forms, which represent opposites along a gradient. Contest competition arises when a single individual claims all of its local resources, leaving other individuals with nothing (Brännström and Sumpter 2005). This allows individuals to potentially monopolize resources because strong individuals continue to claim resources while the weak starve and ultimately perish. Scramble competition, in contrast, occurs when individuals share resources equally and are thus equally penalized by local density increases (Brännström and Sumpter 2005). Competition in EcoSim, like in most ecosystems, is neither purely contest or scramble competition; elements of both forms of competition can be observed.

Stochasticity

To produce variability in the ecosystem simulation, several processes involve stochasticity. At initialization, the number of grass units is determined for each cell following a uniform random distribution (a value between 1 and *MaxGrass*). Similarly, at initialization, individuals are randomly distributed across the world in clusters. The simulation takes as input a clustering radius and a number of prey and predator individuals per cluster (see *Initialization and Input Data*). Let x and y be random coordinates for the center of a cluster, *ClusteringRadius* be the clustering radius, and k be the number of prey individuals in a cluster. Then, for each of the k prey individuals, x_n and y_n (the x and y coordinates for the position of the n^{th} individual in the cluster) are produced by taking x and y and subtracting from or adding to them a random value

between zero and *ClusteringRadius*. This process occurs until the entire initial set of prey individuals is placed in the world. The same process then occurs for the predators. The *Age* of an individual is also determined randomly at birth from a uniform distribution in [1, 24] for prey and [1, 35] for predators. Similarly, the initial energy of an individual is randomly generated in a uniform distribution, ranging from 40% to 100% of the initial maximum energy of the individual. *Age* and *Energy* are randomly generated in this manner to avoid apparition of synchronicity in action selection and death cycles early in runs that would cause instability, leading to extinction of prey or predators. The sex of an individual at initialization or at birth is randomly generated with equal probability to be male or female. Stochasticity is also included in several kinds of actions of the individuals (see *Submodels* for full descriptions of each action). For instance, if a hunting predator cannot find a prey within its vision range, the direction of its movement will be random. Furthermore, the direction of the exploration action is always random.

Mutation and crossover both involve stochasticity, as described under *Adaptation*. Furthermore, when individuals perceive their environment, they perform a radial sweep about their position along the four cardinal directions. The sweep begins at a distance of one and increments to the individual's vision range. The starting cardinal direction and the direction of the radial sweep are randomly generated to remove any biases in perception and movement. Lastly, stochasticity is incorporated into the grass diffusion model (see *Submodels* for elaboration). To understand the extent of stochasticity in EcoSim, Golestani and Gras (2010) examined whether chaotic behavior (one signal of non-randomness) exists in time series generated by the simulation. The authors concluded that the overall behavior of the simulation generates emergent patterns that are non-random and instead like those observed in complex biological systems (Kantz and Schreiber 1997).

Collectives

An EcoSim run persists while there is at least one prey individual. If all prey die, the run is complete due to extinction as the predators can only eat prey. EcoSim can be run with or without predators, though typically there are predators as it is designed to observe predator-prey interaction. A typical EcoSim run has 60000-1000000 prey and 2000-30000 predators at any time step, depending on the parameterization of the run.

In EcoSim, it is necessary to compute the genetic distance between any two genomes of the same type (prey or predator) in order to establish the notion of species. This distance calculation does not include sex-linked genes (see *Reproducing* under *Submodels*). To compute this distance, it is first initialized to zero. For every element of the behavioral genome in its byte vector form, the absolute difference between the pair of corresponding values from each genome is added to the distance. Subsequently, for every

gene of the physical genome, a weight is computed by taking the absolute difference of corresponding floating-point values and then dividing by the range of values for that gene. This weight is then multiplied by the difference between genes, multiplied by five, and added to the distance.

Species emerge from the evolving sets of prey and predators. Species membership is strictly used in data analysis – it is not used to govern any mechanics related to reproduction. There is a separate genetic similarity threshold used for reproduction which is much lower than the speciation threshold, and this allows hybridization (reproduction between members of different species) to occur (see *Reproducing* under *Submodels*). At initialization of EcoSim, there is one species per type. Species can become extinct if all their members die. EcoSim implements a species based on the genotypic cluster definition (Mallet 1995) in which a species is a set of individuals sharing a high level of genomic similarity. In addition, in EcoSim, each species is associated with the average of the genetic characteristics of its members, called the 'species center'. The speciation mechanism implemented in EcoSim is based on the gradual divergence of individual genomes. The speciation method begins by finding the individual *A* in a species *S* with the greatest genetic distance from the species center. Next, the individual *B* in *S* with the greatest distance to *A* is found. If this distance is greater than a pre-defined threshold for speciation, a 2-means clustering is performed (Aspinall and Gras 2010), otherwise *S* stays unchanged.

To initialize the 2-means clustering process, one center is assigned to a random individual, denoted I_r, and the other center is assigned to the individual who is the most genetically different from I_r. After eight cycles of the 2-means clustering algorithm, two new sister species are created to replace *S*. Each species for each type in EcoSim has a unique species identifier, starting at one and incrementing automatically when a new species is formed. Of the two sister species replacing *S*, one retains the species identifier of *S* and the other obtains the next available identifier.

Observation

EcoSim produces a large amount of data at each time step, recording many statistics like the number of individuals, the characteristics of each individual, and the status of each cell of the virtual world. Information regarding individual characteristic include spatial position, level of energy, choice of action, species identity, parents, FCM, *etc.* Information about the individuals, species, and virtual world for every 20 time steps are stored in a file, optionally using the HDF5 format (The HDF Group 2000) with an average size of 6 gigabytes. Also, there is a possibility of storing all of the values of every variable in the current state of the simulation in a separate file, creating the possibility of restoring the simulation from that state afterwards. The overall size of this

file, which is only stored every 20 time steps (by default, this frequency can be modified in the parameters file), is a few gigabytes depending on the numbers of individuals and species. All of the data is stored in a compact special format, to facilitate storage and future analysis. There are also several utility programs that can be used, for example, to analyze the simulation outputs, to calculate the species and individual fitness, to generate images of the world for each time step of the simulation, to generate the video of the world throughout a run or some portion of it, and to draw the FCM of the individuals.

Initialization and Input Data

A parameter file (with filename "*Parameters1.txt*") is defined for EcoSim, which is used to assign the values for each state variable at initialization of the simulation. Example parameters include the width and height of the world, initial numbers of individuals, thresholds of genetic distance for prey/predator speciation, speed of grass growth, probability of grass diffusion, initial maximum age, initial maximum energy, initial maximum speed, initial maximum vision range, initial values of FCM edges for prey/predators, and the characteristics of the fuzzification functions for sensory input. Any of these parameters can be changed for specific experiments and scenarios. Initialization involving stochasticity (such as the initial distribution of individuals in the world) is described under *Stochasticity*, above. Many of these initial parameters are only important in stabilization of the simulation in its early stages, before the emergent properties of the system are observable. These parameters have been tested extensively to ensure that EcoSim is stable in a wide variety of scenarios (if grass levels are low, if grass levels fluctuate regularly over time, if grass diffusion probability is reduced, if prey reproduce asexually rather than sexually, *etc.*). EcoSim is designed to be highly generalized. Typically, the emergent properties of at least two sets of runs initialized identically (or very similarly) with few mechanical differences are studied and compared, to observe the effect of these few mechanical differences on the evolution of the populations. Thus, the physiological scaling rates are informed by empirical biological studies (as noted above under *Individuals*), but the aim of the initial parameters of EcoSim is to produce a stable system and thus they are largely arbitrary. An example of a list of common user-specified parameters for the initial running of EcoSim are presented in Table 3.2.

Table 3.2. Values for user-specified parameters.

User-Specified Parameter	Used Value
Number of Prey	80000
Number of Predators	4000
Max Grass Quantity in each cell	4000
Prey Maximum Energy	2500

Predator Maximum Energy	3000
Prey Vision Range	8
Predator Vision Range	20

Output

EcoSim produces a wide variety of outputs. As EcoSim runs, it prints out updates of its progress. In standard output, it prints out the current time step, followed by the action it is currently processing (e.g. "Individuals updating", "Prey seeing world") and the time it takes to process the action. It also generates three main save files – WorldSaves, MinSaves, and MaxSaves. WorldSaves display the entire "world" vector, and are saved every time step. For each cell in the world, the WorldSave contains its coordinates along with its current level of *Grass*, *Meat*, and *Fertilizer*. MinSaves hold the current dynamic state of every individual alive in the simulation, and they are printed every time step. MinSaves and WorldSaves are used for post-processing and data analyses. MaxSaves are EcoSim's largest outputs – they save the entire state of an EcoSim run and are typically saved more rarely (e.g. every 20 time steps). MaxSaves serve as restore points so that a run can be paused and continued at the user's discretion. Further, MaxSave files can be duplicated to run identical EcoSim runs with different treatments, starting from any time step.

3.2.5 Submodels

Food Sources: Grass and Meat

There are dynamic processes for the resources in each cell, such as grass growth, grass diffusion, and variation in the amount of meat at each time step. At initialization, there is no meat in the world and the amount of grass energy units is randomly determined for each cell, as described under *Stochasticity*.

The grass growth rate in each cell is regulated by several factors: *SpeedGrowGrass* (200 by default), *ProbaGrowGrass* (0.035 by default), *MaxGrass* (4000 by default), and *Fertilizer*. The first, *SpeedGrowGrass*, is a parameter in the EcoSim parameter file that determines the speed of grass growth. For a cell not already containing grass, grass can diffuse from an adjacent cell with a probability of *ProbaGrowGrass* at a rate of *SpeedGrowGrass*, provided that one of the eight cells around the cell contains a non-zero amount of grass. *Fertilizer*, a feature new to EcoSim, is derived from the excretions of individuals. *AmountOfFertilizer*, the amount of fertilizer in a cell, is proportional to the sum of maximum energy (*MaxEnergy*) of the prey and predators residing in that cell, limited to a total of 20000. If *AmountOfFertilizer* is less

than *SpeedGrowGrass*, then the fertilizer does not have any effect. Otherwise, the rate of grass growth is equal to *AmountOfFertilizer* and limited to triple *SpeedGrowGrass*. For a cell already containing grass, the rate of grass growth is simply added to the amount of grass currently in the cell at a given time step. *AmountOfFertilizer* decreases at a rate of 10% per time step. The amount of grass in a cell is limited to *MaxGrass*.

Another new EcoSim feature is that *MaxGrass* can be set to fluctuate cyclically following a cos wave by setting the *FluctuatingResources* parameter in the parameter file. The period, minimum (as a ratio of *MaxGrass*), and amplitude (as a ratio of *MaxGrass*) of the wave can be set using the parameters *FluctuationCycle, FluctuationMinimumRatio*, and *FluctuationAmplitudeRatio*, respectively. Another new feature is that *MaxGrass* can be set such that it creates regularly positioned circular patterns throughout the world using the *CircularFoodGrowth* parameter. The diameter of the circles, the maximum grass level at the center of the circle (as a ratio of *MaxGrass*, though still limited by *MaxGrass*), and the minimum amount of grass in any cell (as a ratio of *MaxGrass*) are set using the *FoodCircleDiameter*, *FoodCircleMaxRatio*, and *FoodCircleMinimumRatio* parameters. *FoodCircleMaxRatio* is used to increase the rate at which *MaxGrass* increases closer toward the center of a circle, and *MaxGrass* increases following a cos wave from *FoodCircleMinimumRatio* to *FoodCircleMaxRatio* from the edge of a circle to the center.

The amount of meat in each cell is limited to *MaxMeat* (4000 by default) and increases every time step by the *Strength* of the prey killed in that cell during that time step. It also decreases at each time step by 1000, even if no meat has been eaten in this cell.

Actions

For each movement action *M*, the movement speed (*Speed*) is equal to *MaxSpeed* × *ActivationLevel*(*M*), thus the speed at which an individual moves during the action depends on its willingness to perform it. *Speed* is the straight-line distance that an individual can move in a single time step. Each action has its own corresponding submodel:

1. *Evading* (for prey only). An evading prey moves in the direction opposite to the barycenter of the five closest predators within its vision range, with respect to its position. If no predator is within the vision range of the prey, the direction is chosen randomly.
2. *Hunting* (for predators only). The predator selects the closest cell (including its current cell) that contains at least one prey and moves toward that cell. If it reaches the corresponding cell based on its *Speed*, the predator selects a prey target and tries to kill

it. When there are several prey in the destination cell, one of them is chosen randomly as the target. If the *Speed* of the predator is not enough to reach the cell, it moves at its *Speed* toward the cell and the hunt has failed. Similarly, the hunt has failed if there is no prey in the vicinity. When a predator's hunt succeeds, the *Strength* of the killed prey is added to the cell in meat energy units. Afterward, the predator consumes the meat to gain its required energy, *min*(*MaxEnergy* – *Energy*, *MeatUnits*), where *MeatUnits* is the number of meat energy units produced by the killed prey. The remaining units of meat energy are allocated to the cell and can be consumed by other predators using their *Eat* action. Prey have a defense capability as well as cooperative defense and use them in a battle against the predator (Arnold 2000).

Prey defense and cooperative defense is passive; prey defend automatically if they have a non-zero *Defense* value and are targeted by a predator, or if they have a non-zero *CoopDefense* value and share a cell with a target. Prey spend energy when trying to defend, and predators receive an energy penalty (P in Eq. 3.4, $AP.D$ and $AP.S$ are *Defense* and the *Strength* of the attacked prey; $CP_i.D$, $CP_i.CD$, and $CP_i.S$ are the *Defense*, *CoopDefense*, and *Strength* of the prey i in the same cell) when they attempt to attack a prey individual with non-zero *Defense* or a cell containing prey defending cooperatively. It is even possible for a predator to be killed by defending prey, particularly if the predator already has low *Energy*. Additionally, the prey that are involved in a cooperative defense also lose some amount of *Energy* based on the strength of the predator ($0.2 \times PredatorStrength \,/\, NumberOfDefenders$). The target prey loses *Energy* equal to 100% of the attacking predator's *Strength* if it is not cooperatively defended, otherwise it loses 80% of the attacking predator's *Strength*. If, after the attack, the prey's *Energy* is greater than zero, the prey survives and the hunt has failed.

$$P = AP.D \times AP.S + \sum_i (CP_i.D \times CP_i.CD \times CP_i.S) \qquad (3.4)$$

3. *Searching for food*. The direction toward the closest food (grass for prey, meat for predators) within the vision range is computed. If the individual's *Speed* is high enough to reach the food, the individual is placed in the cell containing this food. Otherwise, it moves at its *Speed* toward this food. If no food is within vision range, the individual moves in a random direction.
4. *Socializing*. The direction toward the closest possible mate within the vision range is computed. If the individual's *Speed* is high enough to reach this mate, the individual is placed in the cell containing this mate. Otherwise, the individual moves at its *Speed* toward this mate. If no mate is within vision range, the individual moves in a random direction.

5. *Exploring*. A direction is computed randomly. The individual moves at its *Speed* in this direction.

6. *Resting*. Nothing happens.

7. *Eating*. If the current amount of grass (meat) in the prey's (predator's) cell is greater than 0, the prey (predator) consumes the grass (meat) to gain its required energy, *min*(*MaxEnergy* − *CurrentEnergy*, *EnergyUnits*), where *EnergyUnits* is the number of grass (meat) energy units in the cell. *EnergyUnits* is decreased by the amount consumed by the individual.

8. *Reproducing*. Chromosomes in eukaryotic cells are usually present in pairs (diploid organisms). The chromosomes of each pair separate in meiosis, one going to each gamete. In many animal species, sex is determined by a special pair of chromosomes called sex chromosomes (allosomes), the X and Y. All other chromosomes are called autosomes. The sex chromosomes are an exception to the rule that all chromosomes of diploid organisms are presented in pairs of morphologically similar homologs. While females have two X chromosomes, the males have one X chromosome along with a morphologically unmatched chromosome, called the Y chromosome. All somatic cells in male and female organisms have a complete set of autosome and sex chromosomes. Every egg cell contains an X chromosome, while only half of sperm cells contain an X chromosome and the other half contain a Y chromosome. This difference is a chromosomal mechanism for determining sex at the time of fertilization. In other words, while autosome chromosomes are randomly obtained from both parents; the Y chromosome in male offspring is exclusively acquired from the father (Hartl and Jones 2004). Individuals in EcoSim, in contrast to the common case, are haploid. That is, their chromosomes are present as singletons that are generated from specialized evolutionary operations described below. To model more realistic individuals, we made it so that all perception genes, *MaxEnergy* genes, and *StateOfBirth* genes exist on allosomes (that is, they are sex-linked), while all other genes exist on autosomes. Thus, there is an evolving differentiation between male and female behavior.

   As per the section *Process Overview and Scheduling*, females intending to reproduce act first. This is because females initiate reproduction in EcoSim, to simulate female choice. Females can attempt to reproduce with any male in their cell, however, success is not guaranteed and individuals always act in order of decreasing *Strength*. There are several ways a reproduction attempt can fail in EcoSim. Reproduction fails if there are no males in the current cell. Otherwise, the female randomly selects a potential male partner. A reproduction attempt with a single male can fail if: the male has already reproduced (with a different, stronger female), the male has selected a different action (*e.g.*, *Eat* or *Evade*), the male is below reproduction age, the male has insufficient

energy to reproduce, or the genetic distance between the female and male is too great. The genetic distance threshold for reproduction failure is greater than the speciation threshold, therefore individuals from different species can reproduce to generate hybrid offspring. In this case, the hybrid offspring is assigned to the species that has the smaller genetic difference between its average genome and the genome of the offspring. The female can attempt to reproduce with each male in the current cell, but loses two *Energy* for each failed attempt. If reproduction succeeds, the process of generating a new offspring consists of the following steps. When a new offspring is created, it is given a genome that is a combination of the genomes of its parents using a specialized crossover operation along with some possible mutations (as explained under *Adaptation*). The sex of the offspring is randomly determined with equal probability of being male or female. Then, the initial *Energy* (*Energy*$_0$) of the offspring is computed (Eq. 3.5) based on the parents' *MaxEnergy* (abbreviated to *ME* in the equation) and *StateOfBirth* (abbreviated to *SOB* in the equation).

$$Energy_0 = \frac{ME_f \times SOB_f \times ME_m \times SOB_m}{100} \tag{3.5}$$

Finally, the *Energy* of the two parents is decreased. The energy penalty for the mother, *penalty*$_m$, is calculated based on Eq. 3.6, where the subscript *m* and *f* mean mother and father, respectively. The parameter *Energy* is the newborn individual's *Energy*. *FPP* is the first-time pregnancy penalty for the mother, which is five percent of its energy and zero for the subsequent pregnancies. The energy penalty for the father is based on Eq. 3.7.

$$penalty_m = \frac{SOB_m \times \text{Energy} \times 1.05}{\text{SOB}_m + \text{SOB}_f} + FPP \tag{3.6}$$

$$penalty_f = \frac{SOB_f \times \text{Energy} \times 1.05}{\text{SOB}_f + \text{SOB}_m}. \tag{3.7}$$

9. *Move2StrongestPrey/Predator* (for prey/predators, respectively). The direction toward the strongest possible mate within the vision range is computed. If the *Speed* of the individual is high enough to reach the mate, the individual is placed in the cell containing this mate. Otherwise, the individual moves at its *Speed* toward this mate. If no mate is within the vision range of the individual, the direction is chosen randomly.
10. *Move2StrongestPreyCell* (for prey only). This action is similar to *Move2StrongestPrey/Predator,* except that the direction of movement is toward the cell with the highest cumulative *Strength* of prey individuals. This allows prey to benefit from cooperative defense against predators.

11. *Move2WeakestPreyCell* (for prey only). This action is similar to *Move2StrongestPreyCell,* but the direction of movement is toward the cell with the lowest cumulative *Strength* of prey individuals. This allows prey to have a higher chance of success in competition with other prey individuals in accessing food or mates.
12. *Move2StrongestPreyDistance* (for predators only). The predator moves toward the strongest prey individual to acquire more energy after possible hunting. If the *Speed* of the individual is high enough to reach the prey, the individual is placed in the cell containing this prey. If the *Speed* of the predator is not enough to reach the prey, it moves at its *Speed* toward this prey.
13. *Move2WeakestPrey* (for predators only). This action is similar to *Move2StrongestPreyDistance*, with the exception that the direction of movement is toward the weakest prey individual for easier hunting in the future.
14. *Move2WeakestPreyCell* (for predators only). This action is similar to *Move2WeakestPrey*, but the direction of movement is toward the cell with the lowest cumulative *Strength* of prey individuals to minimize the possible effect of cooperative defense by prey individuals.

3.3 Ecological and Evolutionary Properties

Time-series data are generated automatically by EcoSim per time step, as explained above. We computed ten runs of EcoSim in the default configuration (which we hereby refer to as *Default*) to 20000 time steps. Using external tools that already existed, we computed the mean of several important measures for these ten runs. We computed the number of prey and predator individuals, the number of prey and predator species, the mean distance evolved of all female individuals, and three physical attributes for all female individuals (*MaxEnergy*, *MaxSpeed*, and *Vision*). Distance evolved is computed by first computing the mean genome for all individuals at a given time step, and subsequently computing the genetic distance from this genome to the prey genome that the simulation was initialized with.

As expected, there was a dependency between number of prey and predators (Figure 3.4). At initialization of the simulation, the number of prey is greater than the number of predators (80000 and 4000, respectively). Therefore, we tend to observe an early spike in the number of prey, which subsequently sharply declines when the number of predator individuals rises. The increasing number of prey provides a good chance for the predators to have access to more food, resulting in an increase in their *Energy* and reproduction rate. The resulting increase in hunting by predators accompanied by local food resource shortages for prey decreases the number of prey, and consequently the

number of predators, ultimately leading to stabilization of the system. A similar phenomenon occurs at finer spatial scales; local population explosions and extinctions yield fine-scaled fluctuations in numbers of individuals over time, with a time lag between the fluctuations in number of prey and predators. This dependence of predator population on prey population is known as the Lotka-Volterra, model as outlined in Berryman (1992) and empirically corroborated by Piana *et al.* (2006), where they fitted the model to a time series dataset of 16 species of neotropical fish that were classified as either predators or prey. These time series mostly stabilize with these small fluctuations, resulting in 268871 prey (SD = 80804) and 10388 predators (SD = 2613.4). As Britten *et al.* (2014) observed, this stabilization can be jeopardized if there is a sudden decline in predator species in such a predator-prey system.

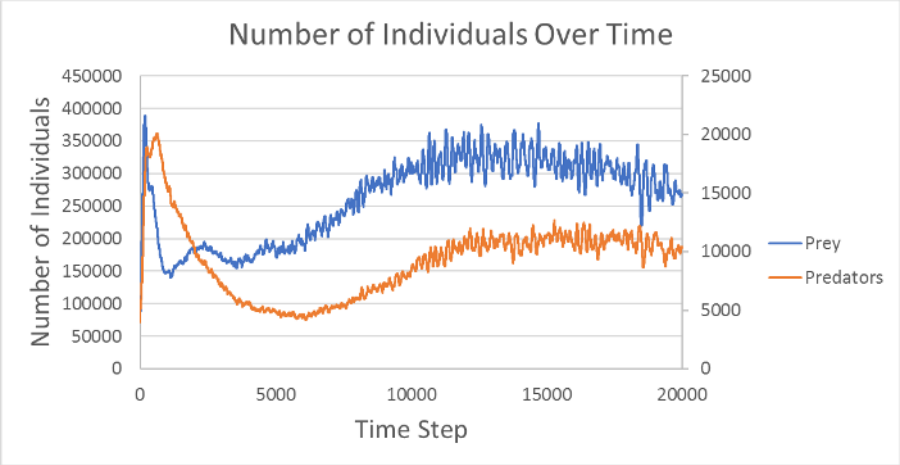


Fig. 3.4. The number of prey (left y-axis) and predators (right y-axis) in the world, over the course of the simulation.

The number of species more strongly correlated with the number of individuals for predators than for prey (Figure 3.5). Generally, an increase in the number of individuals allows for a corresponding increase in diversity within the gene pool, and this increased diversity tends to lead to increased speciation (Khater and Gras 2012). However, with the number of prey individuals so high, the gene flow is also very high, which results in overall genetic convergence. Spatial separation in individuals reduces gene flow. With fewer predator individuals in the world, there is greater spatial separation overall amongst predators, providing a greater opportunity for the subpopulations to genetically differentiate and ultimately yield new species. As Hoskin *et al.* (2005) argued, reduced gene flow in allopatry results in the gradual emergence of reproductive isolation and subsequently new species; this has been observed in EcoSim as well (Golestani, Gras, and Cristescu 2012).

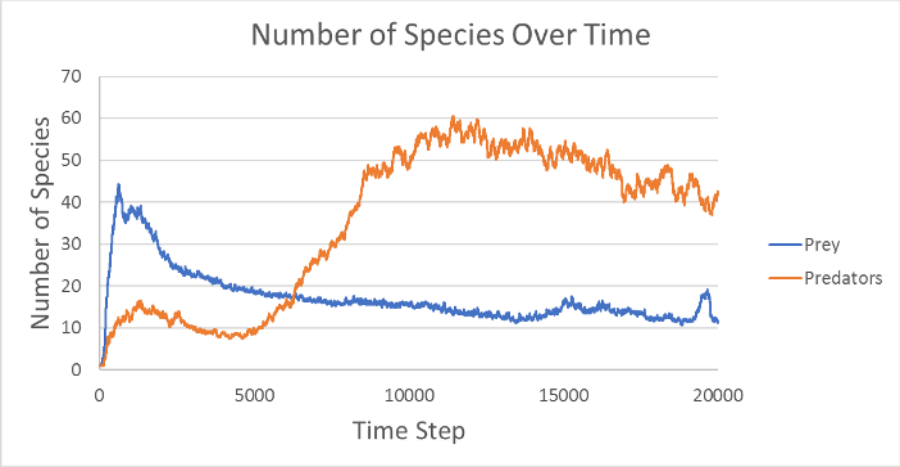


Fig. 3.5. The number of prey and predator species throughout the course of the simulation.

The prey and predator distance evolved were comparable by the end of the simulation (Figure 3.6). However, at the end of the simulation, the rate of predator evolution was greater than that of prey. In fact, nearly halfway through the simulation, the distance evolved for prey hit a plateau. This highlights an important distinction – that the prey (with such a high number of individuals) evolved rapidly but in a convergent manner, whereas the predators evolved more slowly but with high differentiation across all individuals. As Brodie and Brodie (1999), as well as Brodie *et al.* (2002) observe, predators that pursue prey with multiple defenses will tend to adapt evolutionarily, which may in part explain the higher rate of evolution of predators versus prey. Two main factors are responsible for the convergent evolution in prey: the aforementioned high gene flow and the fact that natural selection occurs in EcoSim since there is no pre-defined fitness function (Gras *et al.* 2015, Khater *et al.* 2014). The fitness landscape in EcoSim is dynamic overall; both the prey and predators evolve simultaneously and the world state is constantly changing. However, many aspects of the world remain constant, such as *MaxGrass*, the functions that govern energy expenditure of the individuals, and the rules that govern processes like reproduction. Thus, some genetic convergence should be expected – certain behavioral and physical genotypes will be desirable regardless of the world state at any given time step. The high genetic divergence accumulated early by the predators (apparent in the number of species over time) led to faster overall evolution later in the simulation. Another factor contributory to the fast evolution of predators later in the simulation is that there is more potential for divergence in the predator behavioral genome; the prey behavioral genome has 663 elements, whereas that of predators has 756. It is inevitable that predators will eventually evolve in a more convergent manner as well; this is observable in the subtle decrease in predator evolutionary rate over time.

Fig. 3.6. The distance evolved for prey and predators throughout the course of the simulation.

MaxEnergy evolved similarly for both prey and predators (Figure 3.7). In both cases, it monotonically increased from the initial values of 2500 for prey and 3000 for predators to an average of 3763 (SD = 505.7) and 4310 (SD = 372.3), respectively. As *Strength* is related to *MaxEnergy*, this could represent a type of evolutionary arms race because of the possibility of prey fighting back against predators when they attack. Alternatively, a higher maximum energy capacity may be strictly beneficial for the individuals, because it allows individuals to survive longer between *Eat* actions. Moller (2009) performed estimates of basal metabolic rate (BMR) of 76 bird species that were pursued by predators. The author reports that birds with longer flight initiation distances used to escape predators also had higher BMRs, from which he concludes that predation creates a selection pressure on species to develop higher BMRs (Moller, 2009). Thus, it is possible that the higher maximum energy capacity is necessary in individuals due to an increased BMR. Furthermore, the energy dynamics of each physical attribute is governed in part by the energy consumption functions for prey and predators. Thus, it is possible that with a more heavily penalized *MaxEnergy*, it might be less prone to such a runaway. *Vision* and *MaxSpeed* are related in that individuals must both perceive a resource (a mate, food, *etc.*) and be able to move to it in order to use it immediately. Otherwise, the individual will have to wait for at least one time step until it can use the resource it desires, which may be too late depending on the state of the individual and the environment around it. Thus, we should expect that *Vision* and *MaxSpeed* evolve in a related and intuitive manner. Predator *Vision* and *MaxSpeed* appeared to be heavily related in the way we expected (Figure 3.8). That is, both *Vision* and *MaxSpeed* evolved to slightly increase and then slightly decrease, nearly in unison, with *Vision* always greater than *MaxSpeed*. This is intuitive because it is particularly imperative for predators to perceive their resources; potential mates are far less abundant for predators, and their food resources are constantly changing positions in the world. This observation has been empirically corroborated in a

study of predatory bird species conducted by Garamszegi *et al.* (2002) in which it was found that predatory species evolved increased visual acuity along with larger brains to detect prey. On the other hand, it is less important for prey to perceive their resources, but it is important for prey to move quickly to evade predators. Potential mates and food resources are far more abundant for prey, and their food resources are static in the world (unless a cell's grass is fully consumed before the prey can reach it). Furthermore, over time, we observed that prey tended to perform the *Evade* action decreasingly while they increasingly performed *Explore* instead (Figure 3.9). The directionality of the *Explore* action is randomly generated, and with the high prey density it is possible that when they *Explore* they can randomly discover mates or food resources while they simultaneously evade predators. If all prey in a particular wave performed *Evade* when faced with a predator, many of the prey individuals would move in a similar direction, which could increase competition for resources. On the other hand, increasingly performing *Explore* may be evidence of the evolution of altruism; if a small percentage of prey purposely sacrifice themselves by moving towards the wave of predators (using *Explore* rather than *Evade*), it keeps the wave of predators away from the highest-density prey regions.

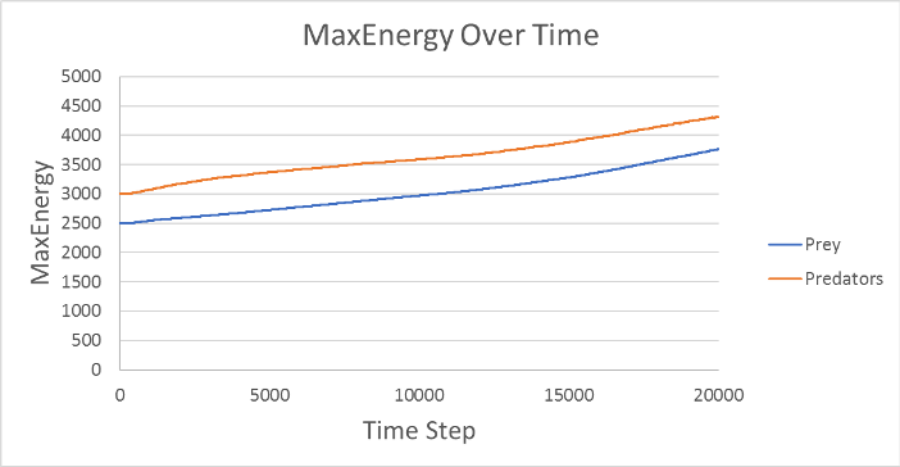


Fig. 3.7. The evolution of *MaxEnergy* for prey and predators throughout the course of the simulation.

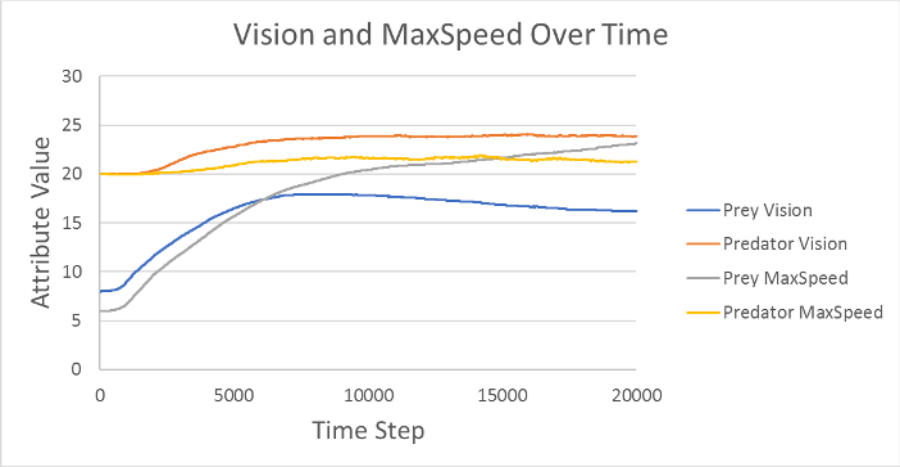


Fig. 3.8. The evolution of *Vision* and *MaxSpeed* for prey and predators throughout the course of the simulation.

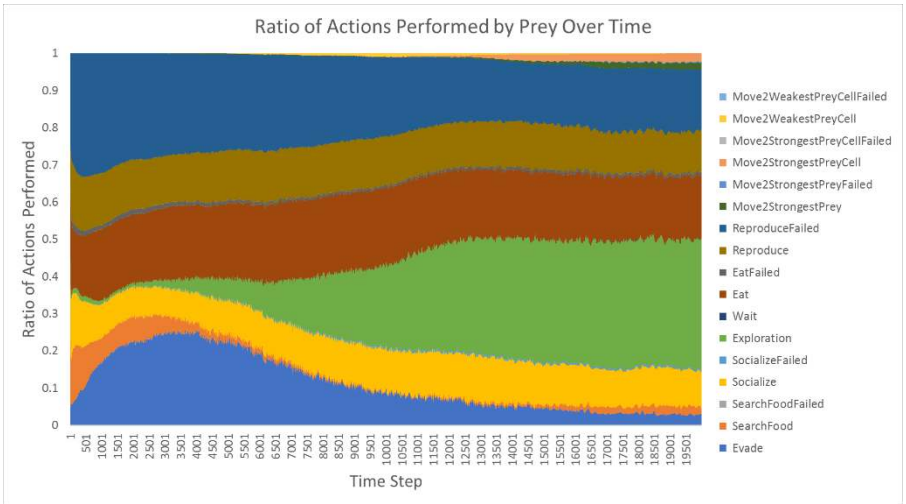


Fig. 3.9. Selection of actions by prey over time. Prey evolved to *Evade* less and *Explore* more, while simultaneously reducing their reproduction failure rate (*ReproduceFailed*). Evolution of an increase in *Move2StrongestPreyCell* and *Move2StrongestPrey* is also observed.

CHAPTER 4

Exploring the Effects of Genetic Diversity on Establishment Success in EcoSim

4.1 Introduction

Studies of biological invasions are certainly important from a practical standpoint; understanding invasions allows us to more effectively minimize the spread of invasive species and their impacts. However, invasions are also becoming increasingly recognized as interesting to study in light of theoretical evolutionary ecology (Sax *et al.* 2007; Lawson Handley *et al.* 2011; Bock *et al.* 2015). Introduced populations are subject to ecological conditions that are likely different from those in their native ranges, and so the introduced organisms must either be exapted (i.e., having evolved traits in the native range that yield fitness advantages in the introduced range, solely by chance; Hufbauer *et al.* 2012) to these circumstances or able to rapidly adapt to them in order to establish. Ecological interactions and their outcomes are undoubtedly influenced by the genetics of the populations involved; in biological invasions, these interactions are unique in terms of the species involved and the genetic and demographic dynamics of the introduced populations (Dlugosch *et al.* 2015).

Evolutionary-ecological studies of species introductions have provided researchers with two non-mutually-exclusive main types of genetic insights (Bock *et al.* 2015). The first is evolutionary rate information (e.g. Simmons and Thomas 2004; Carroll *et al.* 2005), potentially in conjunction with genetic bottlenecks and possibly evolutionary rescue (e.g. from multiple introductions or even multiple sources of introduction, e.g. Kolbe *et al.* 2004). The second type of insight is regarding the phenotypes, corresponding genotypes, or evolutionary processes that allow certain species or populations to establish, expand, and become invasive. There are numerous examples of these studies, including the aforementioned ones (Travis and Dytham 2002; Klopfstein, Currat, and Excoffier 2006; Travis *et al.* 2007; Zayed, Constantin, and Packer 2007; Burton and Travis 2008; Burton, Phillips, and Travis 2010; Colautti and Lau 2015; Peischl and Excoffier 2015; Chen et al. 2018; Lustenhouwer, Williams, and Levine 2019). Kanarek and Webb (2010), for example, showed that evolutionary rescue from Allee effects was possible, and that considering invasions only from ecological perspectives (i.e. without considering evolution) could lead to underestimation of the potential for invasion. Zhang *et al.* (2019) found evidence in the invasive alligator weed for the evolution of increased competitive ability (EICA) hypothesis.

Despite numerous insights from studying invasions, there exist many unanswered questions in the field of invasion biology. An important question that arises in the discussion of establishment success is the role of genetic diversity of introduced

populations. It is theorized that introduced populations should be subject to a demographic and genetic bottleneck (Roman and Darling 2007; Bock *et al.* 2015; Estoup *et al.* 2016; Briski *et al.* 2018); introduced populations can originate from extremely small nonrandom (e.g. spatially biased) samples of a source population and this should have short- and long-term consequences on their establishment success. In the short term, the probability of suitable genotypes for the novel region should increase with genetic diversity of introduced populations (Bock *et al.* 2015; Briski *et al.* 2018). A corollary of this is that the importance of genetic diversity for introduced populations should increase with the degree of adaptive challenge encountered (Estoup *et al.* 2016), for instance, degree of similarity or harshness of environments, or degree of similarity of competitors in native and introduced ranges (Hufbauer *et al.* 2012; Hufbauer *et al.* 2013; Fridley and Sax 2014; Rius and Darling 2014; Estoup *et al.* 2016). Affecting the longer term, low-diversity introduced populations have less "raw material" with which evolution can work to produce well-adapted genotypes in the novel range (Bock *et al.* 2015). This long-term effect need not take many generations to begin to manifest. For instance, Christie *et al.* (2016) found prominent signs of selection occurring in domesticated steelhead trout after only one generation. Similarly, in a lab experiment, Krause, Dinh, and Nielsen (2017) found increased tolerance to oil exposure in *Acartia tonsa* after only the second generation.

There exists empirical evidence for a positive effect of genetic diversity on establishment success (Forsman 2014 and references therein). Hufbauer *et al.* (2013) found in whiteflies that outbred introduced populations established more successfully than inbred populations in a novel, "harsh" environment. In a reciprocal transplant experiment between populations accustomed to two different environments, red flour beetles exhibited increased establishment success with increasing genetic diversity, and this effect was more pronounced with small propagule sizes (Szűcs *et al.* 2017). In a meta-analysis of 18 studies by Forsman (2014), genetic diversity had an overall positive effect on establishment of both plant and animal populations, evidenced in multiple ways (e.g. number of individuals produced after some time, proportion of colonies successfully established, plant biomass produced over time, etc.). Some of this evidence was circumstantial; propagule pressure, for instance, was not controlled across all cited studies (Forsman 2014). Further, some of the experiments had difficulty or failed outright to demonstrate that the actual genetic diversity differed between tested groups (Forsman 2014). Other studies have proposed that high genetic diversity in founding populations or subsequent admixture via multiple introductions may have aided in their establishment (e.g. Kolbe *et al.* 2004; Præbel *et al.* 2013), but these claims are speculative as the hypothesis was not tested outright. Rius and Darling (2014) discussed several important potential consequences of genetic diversity including increased heterozygosity, which

could allow the masking of recessive deleterious mutations, increased hybrid vigour, and the arrival of genotypes non-existent in the parental population which could lead to novel phenotypes that enhance fitness or adaptation.

On the other hand, it is possible that introduced populations, even relatively small ones, may actually not exhibit a substantial loss of genetic diversity compared to their source populations. Roman and Darling (2007) reviewed the genetic diversity and establishment success of aquatic invaders, finding that only 37% of the cases they reviewed involved significant loss of genetic diversity. Further, in successful introductions involving significant diversity loss, 63% of the species could reproduce without sexual recombination whereas only 19% exhibited this capacity in established populations that did not take a significant diversity loss. This suggests that asexual populations may not be impacted by genetic diversity loss. Similarly, Wares, Hughes, and Grosberg (2005) found that introduced populations from a variety of taxa retained over 80% of their genetic diversity as measured by heterozygosity and allelic richness. However, here too, evidence was circumstantial because both reviews included studies for which multiple introductions were either possible or even confirmed.

Some other studies provide evidence that perhaps the consequences of genetic diversity reduction are not always dire, and in fact sometimes beneficial. For instance, Suarez, Holway, and Tsutsui (2008) showed that a genetic bottleneck aided in the invasion of Argentine ants in California. Similarly, Mergeay, Verschuren, and Meester (2006) found that the entire African population of the asexual and invasive American water flea was sourced by a single clone. Zayed, Constantin, and Packer (2007) theorized that the source of an invasive bee population in North America was a single female. Szűcs *et al.* (2014) found no relationship between establishment success and genetic diversity of introduced populations of the red flour beetle; however, they did find that genetic diversity positively affected growth rate of established populations. Briski *et al.* (2018) noted that selection during the transportation phase of an introduction may yield extremely fit but low-diversity introduced populations if the novel environment is similar to that in the transport vector. Rius and Darling (2014) also discussed potential negative consequences of genetic diversity: outbreeding depression can occur from loss of beneficial parental genotypes or from unfit intermediate genotypes, genetic incompatibilities can exist between extremely different genotypes, and dilution of exapted genotypes if native and novel ranges are similar.

One thing that is clear from the above studies is that it is extremely difficult to study the role of genetic diversity on establishment success of introduced populations. There also exist other sources of difficulty (e.g. lack of control over diversity, potential confounding effects of propagule pressure or multiple introductions). One issue is that genetic diversity does not necessarily imply functional diversity. For instance, genetic

diversity is often approximated by analysis of specific neutral genetic markers which may not have any bearing on the success of invasions (Roman and Darling 2007; Wellband *et al.* 2018). Many invasion studies utilize comparisons of closely related but different species, or comparisons of the same species at different locations, or analysis only of successful invasions. This makes it unclear as to whether there is a causal relationship between diversity and success, whether the observed genetic diversity is a byproduct of the success (e.g. via expansion load), or if there is a relationship between genetic diversity and establishment success whatsoever.

One way to circumvent the above difficulties is to test the relationship between genetic diversity and establishment success in an individual-based model (IBM). Many practical studies in invasion biology have been conducted using IBMs. IBMs are a popular choice in predicting spread (Goslee, Peters, and Beck 2006; Phan, Huynh, and Drogoul 2010; Samson *et al.* 2017), preliminarily exploring management regimes (Bonesi, Rushton, and Macdonald 2007; Keith and Spring 2013), and predicting interactions between the introduced species and resident species (e.g. Bonesi, Rushton, and Macdonald 2007; Nguyen *et al.* 2011; Xiao *et al.* 2016). Theoretical studies can also be conducted using IBMs. For instance, studies on expansion load and gene surfing (Klopfstein, Currat, and Excoffier 2006; Travis *et al.* 2007; Burton and Travis 2008; Peischl and Excoffier 2015) were all conducted using IBMs. IBMs have also been used to investigate the evolution of dispersal (Travis and Dytham 2002; Travis *et al.* 2009; Fronhofer, Poethke, and Dieckmann 2015; Henriques-Silva *et al.* 2015), the relative roles of learning and evolution in exploring novel environments (Sutter and Kawecki 2009), the role of sex structure of introduced populations in establishment (Shaw, Kokko, and Neubert 2018), and the ability of populations to persist in changing environments (Santini *et al.* 2016).

In this study, we used EcoSim (see Chapter 3), a predator-prey ecosystem IBM in which its individuals can evolve, to study the effects of genetic diversity on introduced populations. We tested two main hypotheses. Hypothesis I was that, with all other factors held constant, increasing genetic diversity of introduced inocula increases their establishment success. Hypothesis II, a corollary of hypothesis I, was that genetic diversity is more impactful on establishment success when populations are introduced to an environment different from that in which they evolved. In this chapter we also introduce two novel variants of EcoSim – EcoSim Niches and EcoSim Invasions – which were used in this study but can be used in other future studies. EcoSim generates many types of data and large amounts of it – we can directly observe essentially any variable, for the introduced or the natives, at the scale of individuals, species, regionally, or globally. This allowed us to fully account for successful and failed introductions. Further, using EcoSim allowed us to directly circumvent many of the classical problems that

affect similar studies. In addition to testing the two main hypotheses, we used statistical and machine learning methods to determine if there are differences in what drives establishment success given the varying levels of genetic diversity.

4.2 Methods

To simulate biological invasions in EcoSim, we developed a new variant of EcoSim (EcoSim Invasions) which differed from the standard variant only in that it allowed users to transfer prey and predator individuals between two EcoSim runs. With EcoSim Invasions, invasion parameters were designed to be entirely customizable – invasions could involve any number of individuals, subject to any selection process (e.g. transfer a randomly selected subset of individuals with a specific fitness value) or modifications (e.g. set energy level of all introduced individuals to 1000), occurring at any regular or irregular frequency (e.g. every 20 time steps, starting at 10000 time steps, for 2000 time steps). To test hypothesis II, we developed another EcoSim variant – EcoSim Niches – which we also modified such that individuals could be transferred between runs. In 4.2.1, we provide a brief overview of the current study. EcoSim Invasions is detailed in 4.2.2, including all the modifications to standard EcoSim that were made to produce EcoSim Invasions. In 4.2.3, we describe EcoSim Niches, which produced a vastly different environment from that of standard EcoSim. In 4.2.4, we provide the remaining details of the current study; before we get into these details it is necessary to introduce the overall study design, EcoSim Invasions, and EcoSim Niches. Lastly, in 4.2.5, we discuss the data analysis we conducted.

4.2.1 Brief Overview

Our study consisted of a reciprocal transplant experiment with two different environments simulated in EcoSim (standard EcoSim and EcoSim Niches, detailed in Chapter 3 and 4.2.3 respectively), across which prey populations were transplanted with five levels of genetic diversity, with 10 replicates each, occurring over a fixed time interval. With EcoSim Invasions, we held constant propagule size (number of prey individuals introduced per event) and propagule number (number of introductions). In this study, we simulated invasions involving a fixed number of introductions (50), each of same propagule size (100 prey individuals), occurring with the same frequency (every 100 time steps), and over the same time period across all simulations (over 5000 time steps). In total, 10000 introduction events took place. Prey were introduced into four run types, covering each combination of source and receiver of introduced populations. The four types were standard receiving standard prey (hereby denoted S→S), standard receiving Niches prey (hereby denoted N→S), Niches receiving standard prey (hereby denoted S→N), and Niches receiving Niches prey (hereby denoted N→N). For a given

run type, five sets of ten runs were produced such that each set of runs always received introduced prey populations of a fixed genetic diversity level (Figure 4.1). There were 20 EcoSim Invasions runs generating samples – ten standard and ten Niches. There were four types of runs to which prey were introduced, each with five sets of runs of different genetic diversity levels, with ten runs per set, amounting to 200 more EcoSim Invasions runs. Thus, there were 220 EcoSim runs in total.

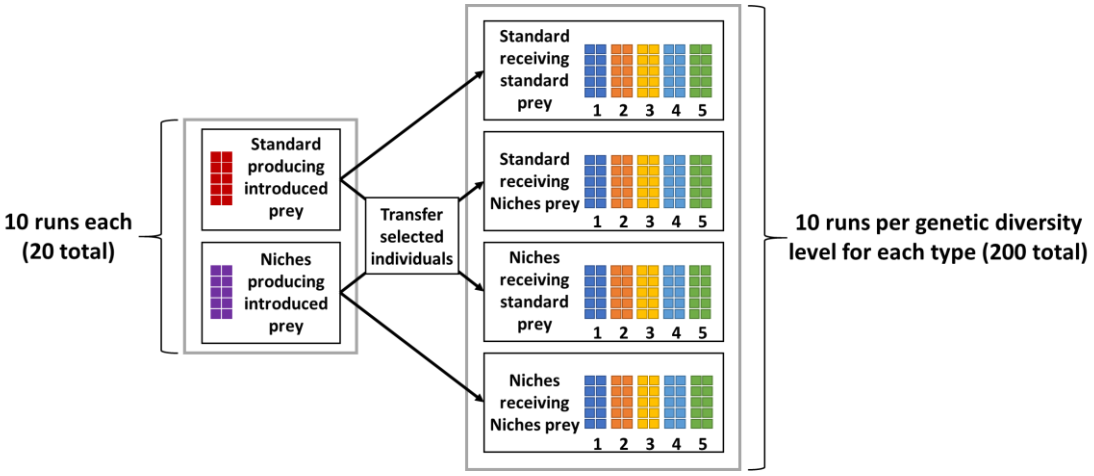


Fig. 4.1. Depiction of the experimental design. Ten standard EcoSim runs (left, "Standard", red boxes) and ten EcoSim Niches runs (left, "Niches", purple boxes) were used to generate prey samples to be introduced into standard EcoSim and EcoSim Niches runs (right). Arrows represent the transfer of prey individuals. There were four types of runs receiving prey populations, with five levels of genetic diversity of introduced prey populations and 10 runs for each combination of type and genetic diversity level (colored boxes on right, numbers indicate level of genetic diversity). Thus, there were 220 total EcoSim runs.

There were several ways we could test Hypothesis I – that establishment success of introduced populations increases with their genetic diversity – in EcoSim. Establishment success is classically difficult to quantify, and presence/absence is often used because of its simplicity. We quantified both the short-term and long-term establishment success of the introduced populations. To quantify short-term establishment success, we recorded presence/absence of introduced individuals 60 time steps after each fresh inoculation (an inoculation into a simulation in which there were no living introduced individuals). Also, we recorded presence/absence of introduced individuals 100 time steps after introductions ceased (time step 20200) to determine long-term establishment of the introduced populations. With hypothesis I, we expected that both short-term and long-term establishment success would increase with genetic diversity, with diminishing returns, for populations introduced into the environment in which they evolved (i.e. populations from standard EcoSim inoculated into another standard EcoSim run, populations from EcoSim Niches inoculated into another EcoSim Niches run). Hypothesis II – that genetic diversity should be more impactful when the environment of the introduced range is dissimilar to that of the native range – was tested in the same manner as hypothesis I. However, with hypothesis II, we instead tested the

establishment success of populations in environments different from that which they evolved in (i.e. populations from standard EcoSim in EcoSim Niches and vice versa) and compared this with what was observed when we tested hypothesis I. We expected that the relationship between genetic diversity and establishment success would be of similar shape in all cases, but that the relationship would be more pronounced for populations introduced into environments different from that in which they evolved. We also analyzed the impact of other factors on establishment success (see Section 4.2.5).

4.2.2 EcoSim Invasions

Here, we describe only the features of EcoSim relevant to this study. EcoSim is described in ODD format in Chapter 3. In EcoSim Invasions, the invasion process (Figure 4.2) occurs as follows. Let NR (native range) and IR (introduced range) be two types of distinct EcoSim runs. NR runs are those that produce individuals that will be introduced to IR runs. At a given time step *t*, a select set of individuals from an NR run are transferred to an IR run. All state variables of the individuals are transferred such that they can be exactly reconstructed in the IR run.

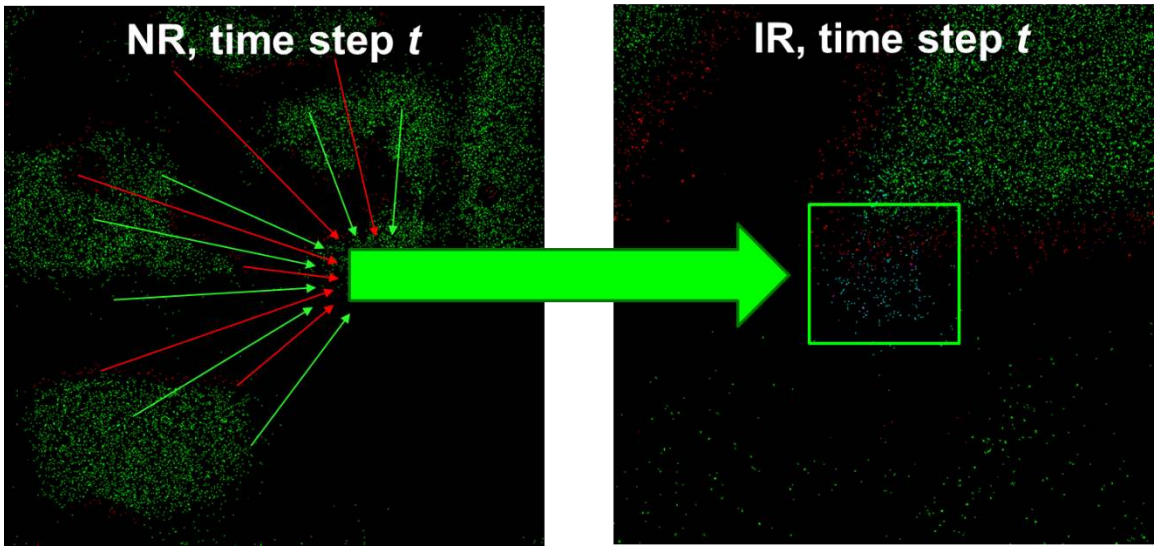


Fig. 4.2. Depiction of the invasion process in EcoSim Invasions. The left and right panels of the image are snapshots of two different EcoSim runs called NR and IR at the same time step (*t*). Green and red dots represent prey and predators native to each respective EcoSim run. Individuals from NR are subsampled at time step *t* – these individuals are introduced to IR at time step *t*. In the right panel, blue and purple dots represent the prey and predators, respectively, that were introduced to IR from NR.

To produce this variant of EcoSim, several important modifications were made which we outline below. We added parameters *isDonator* and *isAcceptor* to determine whether a run would be donating or receiving introduced individuals. Two more parameters, *numberPreyInvaders* and *numberPredInvaders*, controlled the number of prey and predator individuals transferred between runs. Another parameter called *invasionFrequency* was added to determine how often individuals should be transferred

between runs. Finally, a parameter *clonesOnly* was produced, allowing users to create inocula such that a single randomly-selected prey and a single randomly-selected predator are duplicated *numberPreyInvaders* and *numberPredInvaders* times, respectively (when *clonesOnly* = 1). Otherwise, *numberPreyInvaders* randomly-selected prey individuals and *numberPredInvaders* randomly-selected predator individuals comprise the inoculum. Further, custom inocula can be produced from any EcoSim run (standard, Invasions, or other) by writing external scripts that traverse MinSave files and extract prey and predator individuals, subject to any desired selection criteria. For instance, individuals can be sampled from specific physical regions or species, or with a specific fitness or energy level. Further, the introduced individuals can be subject to any modifications; for instance, their energy levels could be all maximized or set to an identical level, or physical properties such as vision could be reduced or enhanced.

We added an *isInvasive* flag to individuals, so that native and introduced individuals could be separately tracked. Importantly, outside of reproduction, the *isInvasive* flag was implemented such that it had no bearing on perception or decision-making of prey or predator individuals in EcoSim Invasions. To elaborate, any prey (native or non-native to an IR run) will perceive all other prey (and similarly, all predators) the same way. This has some important consequences (Table 4.1). For instance, a non-native prey may socialize or try to reproduce with a native prey; a native prey will be equally scared of a native or non-native predator; a non-native predator will be equally willing to hunt a native or non-native prey, and will be equally successful in doing so. Users can modify the code such that the relationships between native and non-native individuals are different from above (perhaps, for instance, such that native and non-native individuals can learn to or already know how to distinguish each other). With respect to reproduction, native and non-native individuals could not reproduce with each other. That is, native individuals can only reproduce with native individuals and non-native individuals can only reproduce with non-native individuals. This was implemented as a reproduction failure condition, thus the individual attempting reproduction incurred an energy penalty but could still attempt reproduction with other local individuals (see *Actions* under *Submodels* in Section 3.2.5).

Table 4.1: Key Simplifications regarding EcoSim Invasions, and their potential consequences.

Regarding…	Simplification	Potential Consequence
Perception	Natives and non-natives register identically to each other	Reduced fitness at expansion front
Social Actions	Natives and non-natives can socialize with each other	Reduced effectiveness of social actions at expansion front
Reproduction	Natives and non-natives may attempt but never succeed in reproduction with each other	Reduced reproduction success along expansion front
Hunting, Escaping,	Any predator can hunt any prey Any prey can escape any pred	Introduced prey/preds always have potential food source

and Eating	Any prey can *eat* any grass	Possible prey enemy release

In this particular study, as we wanted to test the effect of genetic diversity of introduced populations on their establishment, we wanted to remove any biases regarding genetic distance between mating individuals. EcoSim, by default, has a genetic distance threshold that determines whether two individuals can successfully reproduce. For EcoSim Invasions runs, we disabled this threshold for introduced individuals so that high genetic distance between mating individuals, which would be more common in high-diversity introduced populations, would not be disproportionately penalized. On a related note, all introduced individuals were assumed to be of the same species, regardless of species membership in their native range. Species designations are emergent and arbitrary in EcoSim, and purely used for analysis; they have no bearing on the actions of the individuals. This simplified our sampling process, provided us with greater control over the genetic diversity of inocula, and increased the range of genetic diversity that we could explore in this study. Study of the effects of assortative and disassortative mating in introduced populations, by differentially modifying mating success based on genetic distance, could be done in future work.

4.2.3 EcoSim Niches

To test hypothesis II, that genetic diversity is more impactful on establishment success when the introduced range has an environment different from that of the native range, we required an EcoSim variant that produced an environment different from that of standard EcoSim. Thus, we developed EcoSim Niches, which was a variant of EcoSim in which the 2D world was divided into quadrants that each had unique circular *MaxGrass* patterns. In each circular pattern, *MaxGrass* was maximum in the center of the circle, decreasing nonlinearly to some minimum level as distance from the center of the circle increased. The circular patterns were unique in terms of the maximum *MaxGrass* levels they each possessed, as well as the rate at which *MaxGrass* decreased as individuals travelled away from their centers (Figure 4.3). Because the minimum *MaxGrass* level was lowest at the outer edges of the quadrants, they effectively produced barriers that physically and reproductively isolated the populations within them. As each of the quadrants were designed to be drastically different from each other, we anticipated that these regions formed a wide variety of ecological niches to be adapted to by the residents of the simulation – hence the name, EcoSim Niches.

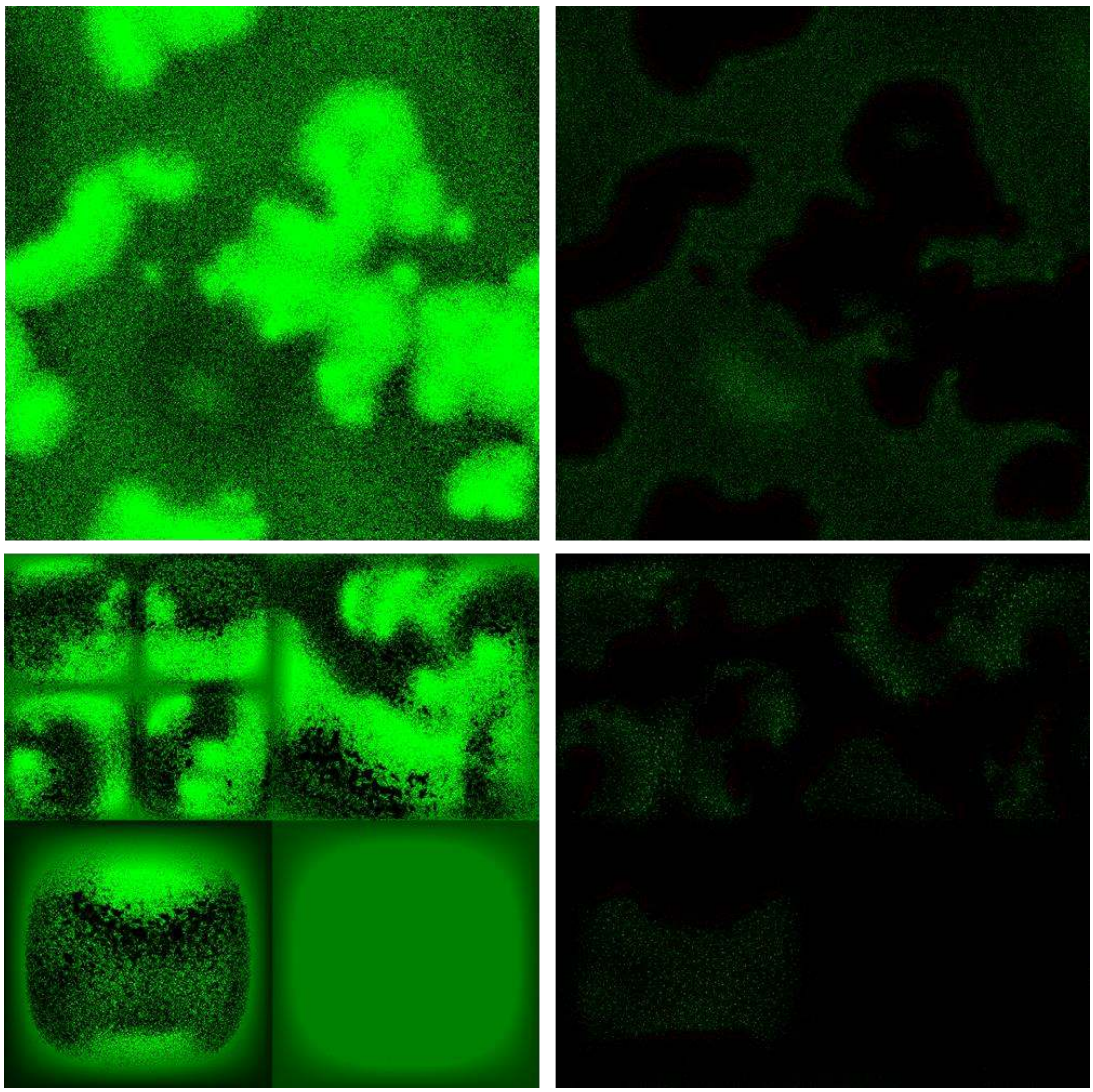


Fig. 4.3. Comparison of *Grass* levels (left) and individual distributions (right) between a standard (top) and a Niches (bottom) EcoSim run, both at time step 15000. On the left, intensity of green increase with *Grass* level of the cell. On the right, green dots represent prey and red dots represent predators. Due to the *Grass* diffusion model, in a given cell, *Grass* levels can reach a maximum level (*MaxGrass*) that is spatially uniform throughout the world in standard runs. In EcoSim Niches runs, however, *MaxGrass* levels in the world are non-uniform and due to their spatial distribution, interesting ecological niches form. Note that the bottom right quadrant in the Niches run at this time step was mostly not yet exploited.

Few changes were required to create the EcoSim Niches variant from the standard variant of EcoSim. Standard EcoSim and EcoSim Niches were parameterized entirely the same. The only difference between standard EcoSim and EcoSim Niches was that *MaxGrass* had an additional multiplier that was computed based on the position of a cell. Where standard EcoSim consulted the *MaxGrass* variable to enforce a uniform *Grass* limit in each cell, EcoSim Niches consulted a *MaxGrassArray* variable which was a 2D array of size equal to that of the world (i.e. 1000x1000 by default). When a run was started or continued, this *MaxGrassArray* was populated using the following function:

(4.1)

$$MaxGrassArray(x,y) = \min(\max(\left(\sin\left(pi \times \frac{x}{D}\right) \times \sin\left(pi \times \frac{y}{D}\right)\right) \times MaxRatio + MinRatio, \\ -\left(\sin\left(pi \times \frac{x}{D}\right) \times \sin\left(pi \times \frac{y}{D}\right)\right) \times MaxRatio + MinRatio), 1.0),$$

where x and y are the x and y coordinates of a cell, D is the size of the quadrants (e.g. D=500 creates four quadrants in a 1000x1000 world, D=250 creates 16 quadrants, etc.), *MaxRatio* controls the rate at which *MaxGrass* increases towards the center of a circle, *MinRatio* controls the minimum value of *MaxGrass*, which occurs at the edges of quadrants. A cross-sectional view, across the center of a quadrant with D=250, provides a demonstration of the effect of the parameters *MaxRatio* and *MinRatio* (Figure 4.4).

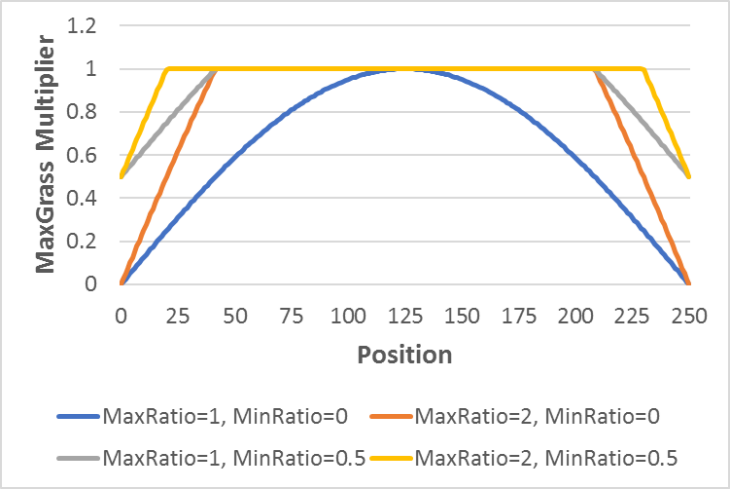


Fig. 4.4. Demonstration of the function used to initialize *MaxGrassArray* in EcoSim Niches. In this demonstration, D=250 and a cross-section across the center of a quadrant is depicted (i.e., forcing the multiplier produced by the sin function to 1 across the entire cross-section). The function limits *MaxGrassArray* at any index to 1.0, occurring at the center of a quadrant. *MinRatio* controls the minimum multiplier held in *MaxGrassArray*, which occurs at the edges of quadrants. *MaxRatio* increases the rate at which *MaxGrass* increases as individuals move towards the center of the quadrants.

To produce subquadrants in EcoSim Niches, as observed in Figure 4.3, conditional logic based on cell position was used to parameterize Eq. 4.1 in different ways depending on location. The bottom-right quadrant for instance, as in Figure 4.3, was purposely designed with a low *MaxGrassRatio*. Interestingly, prey individuals are initially unable to exploit this resource but can evolve the capacity to exploit it. Further, with the parameterization of each quadrant, prey individuals tend not to cross the harsher boundaries until they evolve sufficiently high *MaxSpeed* and *Vision*, or the capacity to exploit the relatively low-resource regions between quadrants. This produces physical and reproductive isolation.

4.2.4 Details of the Current Study

In this section, we will first elaborate further on the study design. We will then describe the process we used for selection of prey to be introduced across simulations, given that we needed to control genetic diversity of the introduced prey populations while simultaneously maintaining fitness distributions as much as possible.

Study Design

To be consistent with the abbreviations in 4.2.2, let NR*j* be a source run and IR*k-l* be a destination run, with respect to transferred prey samples. NR*j* produced five samples of 100 prey individuals (i.e. holding propagule size constant) every 100 time steps over 5000 time steps (from time step 15100 to 20100; i.e. holding propagule number constant). Each of the five samples at time step *t* targeted a specific level of genetic diversity, ranging from zero (i.e., all prey individuals in the sample are clones) to a maximum determined empirically for samples of 100 individuals (details of sampling process and empirically determined maximum diversity follow under *Sampling Process*). In preliminary work, we tested several propagule sizes and determined that establishment success widely varied with 100 individuals – we did not want establishment to be too easy (e.g. 100% success rate) or too difficult (e.g. 0% success rate). Consider a sample from NR*j*, InvaderSave*j-k-t*, where *j* was the run number of origin, *k* was the level of genetic diversity (increasing from one to five), and *t* was the time step during which the sample was taken. When run IR*k-l* reached time step *t*, it loaded InvaderSave*j-k-t,* where *l*=11-*j* (Figure 4.5). The rationale for the relationship between NR and IR run numbers follows.

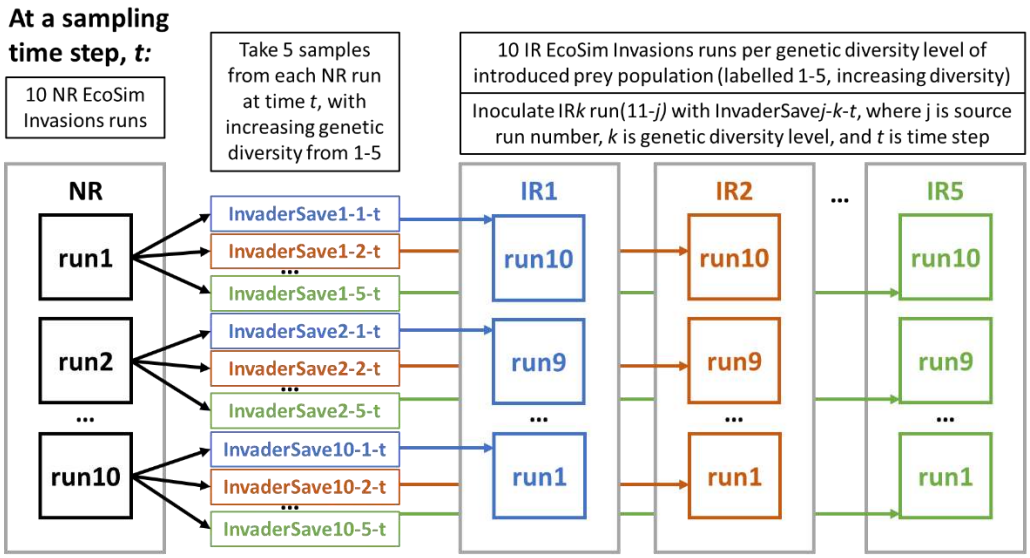


Fig. 4.5. Depiction of the sample transfer process. Prey samples taken from *NR* simulations at time step *t* are transferred to *IR* simulations and loaded in them when they reach time step *t*. There are five samples taken, corresponding to five sets of *IR* simulations – each sample and corresponding *IR* run are dedicated to a level of genetic diversity (labelled 1-5, from lowest to highest genetic diversity). Samples produced in *NRj* is transferred to corresponding *IRl*, where $l = 11-j$.

All standard EcoSim simulations and all EcoSim Niches simulations were duplicates of a single set of standard and Niches simulations, respectively. Let SO and NO be the set of original standard and EcoSim Niches simulations, respectively. SO and NO simulations were executed to 15000 time steps and then duplicated into all other simulation directories (NR standard, NR Niches, IR standard receiving standard, IR standard receiving Niches, etc.). From time step 15000 they were all executed as EcoSim Invasions simulations. Transfers always occurred from some simulation number *j* into another simulation number $l=11-j$, as mentioned above, so that no individuals would be transferred into the exact simulation in which they were produced.

Sampling Process

As mentioned in 4.2.2, custom scripts can be written to sample in any way a user may need. We leveraged this capability to perform our sampling for this study. As per above, sampling was carried out at specific time steps, and samples were taken of prey individuals that were alive in the simulation at the given time step. The sampling process consisted of three main steps – fitness-based sampling, hierarchical clustering on Shannon entropy of individual genomes, and resampling from a selected cluster with a target entropy level. The steps are details below.

The goal of the fitness-based sampling was to ensure that the fitness distribution of the sample reflected that of the population. The output of the fitness-based sampling was a sample of 1000 prey individuals, whose fitness distribution closely reflected that of the population. To obtain this sample, we first computed fitness of all prey individuals at time *t* (sample fitness histogram in Figure 4.6). We defined fitness for an individual as the number of direct children plus the number of grandchildren as per Barbosa *et al.* (2012). Subsequently, a fitness histogram was generated with seven bins (Figure 4.7). The bins were 0 fitness, 1-4 fitness, 5-8 fitness, 9-16 fitness, 17-32 fitness, 33-64 fitness, and 65+. These bins were designed in preliminary work such that all bins were likely to have non-zero representation if resampled with 1000 individuals while also considering resolution across the possible fitness values, which we have observed to range from zero to approximately 250. Using these histograms, the percentage of representation of each bin in the original sample was computed, and quotas were generated for a sample of 1000 prey individuals (Figure 4.8). That is, if the 65+ bin represented 9% of the original sample, 90 individuals would be sampled from the set of individuals with fitness over 65 to satisfy the quota for a sample of 1000 individuals. Finally, sampling took place such that 1000 individuals were obtained, adhering to these quotas.

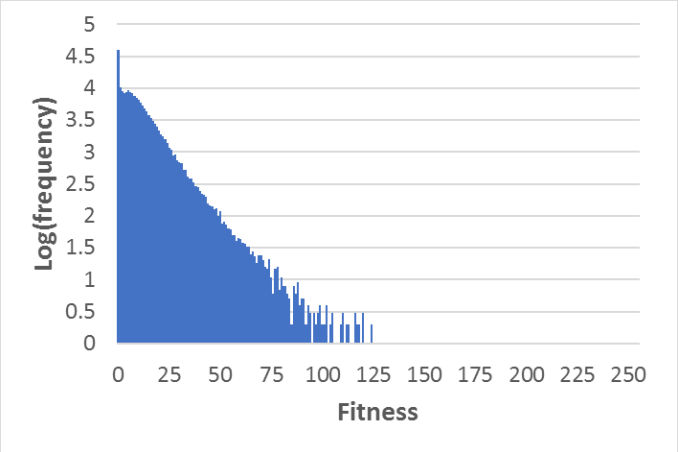


Fig. 4.6. Fitness histogram (log scaled on y-axis) at time step 15100 for a run that produced introduced populations. Same run and time step used in Figures 4.7 and 4.8.

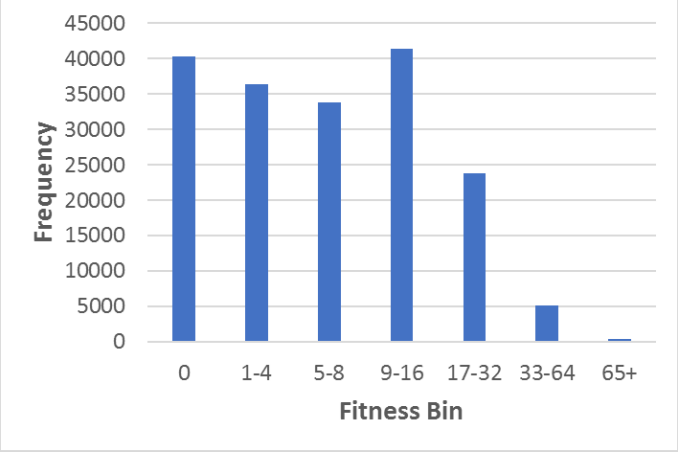


Fig. 4.7. Fitness histogram rebinned for fitness-based sampling. Same run and time step as Figures 4.6 and 4.8.

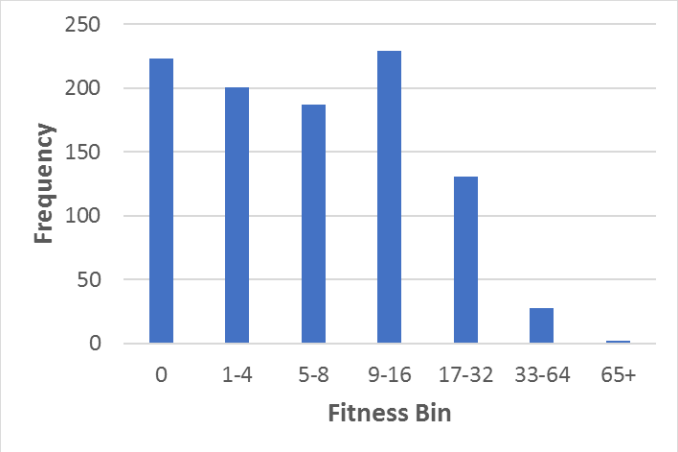


Fig. 4.8. Quotas for each fitness bin for fitness-based sampling, to produce a sample of 1000 individuals with which to hierarchically cluster based on genetic diversity. Same run and time step as Figures 4.6 and 4.7.

Once fitness-based sampling was complete, hierarchical clustering of the sample was performed. Hierarchical clustering is an algorithm that aims to produce a hierarchy of clusters, such that the clusters are produced in order of an increasing or decreasing metric. This was ideal for our purposes, as we needed to produce subsamples of prey individuals that existed on a high-resolution gradient of genetic diversity. Thus, we chose Shannon entropy of prey genomes (GE, Eq. 4.2; hereby referred to as genetic entropy), measured in bits, as our measurement of genetic diversity as per Khater, Salehi, and Gras (2011).

Hierarchical clustering can be computed in an agglomerative (bottom-up) or divisive (top-down) manner. The agglomerative algorithm, which we used, was computed as follows:

1. Randomly partition original set of 1000 individuals into n clusters containing k individuals each (we used $k = 4$ to produce $n = 250$ initial clusters, informed by preliminary work described below), add each cluster to vector of clusters, *clusters*
2. While size of *clusters* > 1:
    a. Compute pairwise genetic entropy for all pairs of clusters as if they were combined
    b. Find the pair of clusters minimizing genetic entropy when combined, C_1 and C_2
    c. Combine C_1 and C_2 into C_{new}
    d. Remove C_1 and C_2 from *clusters*, add C_{new} to *clusters*
3. Output all clusters generated throughout the process, sorted in terms of increasing genetic entropy

We kept a record of every cluster that was created during this process, along with its genetic entropy, the individuals it contains, and its fitness histogram using the fitness bins defined above. This allowed us to subsequently select clusters with genetic entropy levels closest to target genetic entropy levels, which we could resample and use as populations to introduce across EcoSim simulations (Figure 4.9). We used empirical data to devise our entropy targets for the introduced samples, described below. Thus, we controlled for genetic diversity of our introduced populations.

Genetic entropy was computed as follows:

$$GE(X) = -\sum_{i=0}^{n-1}\sum_{j=0}^{m-1} p_{ij} log_2(p_{ij}), \tag{4.2}$$

where X is a set of genetic sequences (i.e. vectors of bytes) of length n, i is a particular gene (i.e. zero-indexed position in the vector), and there are m potential alleles following the continuum-of-alleles model (Kimura 1965). As there are 663 elements in the prey behavioural genome and 8 elements in the prey physical genome, n=671. As the genetic sequences are byte vectors, m=256. Finally, p_{ij} is thus the probability of allele j at gene i in the set of sequences X (i.e. the probability of a particular byte value at a given index of the sequence vector). The minimum GE for a set of genetic sequences is zero, occurring when all sequences are exactly the same. As we know the size of the space of possibilities for genomes, we can compute a theoretical maximum GE. The theoretical maximum GE represents the GE when there is perfectly even representation of every possible genome in the set of sequences. With n=671 and m=256, the maximum GE is 5368.

In preliminary work, we performed hierarchical clustering on samples of 1000 prey individuals, produced with our fitness-based sampling method (described above). The samples were obtained from ten EcoSim simulations at time steps 15100 to 20100, every 100 time steps. Our aim was to determine potential genetic entropy targets for our introduced populations while simultaneously tuning k and n of our hierarchical clustering algorithm to maximize efficiency and maintain a high resolution of genetic diversity across the clusters we produced. We found that k = 4 and n = 250 produced a reasonably high-resolution gradient of genetic diversity, and that genetic diversity targets of 0 (hereby denoted L1), 105 (hereby denoted L2), 210 (hereby denoted L3), 315 (hereby denoted L4), and 420 (hereby denoted L5) bits would be feasible on most population samples (Figure 4.10). Hierarchical clustering was only used to obtain samples for L2-L5, while L1 samples were obtained by randomly selecting a single individual as described below. The output of the hierarchical clustering process was the sorted list of all clusters generated throughout the process, which we ranked from 1 to 499 in terms of increasing genetic entropy. With respect to the edge cases, obtaining a rank-1 cluster with ≤105 bits of genetic entropy was feasible in 92.9% of samples, while obtaining a rank-499 cluster (out of 499 clusters) with ≥420 bits of genetic entropy was feasible in 84.1% of samples. Because of the high resolution of the diversity gradient with k = 4 and n = 250, this meant that the above diversity targets would likely be sufficiently achieved. The mean genetic entropy of the resultant selected clusters were 0, 107.33, 210.12, 315.01, and 419.52 bits, with respective standard deviations of 0, 11.55, 1.86, 0.23, and 2.91 bits.

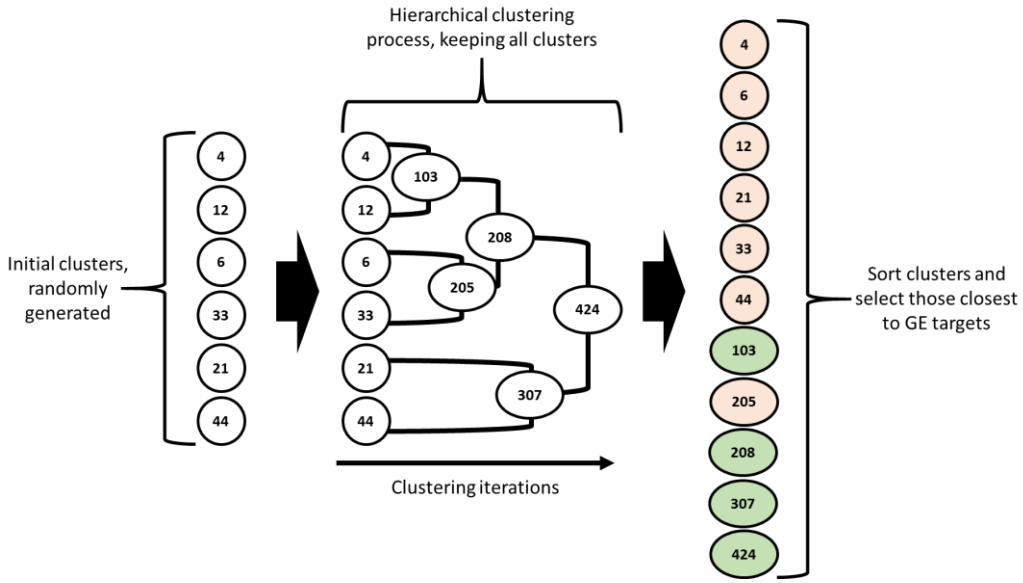


Fig. 4.9. Hierarchical clustering toy example, for illustrative purposes. Circles represent initial clusters, ovals represent intermediate clusters. Numbers within circles and ovals represent cluster genetic entropy (GE). Start with six initial clusters each containing four randomly selected prey individuals. During hierarchical clustering process, per iteration, combine the two clusters yielding minimum genetic entropy when combined. Continue this process until only one cluster remains. When clustering is complete, sort clusters based on genetic entropy and select those exhibiting genetic entropy closest to target entropy values (105, 210, 315, and 420 bits). Selected clusters are shown in green, discarded clusters are shown in red.

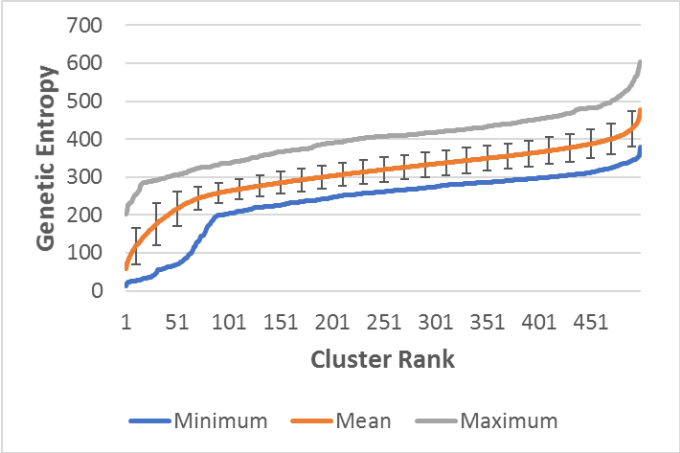


Fig. 4.10. Genetic entropy of clusters by rank, in order of increasing genetic entropy, with $k = 4$. Minima, maxima, mean, and standard deviation of genetic entropy are shown for each rank, across 510 samples from ten EcoSim simulations.

The ranked clusters produced by the hierarchical clustering process had a variable number of individuals per cluster. We required that our prey samples had a fixed number of individuals (100). Thus, we resampled the four selected clusters (the clusters with genetic entropy closest to our non-zero target entropy levels, L2-L5), with replacement, to produce four corresponding samples to introduce into other EcoSim simulations. By definition of genetic entropy, assuming perfectly even resampling of the individuals, entropy should remain exactly the same after the resampling process. Although our

sampling was naturally imperfect in terms of evenness, genetic entropy of the resultant samples was almost completely unaffected by this resampling process (Figure 4.11). To obtain the sample with zero bits of entropy at each time step, we simply selected a single individual and resampled it 100 times. When generating samples, only genetic information (i.e. physical and behavioural genomes) of the individuals was copied. All other state variables (i.e. sex, energy level, speed, etc.) were obtained by randomly sampling the variables from the original source population and, where necessary, clipped to be within the range specified by their associated genes (e.g. *MaxEnergy*, *MaxSpeed*, etc.). Thus, the distribution of all non-genetic state variables in the original population was approximately preserved in the introduced sample.

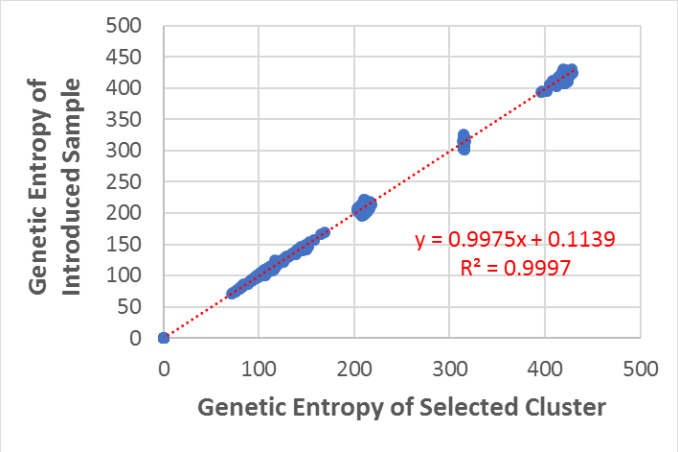


Fig. 4.11. Strong concordance of genetic entropy of clusters from hierarchical clustering process (x-axis) and that of corresponding introduced samples (y-axis) after resampling the clusters to produce samples of 100 individuals.

4.2.5 Data Analysis

To compare between standard EcoSim and EcoSim Niches, we collected various time-series data up to 15000 time steps (i.e. just prior to reciprocal transplantation between the simulations) and examined them for significant differences between the two EcoSim variants. Behavioural and physical genome elements (i.e. genes) were also obtained from a sample of 40000 prey and 1500 predator individuals that were alive at time step 15000 from a single run of each variant. For each gene, between the two variants, we conducted a t-test and Levene's test to test for difference in mean and variance, respectively. Bonferroni correction was used to account for the extremely high number of comparisons (663 for prey, 756 for predators).

With data from IR simulations, we first analyzed the time series data for introduced log(abundance) and genetic entropy throughout the course of introductions for 3500 time steps (as eight simulations did not progress much further due to extinctions), per IR run type and across the five genetic diversity levels. This provided us with a high-level view of how the invasions progressed and how the genetic diversity of the

introduced populations changed over time. The data were highly variable and noisy due to repeated failed introduction attempts, so each time-series was smoothed using a rolling average with a window size of 100. We then computed the means and standard deviations for each time step across all simulations of a given treatment (combination of run type and genetic diversity) and plotted the results as time series.

With respect to hypotheses I and II, biologists typically use presence/absence or abundance estimates to measure the ability of an introduced population to establish. With EcoSim, we can assess establishment success in a variety of ways that are typically infeasible in studies of real invasions. We quantified short-term establishment success in IR simulations by recording the proportion of introduction events, during which the current abundance of introduced individuals in the run was zero (which we henceforth refer to as a fresh inoculation), in which the introduced population persisted after 60 time steps (henceforth referred to as short-term establishment success). Further, we quantified the proportion of simulations of a given group in which the abundance of introduced individuals was greater than 1000 at time step 20200, as a proxy of its long-term establishment (henceforth referred to as long-term establishment success). For short-term and long-term establishment success, we conducted z-tests for proportions within groups created by run type to determine significance of difference in proportions across genetic diversity levels. Similarly, we conducted z-tests for proportions between groups created by run type to test for significant differences in overall proportions across run types, for both establishment success measures.

To determine how other factors significant for establishment success might change given genetic diversity, we gathered mean values for select features descriptive of the introduced population for every fresh inoculation. The means were obtained over five time steps, starting five time steps after the inoculation event, to allow the population to settle into their novel environment and to reduce stochasticity. We analyzed these data, per run type, for differences in distribution across genetic diversity levels, separately for all, successful, and failed short-term establishment attempts (as defined above), using Kruskal-Wallis tests. Upon rejection of the null hypothesis for a given feature using Kruskal-Wallis (i.e., the feature showed a significantly different distribution for some genetic diversity level across all, successful, or failed establishments attempts), we conducted pairwise Conover's *post hoc* tests, with p-values adjusted using Holm correction, to find which genetic diversity levels that yielded significant differences in distribution for the given feature.

To select the features that we analyzed using the above process, we used a combination of permutation importance (Strobl *et al.* 2007) and Spearman's rank-order correlation as follows. We first devised a set of 29 population-wide features that we

anticipated to have bearing on the establishment success of the introduced populations (e.g. speed, compactness, *Energy*, *MaxSpeed*, *MaxEnergy*, *Vision*, proportion of individuals performing each action, etc.). In addition to these 29 features, we used the short-term establishment outcome as the classification target. We computed permutation importance of each feature with random forests (Breiman 2001) using the Scikit-Learn and ELI5 Python packages (https://scikit-learn.org/, https://eli5.readthedocs.io/en/latest/) to classify the samples, as follows. We first split the entire dataset into a training set and testing set, and developed a random forest trained using the training set. The random forest had 1000 trees, and all other parameters were left to their default values. With this random forest as a benchmark, we obtained the classification area under the receiver operating characteristic curve (AUROC), computed using the testing data. For each feature, we randomly permuted the data for the selected feature while holding all other features as they were, and trained sets of ten random forests with the same parameterization to observe the loss in testing AUROC of these random forests. The greater the loss in testing AUROC, the more important the feature was in generally characterizing establishment success. The output of the process was the ranked importance of each feature along with its corresponding mean and standard deviation of loss, when randomly permuted, over the ten random forests. Correlated features suffer asymmetrically using permutation importance (Strobl *et al.* 2007); that is, both correlated features may suffer some loss in perceived importance, but the loss in importance for each feature is unequal. So, in conjunction with this process we computed pairwise Spearman's rank-order correlations across all features. For each pair of features yielding absolute(rho)$\geq$0.5, we removed the less important feature given the ranking from above. With the less-important correlated features removed, we then recomputed the feature importance using permutation importance again and selected the most important features.

4.3 Results

We divide the results into four main sections, and the discussion loosely follows this format as well. First, we provide results pertaining to the differences between standard EcoSim and EcoSim Niches. We then provide a high-level view of the invasion progress for each run type and genetic diversity level. We then present results pertaining to hypothesis I – that genetic diversity has a positive relationship with establishment success – and hypothesis II – that genetic diversity has a stronger impact when the introduced populations' native and novel regions greatly differ. Finally, we provide other insights about a variety of factors leading to success or failure of establishment, in light of the amount of genetic diversity of the introduced populations.

4.3.1 Comparison of standard EcoSim and EcoSim Niches

Standard EcoSim and EcoSim Niches produced different environments and consequently different individuals. Based on the formulation of EcoSim Niches (Section 4.2.3) there were clear differences in resource availability and distribution. We sought to determine whether these differences impacted the evolution of prey and predator individuals in the simulations.

Grass per cell differed significantly between the two variants throughout the simulations (Figure 4.12). Standard EcoSim yielded higher *Grass* levels than EcoSim Niches, as intended. Having fewer global resources in EcoSim Niches effectively reduces its carrying capacity, and consequently there were significantly fewer prey (Figure 4.13) and predator (Figure 4.14) individuals. Despite the disparity in number of prey and predators between the two variants, differences between the two variants in prey species richness (number of prey species in the virtual world; Figure 4.15) and predator species richness (not depicted) were often insignificant.

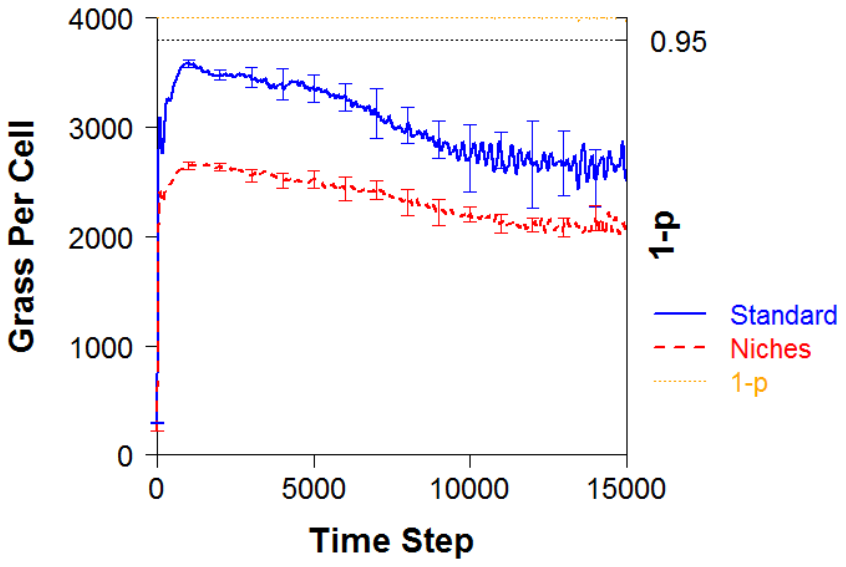


Fig. 4.12. Comparison of amount of *Grass* per cell (left y-axis) between standard EcoSim and EcoSim Niches. Difference between variants was significant for nearly the entirety of the simulations. *Grass* levels in both variants stabilized at approximately 10000 time steps. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

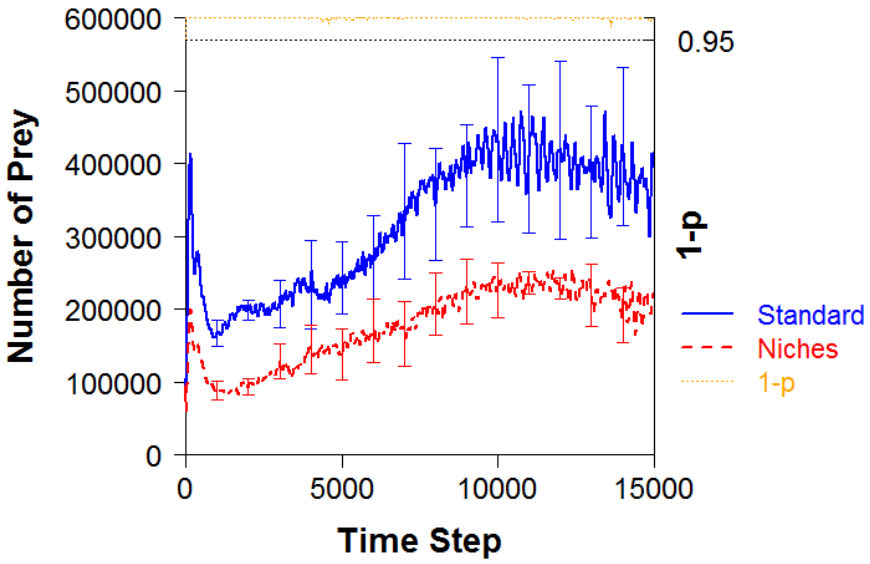


Fig. 4.13. Comparison of number of prey individuals (left y-axis) between standard EcoSim and EcoSim Niches. Difference between variants was significant for nearly the entirety of the simulations. Quasi-stationarity was observed from approximately 10000 time steps. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

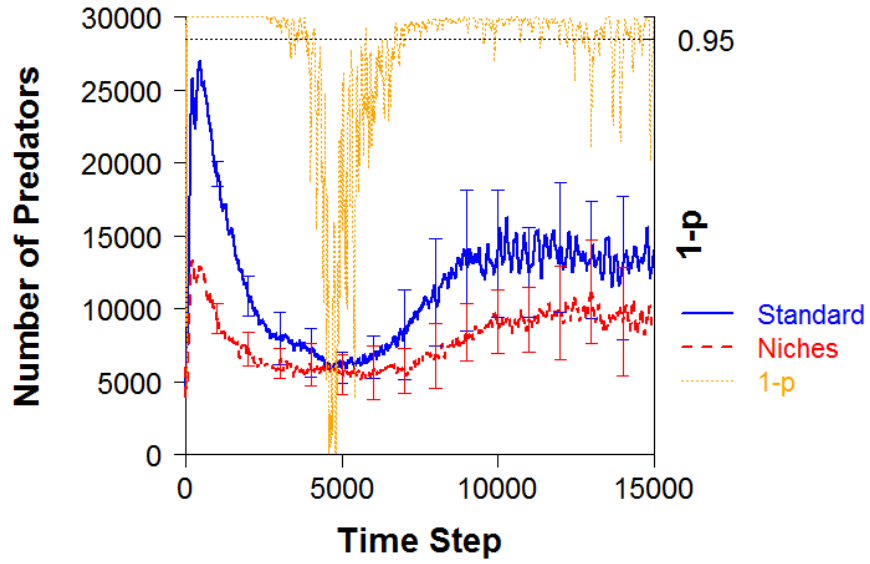


Fig. 4.14. Comparison of number of predator individuals (left y-axis) between standard EcoSim and EcoSim Niches. Difference between variants was significant for nearly the entirety of the simulations. Quasi-stationarity was observed from approximately 10000 time steps. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

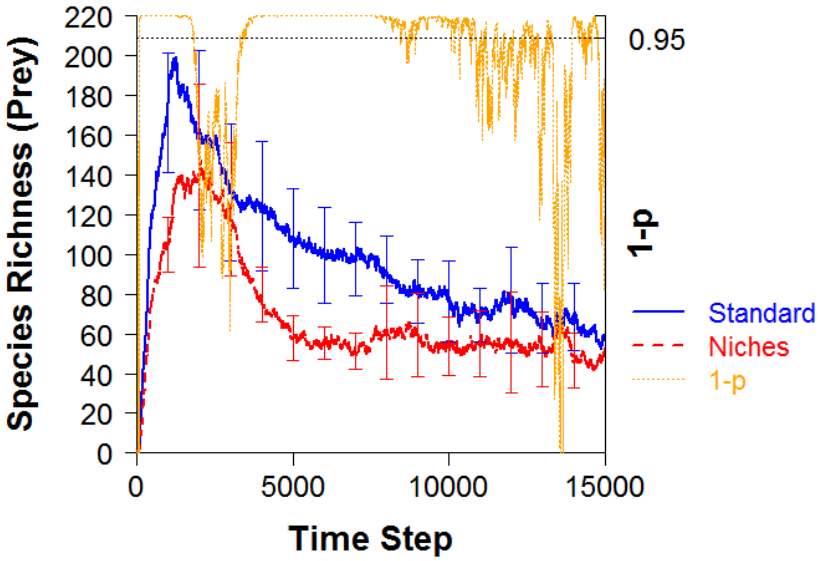


Fig. 4.15. Comparison of prey species richness (number of prey species in the virtual world; left y-axis) between standard EcoSim and EcoSim Niches. The disparity in prey species richness was disproportionate to the disparity in number of prey individuals (Figure 4.14), and convergence was apparent in species richness. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

Energy spent by prey per time step differed significantly for most of the duration of the simulations (Figure 4.16), and prey in standard EcoSim tended to spend more energy than those in EcoSim Niches. Differences in *Energy* spent by predators (not depicted) was sometimes significant, with those in standard EcoSim spending more than those in EcoSim Niches. Further, predators in standard EcoSim usually had significantly more *Energy* than those in EcoSim Niches (Figure 4.17), while the differences in prey *Energy* (not depicted) were mostly insignificant.

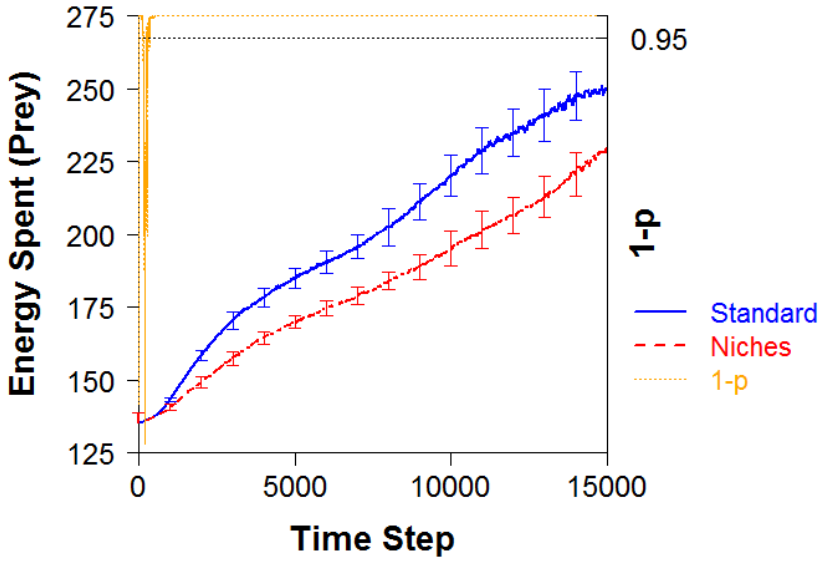


Fig. 4.16. Comparison of *Energy* spent per time step by prey (left y-axis) between standard EcoSim and EcoSim Niches. Difference between variants was significant for nearly the entirety of the simulations. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

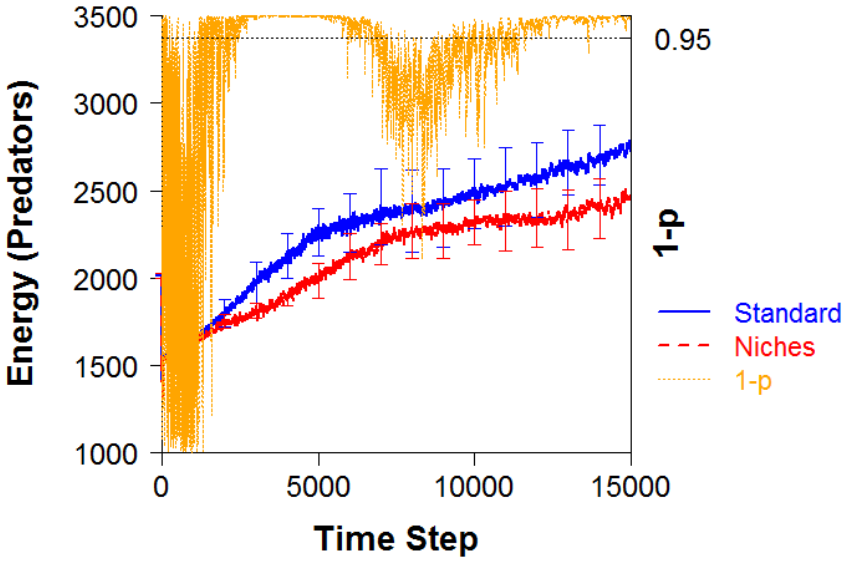


Fig. 4.17. Comparison of predator *Energy* levels (left y-axis) between standard EcoSim and EcoSim Niches. Difference between variants was usually significant. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

Between the two variants, there were significant differences in *MaxSpeed* for prey (Figure 4.18) and predators (Figure 4.19). Similarly, prey *Speed* exhibited significant

differences between the two variants (Figure 4.20), as did *Vision* for prey (Figure 4.21) and predators (Figure 4.22). In all cases, the value in standard EcoSim was greater than that in EcoSim Niches. Predator *Speed* did not show significant difference (not depicted). Distance evolved for prey (Figure 4.23) was significantly greater in standard EcoSim than in EcoSim Niches, while predator compactness (the mean number of individuals per cell containing at least one individual, Figure 4.24), predator distance evolved (Figure 4.25), prey number of FCM edges (Figure 4.26), and predator number of FCM edges (Figure 4.27) were all significantly greater in EcoSim Niches than in standard EcoSim.

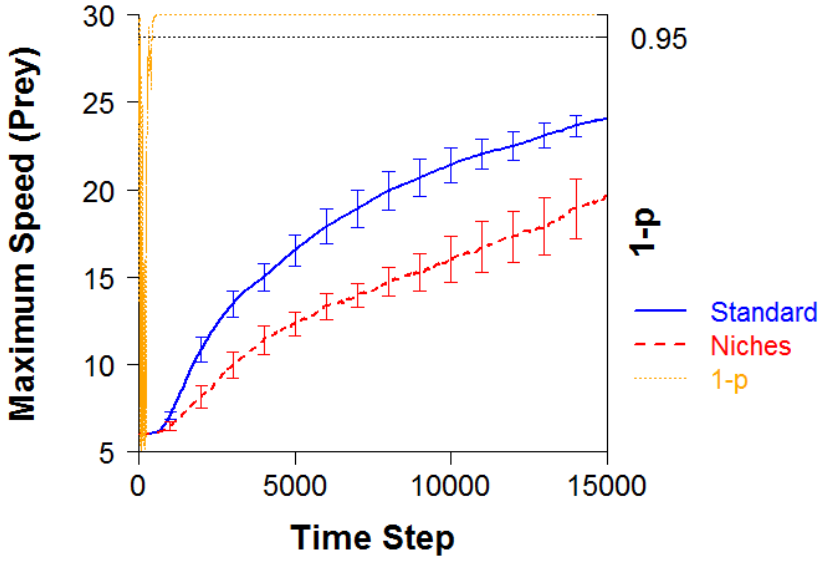


Fig. 4.18. Comparison of prey *MaxSpeed* (left y-axis) between standard EcoSim and EcoSim Niches. Difference between variants was significant for nearly the entirety of the simulations. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

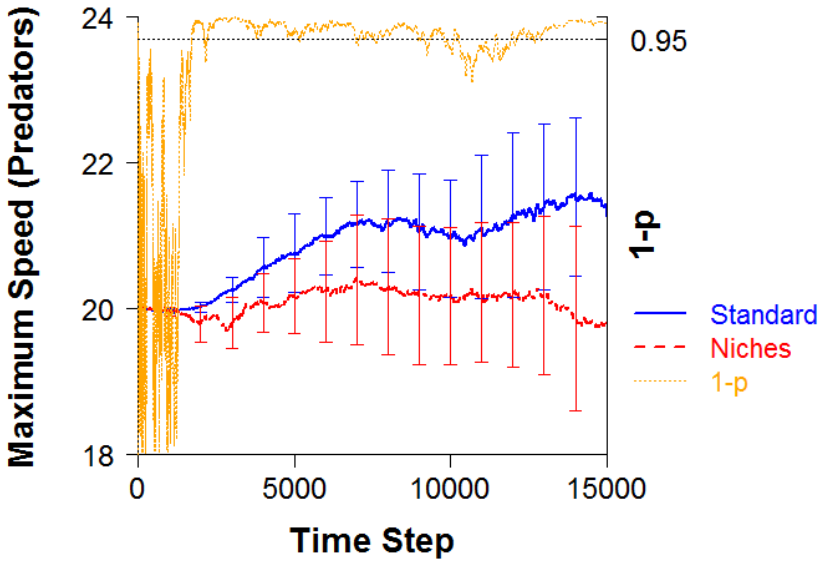


Fig. 4.19. Comparison of predator *MaxSpeed* (left y-axis) between standard EcoSim and EcoSim Niches. Difference between variants was significant for nearly the entirety of the simulations, and in general high variance was observed for both variants. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

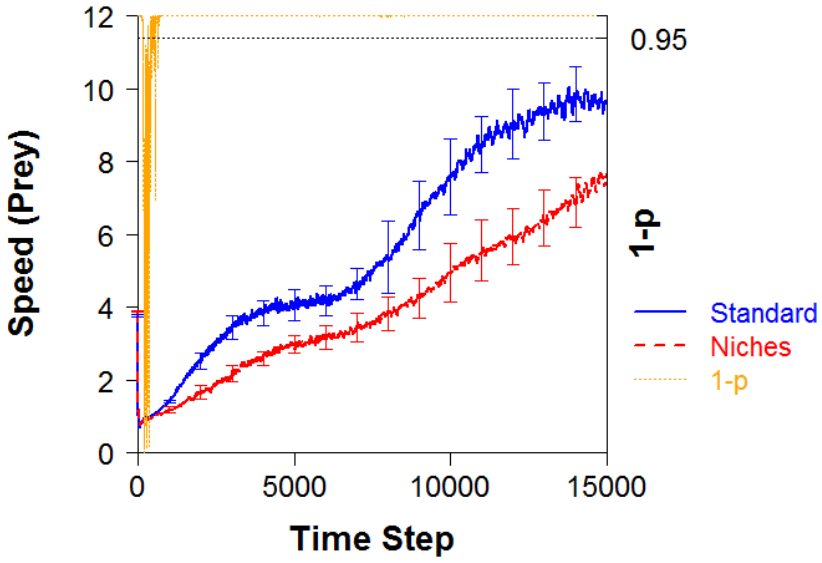


Fig. 4.20. Comparison of prey *Speed* (left y-axis) between standard EcoSim and EcoSim Niches. Difference between variants was significant for nearly the entirety of the simulations. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

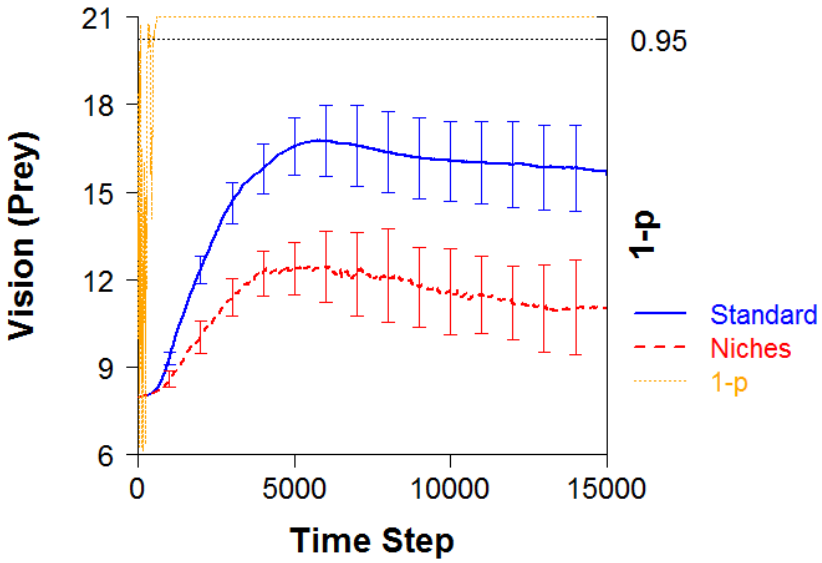


Fig. 4.21. Comparison of prey *Vision* (left y-axis) between standard EcoSim and EcoSim Niches. Difference between variants was significant for nearly the entirety of the simulations. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

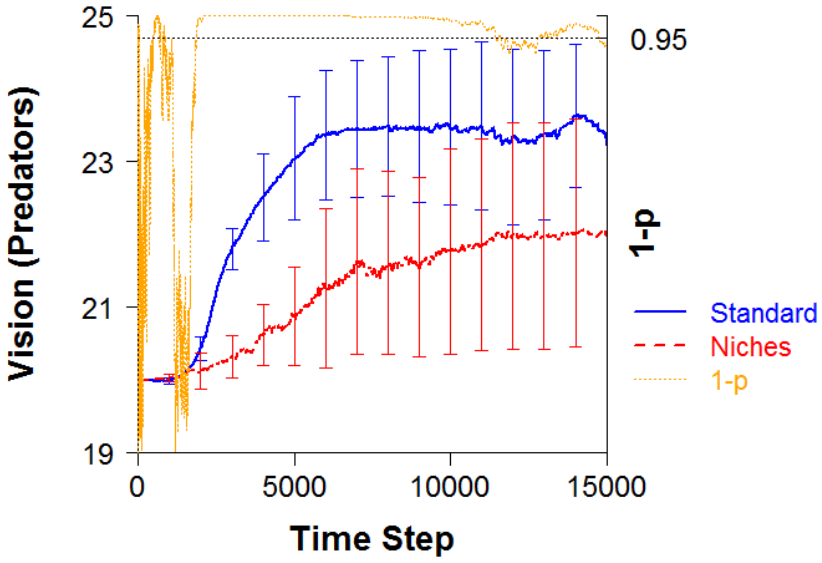
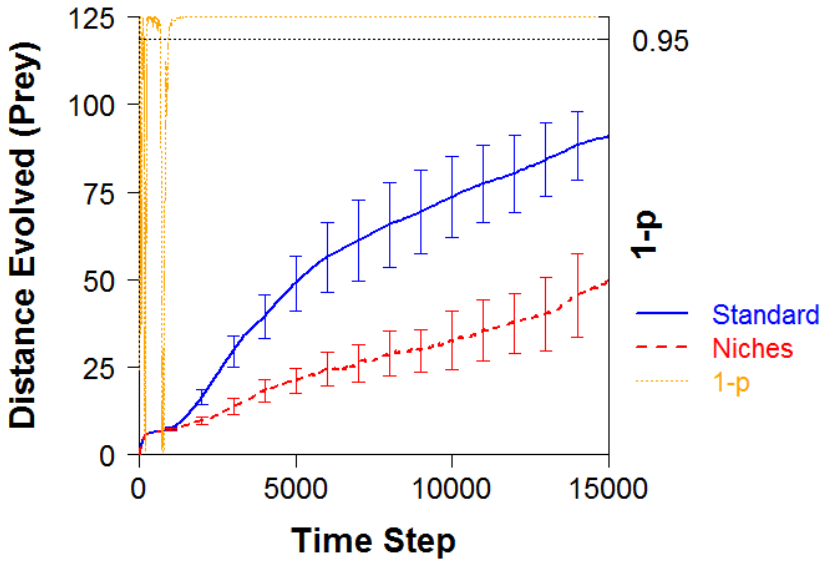


Fig. 4.22. Comparison of predator *Vision* (left y-axis) between standard EcoSim and EcoSim Niches. Difference between variants was typically significant. High variance in *Vision* was generally observed. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.



A

Fig. 4.23. Comparison of prey distance evolved (left y-axis) between standard EcoSim and EcoSim Niches. Prey individuals in standard EcoSim evolved from their initial genome faster than those in EcoSim Niches. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

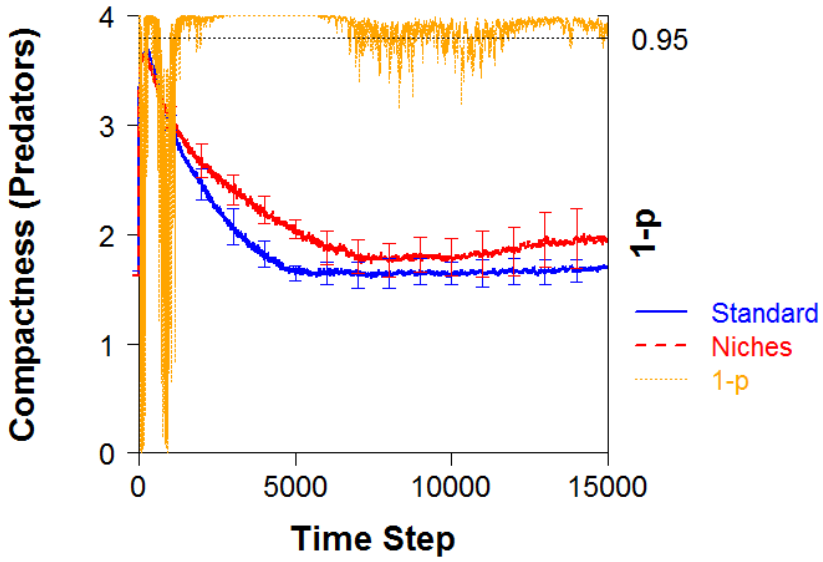


Fig. 4.24. Comparison of predator compactness (left y-axis) between standard EcoSim and EcoSim Niches. Predator compactness was significantly greater in EcoSim Niches for the majority of simulation duration. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

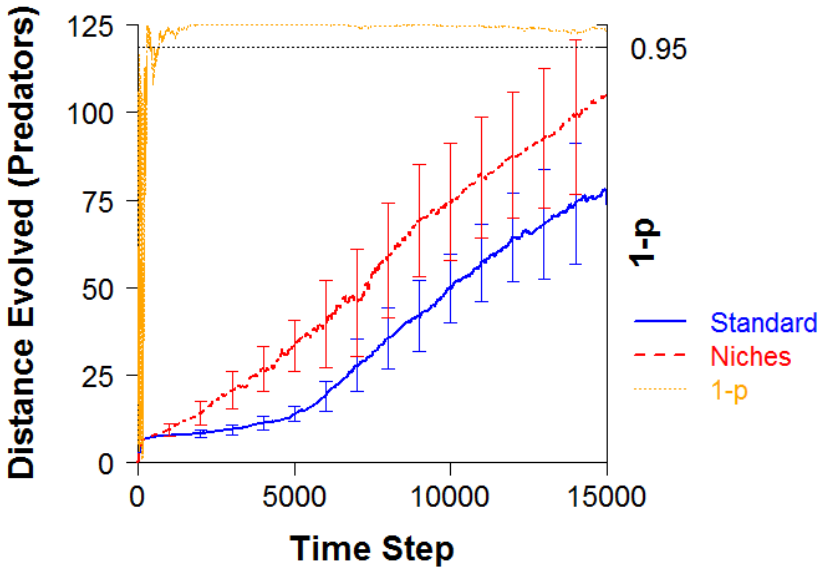


Fig. 4.25. Comparison of predator distance evolved (left y-axis) between standard EcoSim and EcoSim Niches. Predator individuals in EcoSim Niches evolved from their initial genome faster than those in standard EcoSim initially, but ultimately their evolutionary trajectories were almost identical. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

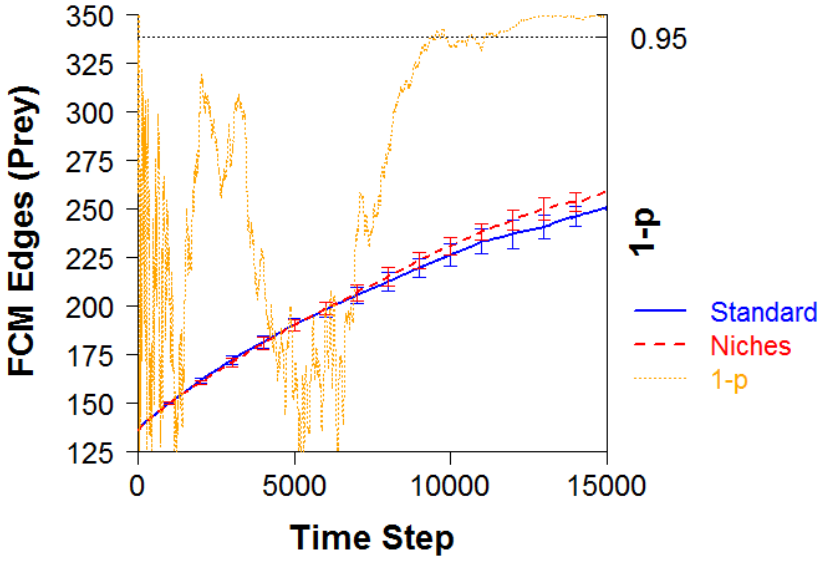


Fig. 4.26. Comparison of number of FCM edges for prey (left y-axis), between standard EcoSim and EcoSim Niches. Difference between variants was significant at approximately 10000 time steps, and divergence was apparent as time progressed. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

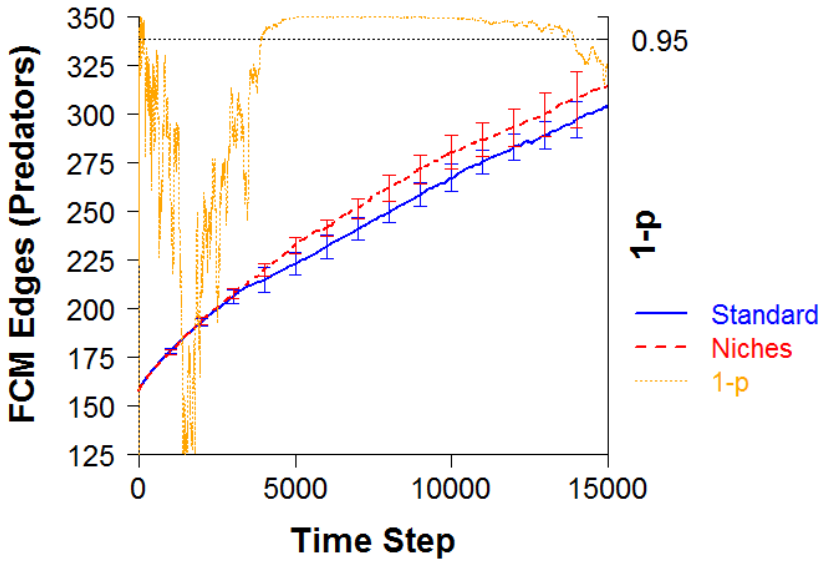


Fig. 4.27. Comparison of number of FCM edges for predators (left y-axis), between standard EcoSim and EcoSim Niches. Difference between variants was significant at approximately 3000 time steps, though convergence seemed to be occurring from approximately 13000 time steps and onward. Significance shown in orange as 1-*p* from t-test (right y-axis). Error bars denote one standard deviation of the mean.

In addition to the above differences between the two variants, we discovered that behavioural and physical gene distributions exhibited numerous differences in mean and variance. Overall, 93.4% of prey behavioural genes exhibited significant differences in mean ($p < 0.05/663$, t-test with Bonferroni correction) while 81.7% of behavioural genes showed significant differences in variance ($p < 0.05/663$, Levene's test with Bonferroni correction). Further, in genes where the difference in variance was significant, prey from EcoSim Niches exhibited the greater variance 55.4% of the time. For predators, 90.3% of behavioural genes had significant differences in mean ($p < 0.05/756$, t-test with Bonferroni correction) while 87.8% of behavioural genes exhibited significant differences in variance ($p < 0.05/756$, Levene's test with Bonferroni correction). Of the genes exhibiting significant differences in variance, 66.0% showed greater variance in EcoSim Niches than in standard EcoSim.

4.3.2 Invasion progress over time

The invasions for each run type progressed in different ways, and temporal relationships between levels of genetic diversity and abundance of the introduced individuals were inconsistent across the different run types (Figure 4.28). The abundance of introduced individuals in N→S (8429) was approximately five times greater than that of S→S

(1520), on average, 3500 time steps into the introductions. A similar relationship was observed when comparing the number of introduced individuals in N→N (72 individuals) and S→N (17 individuals), at the same point in simulation time. S→S versus S→N and N→S versus N→N each differed by approximately two orders of magnitude. In all cases, variation in the time-series data was extremely high; in many cases the abundance in a given run was upwards of 400,000 individuals while in another run it was zero. In all but S→S, L1 (i.e., each inoculation was a population of clones) and L5 (i.e. each inoculation was of maximal genetic diversity) genetic diversity levels yielded the highest mean abundance after 3500 time steps. On the other hand, intermediate diversity levels (particularly L2 and L3) tended to yield lower abundances. Particularly in S→S and N→S, it appears as though the introduced abundance has reached the carrying capacity for the respective simulations.

In terms of genetic diversity of introduced populations, S→S and N→S yielded increasing trends for all inoculation diversity levels (Figure 4.29) with L1 showing the greatest increases – even temporarily surpassing the diversity of L2 in S→S and surpassing that of L2 and L3 in N→S. In S→N, only L1 showed an increasing trend, eclipsing the genetic diversity of L2 populations, while all other diversity levels were stable. N→N, genetic diversity of L1, L2, and L4 showed increasing trends while L3 and L5 were stable.

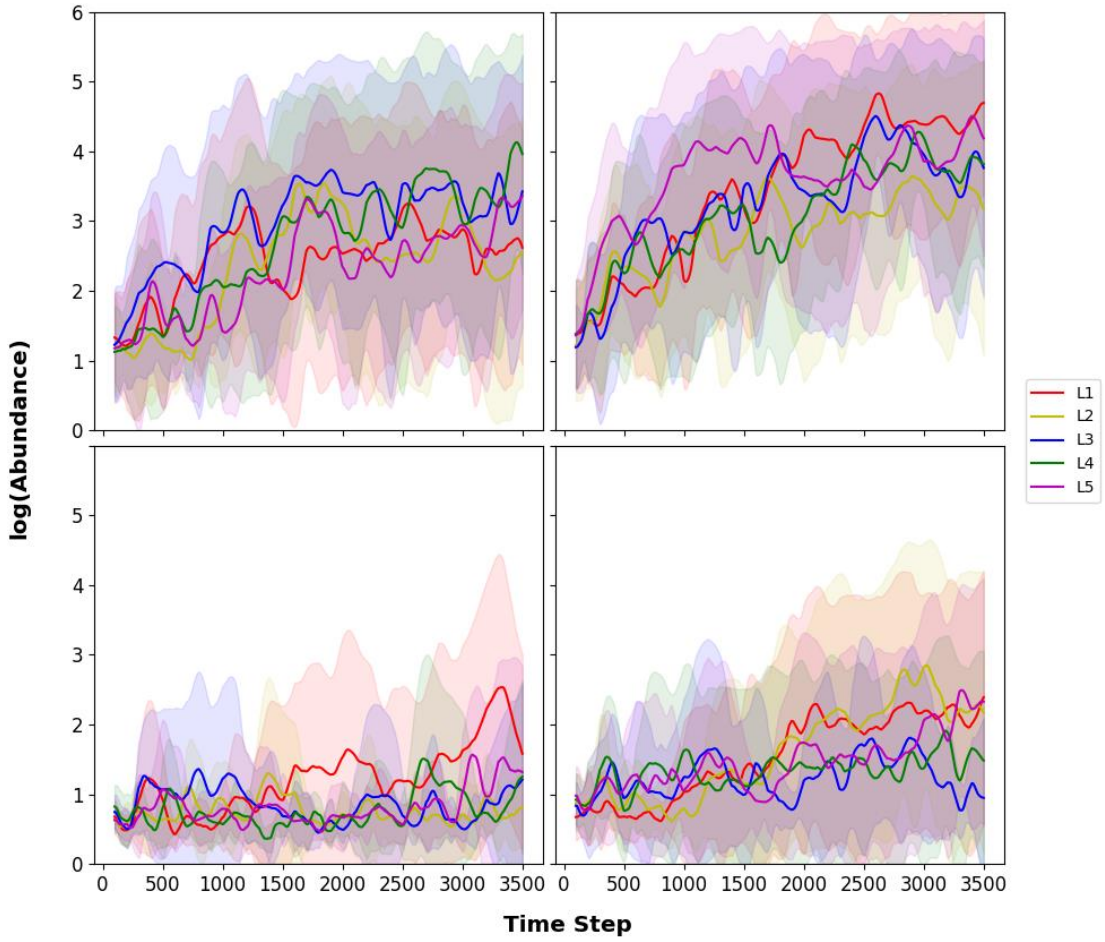


Fig. 4.28. Introduced individual abundance, $\log_{10}$-scaled, over time. Top left: S→S. Top right: N→S. Bottom left: S→N. Bottom right: N→N. Lines represent means for different genetic diversity levels, while corresponding shaded areas represent mean ± one standard deviation.

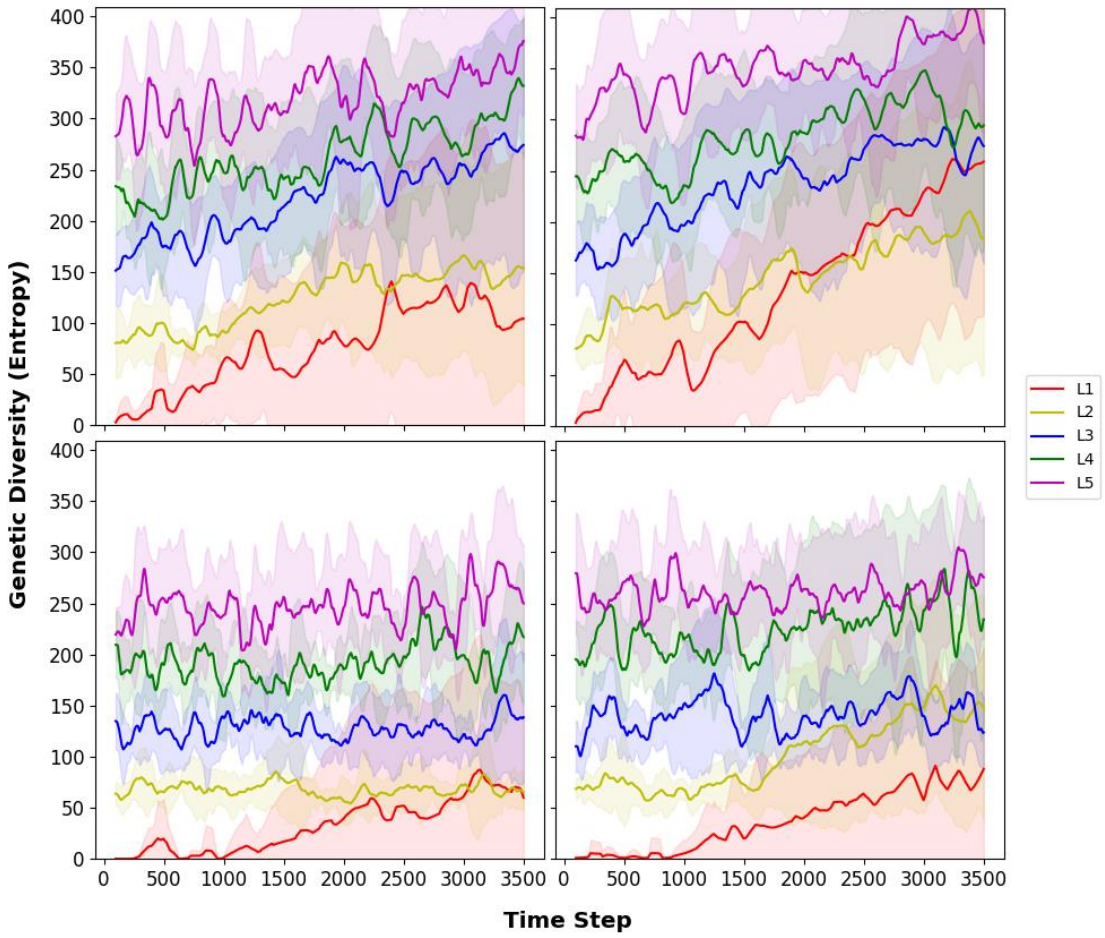


Fig. 4.29. Genetic diversity of introduced populations over time. Top left: S→S. Top right: N→S. Bottom left: S→N. Bottom right: N→N. Lines represent means for different genetic diversity levels, while corresponding shaded areas represent mean ± one standard deviation.

4.3.3 Hypotheses I and II

The differences in short-term establishment success between run types (i.e. aggregated across genetic diversity levels) were all highly significant ($p \leq 0.0014$ in all comparisons, z-test for proportions; Figure 4.30). Standard EcoSim yielded environments that were more invasible in the short term than EcoSim Niches as evidenced by significant differences in short-term establishment success between the two environments for each source of introduced individuals. Similarly, EcoSim Niches produced populations that were more capable of establishing over the short term, as evidenced by significant differences between sources when transferred to the same environments. Overall, N→S yielded the highest short-term establishment success with a success rate of 77.9%, S→S yielded 70.7%, N→N yielded 52.6%, and S→N yielded 38.8%. There were some weakly significant differences across genetic diversity levels for short-term establishment success. In S→S, L3 (76.7%) short-term establishment success was slightly higher than

that of L2 (68.9%; $p = 0.093$, z-test for proportions) and L5 (68.5%; $p = 0.077$, z-test for proportions). In S→N and N→N, short-term establishment success slightly improved with increased genetic diversity. In S→N, L1 (36.5%) and L5 (43.1%) yielded different short-term establishment success ($p = 0.067$, z-test for proportions). In N→N, that of L1 (47.2%) was lower compared to L3 (54.7%; $p = 0.057$, z-test for proportions), L4 (55.7%; $p = 0.033$, z-test for proportions), and L5 (56.2%; $p = 0.026$, z-test for proportions), while that of L2 (48.6%) was lower compared to L4 ($p = 0.071$, z-test for proportions) and L5 ($p = 0.057$, z-test for proportions).

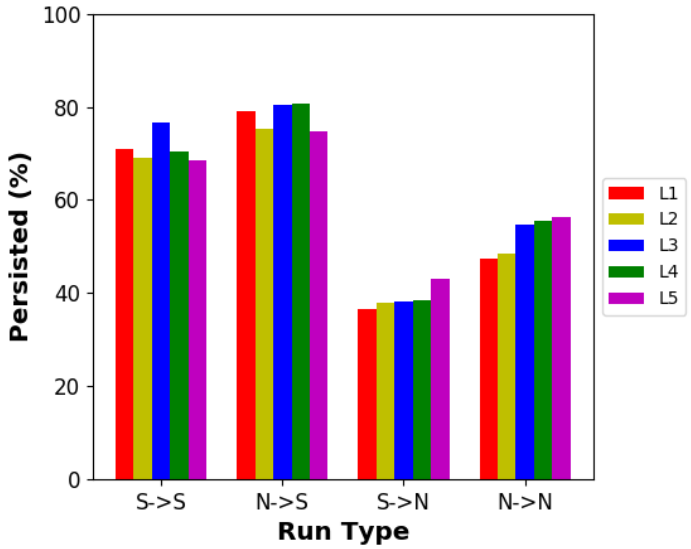


Fig. 4.30. Short-term establishment success, measured by the proportion of inocula persisting for 60 time steps in simulated environments otherwise uninhabited by introduced individuals. S→S yielded a shape indicating that a moderate amount of genetic diversity was most favourable. S→N, and N→N yielded similar shapes, with high genetic diversity values being most favourable. Differences between run types (aggregated across genetic diversity levels) were all highly significant. Standard EcoSim environments were more invasible than EcoSim Niches, while individuals from EcoSim Niches showed greater invasiveness than those from standard EcoSim.

Long-term establishment success (Figure 4.31) yielded patterns similar those observed in analysis of overall establishment success, in terms of comparisons made between run types. Significant comparisons were of S→S (74%) to N→S ($p = 0.016$), S→S to S→N (38%; $p = 0.00029$), S→S to N→N ($p = 0.037$), N→S (92%) to S→N ($p = 1.5\text{x}10^{-8}$), N→S and N→N (54%; $p = 1.9\text{x}10^{-5}$). The comparison between S→N and N→N yielded insignificant difference ($p = 0.11$). Due to the small sample size, comparisons within a given run type (across genetic diversity levels) were not statistically significant, though the patterns observed in long-term establishment were quite different from those observed in short-term establishment. L2 of S→N yielded no long-term establishment.

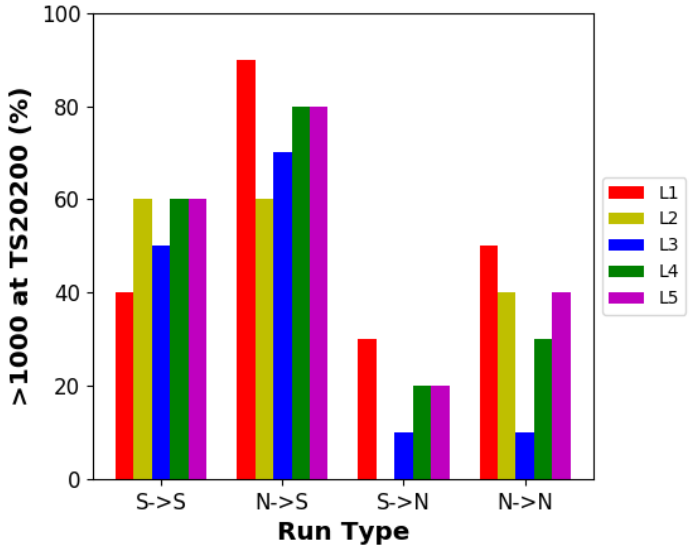


Fig. 4.31. Long-term establishment success, measured by the proportion of simulations in which the abundance of introduced individuals was greater than 1000 at time step 20200. Aside from S→S, all run types yielded similar shapes, indicating that extreme diversity levels were most favourable while moderate genetic diversity was unfavourable. Standard EcoSim environments were more invasible than EcoSim Niches, while individuals from EcoSim Niches showed greater invasiveness than those from standard EcoSim.

4.3.4 Other factors affecting establishment success

Compactness (0.08±0.015; mean ± standard deviation of random forest AUROC loss when feature was randomly permuted), proportion of individuals attempting to eat (performance loss of 0.024±0.0041), and reproduction efficiency (performance loss of 0.013±0.0068) were amongst the most important features in predicting establishment success prior to removal of correlated features. Due to high correlation, we removed proportion of population exploring (correlated with speed, ρ=0.93), distance evolved (correlated with *MaxSpeed*, ρ=0.93), energy spent per time step (correlated with speed, ρ=0.91), *Energy* (correlated with eat attempts, ρ=0.89), *Vision* (correlated with *MaxSpeed*, ρ=0.8), proportion of individuals failing to move to the strongest prey in their vicinity (correlated with proportion of individuals moving to the strongest prey in their vicinity, ρ=0.75), *MaxEnergy* (correlated with *Strength*, ρ=0.6), *Speed* (correlated with proportion of individuals attempting to reproduce, ρ=0.56), proportion of individuals attempting to eat (correlated with compactness, ρ=0.56), and reproduction efficiency (correlated with compactness, ρ=0.52). After removing these features, the most important features in determining establishment success were compactness (performance loss of 0.23±0.020; Figure 4.32), proportion of population failing to socialize (performance loss of 0.017±0.0087; Figure 4.33), proportion of individuals socializing (performance loss of 0.0125±0.006; Figure 4.34), and the proportion of individuals escaping from predators (performance loss of 0.0118±0.0027; Figure 4.35). The remaining factors were of

extremely low importance (i.e. <1% performance loss), and hence we restricted subsequent analysis to these four features. The benchmark random forest had an AUROC of 0.89 and an F1 score of 0.87 on the testing dataset, showing that the model performed extremely well and generalized to unseen data.

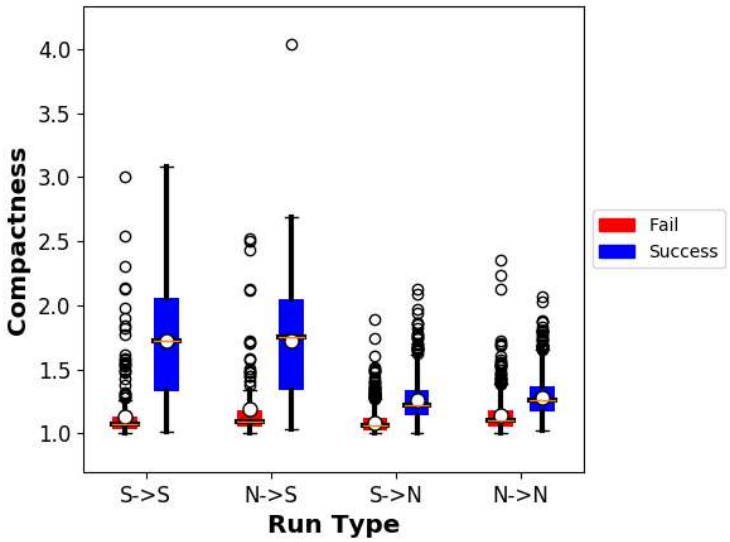


Fig. 4.32. Compactness, the number of introduced individuals per cell in cells containing at least one introduced individual, for successful (blue) and failed (red) establishment attempts, grouped by run type. Horizontal line in interquartile range represents (IQR) median, white dot represents mean. Whiskers extend to 1.5xIQR, white dots outside whiskers are outliers. Compactness was the best predictor for short-term establishment success. Successfully-established introduced populations from EcoSim Niches exhibited significantly greater compactness than those from standard EcoSim, when compared within each introduced environment.

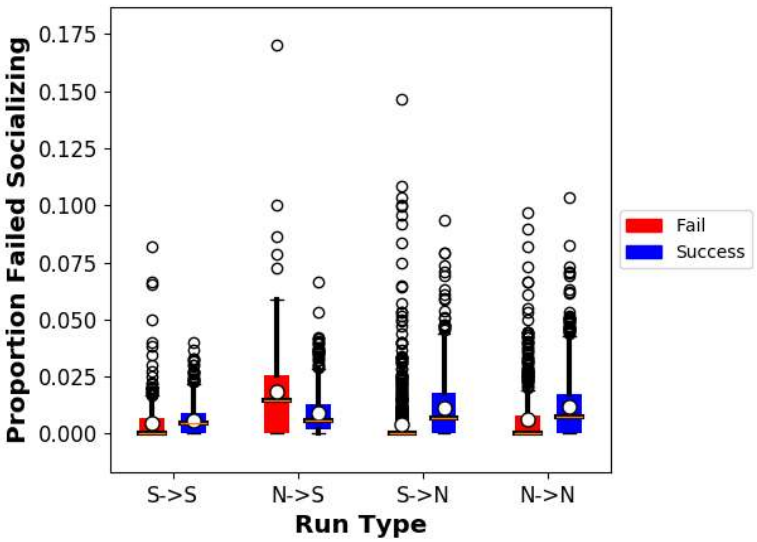


Fig. 4.33. Proportion of individuals failing to socialize for successful (blue) and failed (red) establishment attempts, grouped by run type. Horizontal line in interquartile range represents (IQR) median, white dot represents mean. Whiskers extend to 1.5xIQR, white dots outside whiskers are outliers. This feature was ranked second for prediction of short-term establishment success. Proportion of successfully-established introduced populations failing to socialize was significantly lower in S→S compared to all other run types.

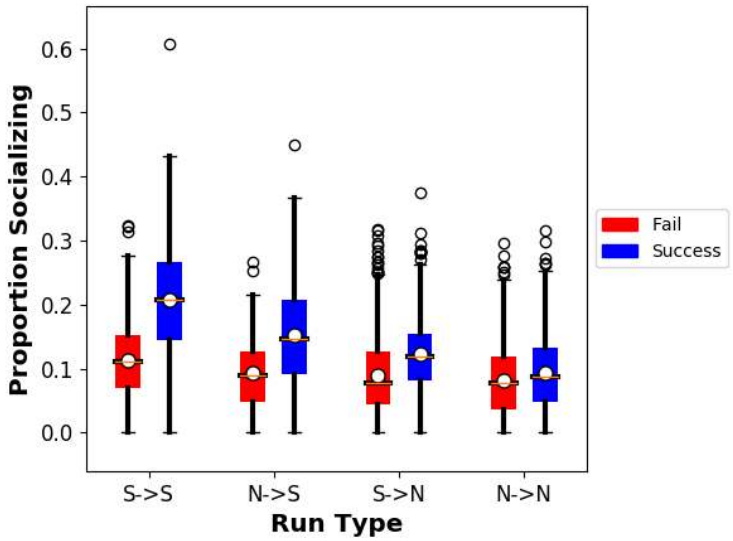


Fig. 4.34. Proportion of individuals socializing, for successful (blue) and failed (red) establishment attempts, grouped by run type. Horizontal line in interquartile range represents (IQR) median, white dot represents mean. Whiskers extend to 1.5xIQR, white dots outside whiskers are outliers. This feature was ranked third for prediction of short-term establishment success. In all run types, the proportion of individuals socializing was greater in successful establishments than failed. Successfully-established introduced populations from standard EcoSim exhibited significantly greater use of the socialize action compared to those from EcoSim Niches, when compared within each introduced environment.

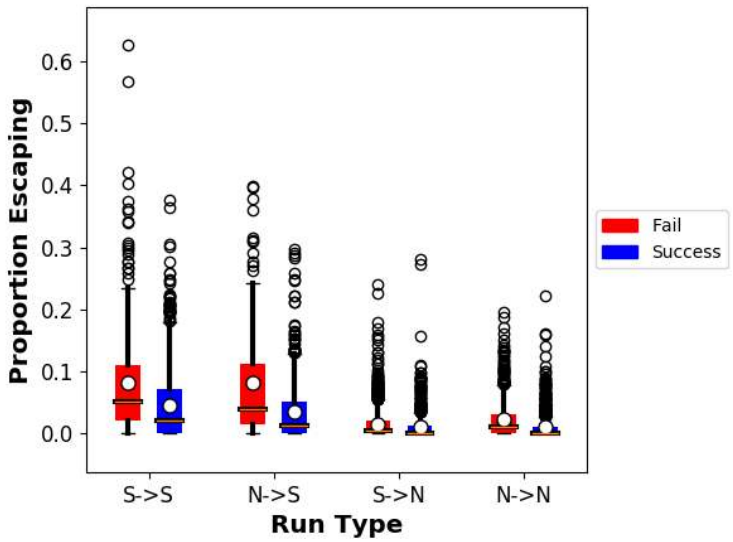


Fig. 4.35. Proportion of individuals escaping from predators, for successful (blue) and failed (red) establishment attempts, grouped by run type. Horizontal line in interquartile range represents (IQR) median, white dot represents mean. Whiskers extend to 1.5xIQR, white dots outside whiskers are outliers. This feature was ranked fourth for prediction of short-term establishment success. In all run types, the proportion of individuals escaping was greater in failed versus successful establishments. The escape action was used significantly more often in populations introduced to standard EcoSim than to EcoSim niches due to greater predator density.

Subsequent analysis of these features yielded several significant differences in distributions across genetic diversity levels for certain run types, except for S→S. For N→S, the proportion of individuals socializing showed significant differences over genetic diversity (Figure 4.36; p = 0.0015 on combined successful and failed attempts; Kruskal-Wallis) with L2 yielding significant comparisons against L3, L4, and L5 (p = 0.024, p = 0.0025, and p = 0.024 respectively; Conover's test with Holm correction). For S→N, compactness yielded significant differences across genetic diversity (Figure 4.37; p = 0.038 on establishment failures, p = 0.030 on combined successful and failed attempts; Kruskal-Wallis) with L1-L4 yielding a significant comparison (p = 0.016 on establishment failures, p = 0.015 on combined successful and failed attempts; Conover's test with Holm correction). In N→N, the proportion of individuals escaping from predators (p = 0.023 on failed attempts; Kruskal-Wallis) showed significant differences over genetic diversity with L1-L5 exhibiting significant difference (p = 0.026; Conover's test with Holm correction). The proportion of individuals socializing showed significant differences in distributions in combined successful and failed establishments (p = 0.028; Kruskal-Wallis) with significant comparisons between L1-L2 (p = 0.04; Conover's test with Holm correction). Similarly, compactness yielded significant differences over genetic diversity in combined failed and successful establishments (Figure 4.38; p = 0.027, Kruskal-Wallis), with comparisons between L1-L2 yielding weak significance (p = 0.08; Conover's test with Holm correction).

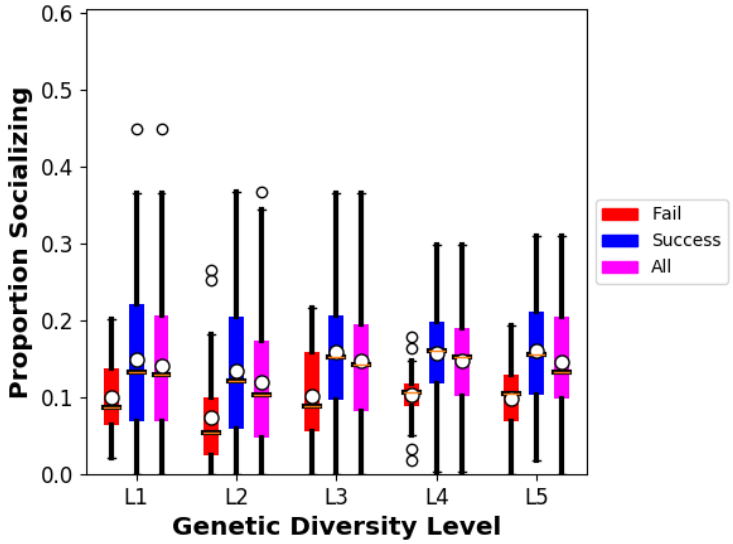


Fig. 4.36. Proportion of individuals socializing in N→S, for successful (blue), failed (red), and all (magenta) establishment attempts, grouped by genetic diversity. Horizontal line in interquartile range represents (IQR) median, white dot represents mean. Whiskers extend to 1.5xIQR, white dots outside whiskers are outliers. L2 proportion socializing was significantly different from L3, L4, and L5 for combined successful and failed establishments.

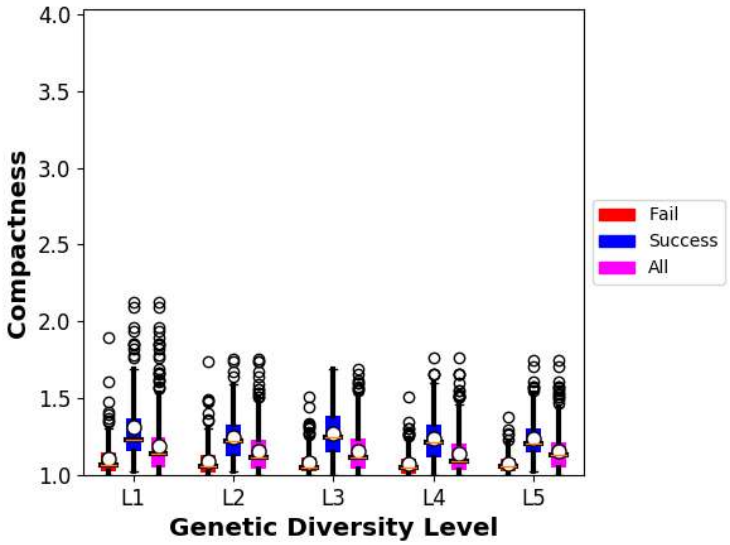


Fig. 4.37. Compactness, the number of introduced individuals per cell containing at least one introduced individual, for S→N successful (blue), failed (red), and all (magenta) establishment attempts, grouped by genetic diversity. Horizontal line in interquartile range represents (IQR) median, white dot represents mean. Whiskers extend to 1.5xIQR, white dots outside whiskers are outliers. Comparing L1 versus L4 overall and for failed establishments yielded significant difference in distribution.

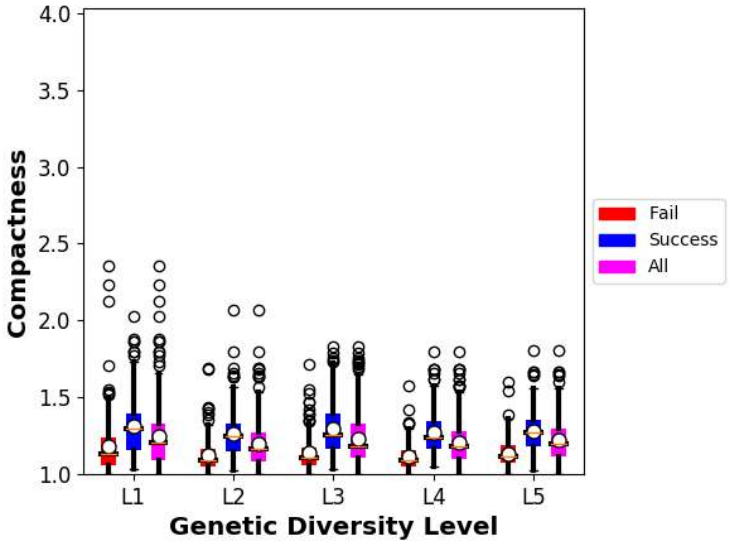


Fig. 4.38. Compactness, the number of introduced individuals per cell containing at least one introduced individual, for N→N successful (blue), failed (red), and all (magenta) establishment attempts, grouped by genetic diversity. Horizontal line in interquartile range represents (IQR) median, white dot represents mean. Whiskers extend to 1.5xIQR, white dots outside whiskers are outliers. Comparison between L1 and L2 yielded significant difference in combined successful and failed establishments (magenta).

4.4 Discussion

Hypothesis I stated that establishment success of introduced populations would increase with increasing genetic diversity of the inocula, with all other invasion parameters held constant. Hypothesis II, a corollary of hypothesis I, stated that this relationship would be stronger when the source and introduced ranges were vastly different. We found limited and circumstantial evidence of each hypothesis using our individual-based modelling approach. Short-term establishment in S→N and N→N showed significant differences in the extreme values of genetic diversity, with nearly monotonically-increasing establishment success with increasing genetic diversity (Figure 4.30). With S→S and N→S, however, such a relationship was not observed and instead inocula of extremely low diversity (i.e. inocula were formed of clonal populations) and medium-diversity (i.e. genetic entropy of ~210 bits) were favorable. These results were similar to those of Hufbauer *et al.* (2013), in which a similar experiment was conducted on whiteflies and differences in establishment success of introduced populations were attributable to genetic diversity more strongly when the populations were introduced to a "harsh" environment. Similar observations were made in a reciprocal transplant experiment by Szűcs *et al.* (2017) on red flour beetles, however in that study each transplanted population performed better in its natal environment. On the other hand, Szűcs *et al.* (2014) found no effect of genetic diversity on establishment success of red flour beetles in an experiment similar to that of Hubauer *et al.* (2013) and Szűcs *et al.* (2017) with a relatively less "harsh" environment, while demographics had an effect at low propagule size and propagule size was always extremely important. The relationship observed over genetic diversity in N→N supports hypothesis I, and the difference in relationship observed between S→S and S→N supports hypothesis II. However, the difference in N→S versus N→N directly contradicts hypothesis II. The observations in short-term establishment indicate that increasing genetic diversity of introduced populations aided more strongly in difficult (e.g. complex, heterogeneous, low-resource, as is EcoSim Niches) environments – rather than of those that are simply different from the origin of the introduced population. It has long been speculated that genetic diversity of introduced populations should positively influence their establishment success either as the difference between origin and novel environment increases, or as the complexity of the novel environment increases (Sakai *et al.* 2001; Bock *et al.* 2015; Dlugosch *et al.* 2015). We found evidence for the latter, and evidence directly against the former, at least in the earliest stages of establishment.

Subsequent analysis of the most impactful features in short-term establishment success mediated some potential insights regarding the above relationships. Introduced prey compactness is the number of introduced prey in a cell on average, for cells containing at least one introduced prey individual. Compactness was by far the most

important factor in determining short-term establishment success, based on our analysis using permutation importance on random forests. This was an expected result; the impact of Allee effects (positive density-dependent population growth at low abundance) in early invasions has long been studied in the invasion literature (Leung, Drake, and Lodge 2004; Taylor and Hastings 2005; Kanarek and Webb 2010). Compactness was much greater in successful populations introduced to standard EcoSim, owing to the increased abundance of resources overall (Figure 4.32). Both successful and failed introduced populations from EcoSim Niches exhibited greater compactness than those from standard EcoSim in each introduced range. Patterns for successful usage of the socialize action (Figure 4.34), were similar to those of compactness, however successfully-establishing introduced populations from standard EcoSim used the socialize action more often than those from EcoSim Niches in each environment. These results indicate that Allee effects are clearly important in the establishment of the sexually reproducing populations in EcoSim, and better invaders (e.g. those from EcoSim Niches) are better able to maintain compactness to guard against Allee effects in the earliest stages of an invasion. In S→N and N→N, compactness showed significant differences between L1 and other genetic diversity levels overall, and in both cases L1 was the genetic diversity level that fared the worst in short-term establishment (Figures 4.32 and 4.33). The significant difference between L1 and others is because introduced populations are entirely clonal, and clonal individuals in a common location (e.g. under a similar set of stimuli) should behave similarly as they are genetically predisposed to, outside of differences in their state (i.e. *Energy* levels, *Age*, etc.). This leads to generally stronger synchronization of action selection, which is beneficial in terms of combatting Allee effects (i.e. by synchronizing socialization or reproduction actions and therefore increasing reproduction efficiency) but also more costly in the face of extremely low resources. The result of this effect is the emergence of temporally cyclical compactness, especially in the earliest stages of establishment, which was especially pronounced in L1 individuals from standard EcoSim (S→S, Figure 4.39; S→N, Figure 4.40).

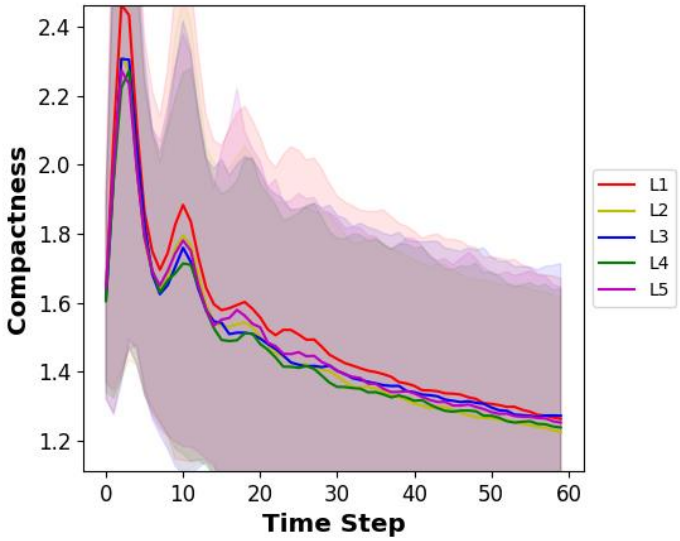


Fig. 4.39. Resultant synchronization of compactness due to synchronization of action selection in S→S introduced prey. Stronger synchronization in L1 is due to the populations being clonal. Individuals with similar genomes behave similarly in similar circumstances in EcoSim.

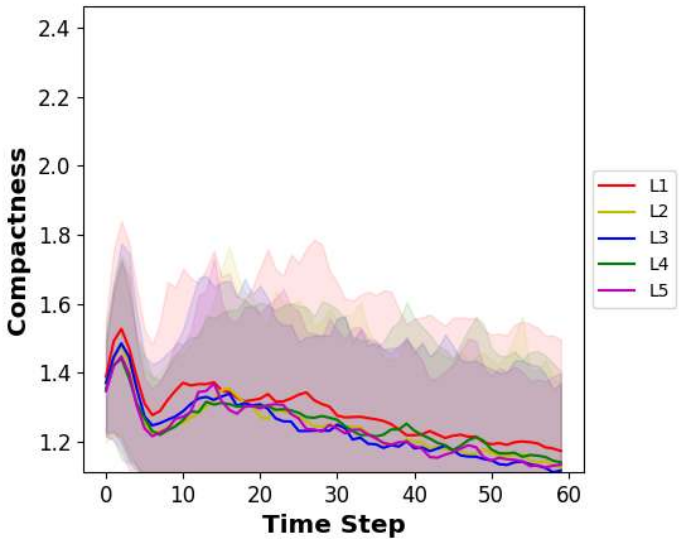


Fig. 4.40. Resultant synchronization of compactness due to synchronization of action selection in S→N introduced prey. Stronger synchronization in L1 is due to the populations being clonal. Individuals with similar genomes behave similarly in similar circumstances in EcoSim.

With respect to long-term establishment (Figure 4.31), dominance of extreme genetically-diverse inocula was apparent aside from in S→S, though the comparisons between genetic diversity levels yielded insignificant results because of low sample size. Observations regarding genetic diversity of established populations over time indicated that, at the very least, it is possible for consistent extremely low-diversity inocula from a

single source to produce an extremely high-diversity established population – even exceeding diversity levels observed in other established populations produced from inocula of significantly greater genetic diversity – while all other invasion parameters are held constant. Other studies (e.g. Kolbe *et al.* 2004; Dlugosch and Parker 2008) have demonstrated examples of multiple introductions from single or multiple sources, leading to genetic diversity levels in the introduced range that are unprecedented in any one native range. To our knowledge, however, this is the first example in which such a comparison was made with invasion parameters fixed and genetic diversity levels of individual inocula controlled. Though it was an interesting observation, it was not the purpose of this study and certainly more experimentation on this front is warranted. In the case of N→S, for example, genetic diversity of the established L1 population nearly reached that of L4 after 3500 time steps (i.e. 35 inoculations; Figure 4.29). Populations stemming from inocula of other genetic diversity levels did see some gain in diversity in some cases (e.g. L2 in all but S→N, L3 and L4 in S→S and N→S), but generally none received gains in diversity like L1. Of course, L1 stands to gain the most genetic diversity through repeated introductions (as every inoculation itself has a genetic diversity of zero, so we should expect genetic diversity of the established population to increase with any subsequent introduction), but that alone does not explain the ability of L1 populations to ultimately reach or exceed the diversity levels of the other populations (i.e. in which each independent inoculum was far more genetically diverse). Thus, an interesting extension to this study would be to investigate this phenomenon further to explicitly observe the role of genetic admixture via multiple introductions in the long-term evolution of the introduced populations. It is probable that genetic admixture affects different established populations in different ways in the long term; perhaps, in cases of multiple introductions, there exists a relationship between genetic diversity of inocula and the relative benefit they provide to the pre-existing established population.

On the other hand, perhaps, in multiple introductions, there is a relationship between genetic diversity of the established population and the relative benefit provided by each inoculum. It is possible that the resultant populations at 3500 time steps, stemming from inocula of lower diversity, contained alleles from a greater number of sources (i.e., genetic admixture contributed disproportionately to their ultimate gene pools). While we did not investigate this further in this study, it would certainly be an interesting subject of future investigation.

All measurements of establishment success (abundance, short-term success, and long-term success) exhibited differences between the different IR run types (i.e. S→S, N→S, S→N, N→N). With respect to abundance of introduced individuals over time, the difference between any pair of run types was approximately an order of magnitude in most cases. It was anticipated that introduced populations would perform best in the

environments from which they were taken (i.e., standard EcoSim individuals would perform better in standard EcoSim and EcoSim Niches individuals would perform better in EcoSim Niches), but this was not observed. Instead, an asymmetrical relationship was observed in which EcoSim Niches individuals performed better than standard EcoSim individuals in both standard EcoSim and EcoSim Niches. Such asymmetrical relationships are observed in nature. Fridley and Sax (2014) highlighted asymmetry in reciprocal number of invaders between the Red Sea and the Mediterranean Sea as well as between Lake Ontario and the Hudson River. They proposed the evolutionary imbalance hypothesis, which stated that regions of higher diversity should not only produce better invaders – and more of them, even after accounting for differences in species richness – but that they should also be less invasible. They theorized that the biodiversity in a region is representative of the time that evolution has had to take its course on it; thus, greater diversity in a region is a manifestation of a greater number of "evolutionary experiments". According to Fridley and Sax, one consequence is that the species inhabiting diverse regions will be more optimized to their respective niches. Another consequence is that there would be more genetic diversity not only among species but also within them (i.e. phylogenetic diversity), and this was anticipated to aid in invasion success as well. In EcoSim Niches, there was greater genetic diversity compared to standard EcoSim; behavioural genes exhibited more variance and there was a greater ratio of species richness to abundance for predators and prey. However, both run types were given the same amount of evolutionary time before reciprocal transplants occurred. Thus, our study corroborates the evolutionary imbalance hypothesis in observed outcome, but not necessarily in all theorized explanations (i.e., diversity did not indicate evolutionary time).

Mechanically, among other factors, Fridley and Sax (2014) discussed the relative intensity of competitive environments in which the populations evolved as a potential explanation for such imbalance. Based on data presented in 4.3.1 and Appendix A, we conclude that standard EcoSim and EcoSim Niches produced significantly different environments and consequently the individuals produced by the two variants were also different behaviourally, physically, and in terms of measured properties that emerged due to their differences (e.g. *Speed*, *Energy*, energy spent, species richness). Regarding physical characteristics of the prey and predator individuals, in all cases where significant differences were observed (Figures 4.18, 4.19, 4.21, and 4.22), the magnitude of the property was larger in the individuals from standard EcoSim than in EcoSim Niches. As discussed in Section 3.2.2 under *Individuals*, all physical characteristics carry an energy cost that is exacted every time step, and this cost increases nonlinearly with increasing magnitude of each physical characteristic. These reduced physical characteristics in the prey from EcoSim Niches contributed to an efficiency advantage (Figure 4.16).

As previously mentioned, EcoSim Niches features a reduction in resource abundance compared to standard EcoSim (in addition to highly variable resource abundance per cell). This presents a challenge for prey individuals as they must compete amongst each other for sometimes minimal resources. Compactness of introduced prey from EcoSim Niches was greater than that of introduced prey from standard EcoSim in each environment (Figure 4.32), matching what was exhibited by the native populations in their respect native ranges. The benefit of maintaining high compactness is maintaining the ability to find a mating partner that is willing and able to reproduce, especially at low population densities as observed in EcoSim Niches. However, the cost to maintaining high compactness is increased intraspecific competition. The efficiency advantage due to reduced physical capacities in individuals from EcoSim Niches, noted above, reduced the cost associated with maintaining high compactness. This likely yielded a competitive advantage for populations from EcoSim Niches in the early stages of establishment, as they were able to effectively reduce the cost of guarding against Allee effects. Thus, our study provides corroboration for the mechanical explanation of the evolutionary imbalance hypothesis that environments with intense competition simultaneously yield better invaders and are less invasible themselves.

A valid follow-up question is "why wouldn't standard prey simply evolve to be like EcoSim Niches prey if their strategy is so advantageous in the standard EcoSim environment?" A potential explanation is that evolving the general strategies adopted by EcoSim Niches prey – which seem to align more with r-selection than k-selection – may require the traversal of a low-fitness region of the evolutionary fitness landscape in the general standard EcoSim environment that is not present in the general EcoSim Niches environment. That is, genome *z* (producing the strategy most common in EcoSim Niches) might be highly favourable in both standard EcoSim and EcoSim Niches, but in order to reach it from genome *x* (e.g. the initial genome of all prey), it would be necessary to cross genome *y*, which is favourable at some point in EcoSim Niches but not in standard EcoSim.

There are some clear limitations to this work. Most importantly, this was a simulation study and although we showed that standard EcoSim and EcoSim Niches produced vastly different environments and individuals, it was not possible for us to say just how different they were. Real invasions involve the passing of introduced populations through a variety of potentially strenuous filters (Kolar and Lodge 2001; Colautti and MacIsaac 2004; Blackburn *et al.* 2011); in the novel territory, the first filter is the environment. In reality, introduced species need to be able to cope with temperature, chemistry, and a wide variety of other physical factors before establishment is even a possibility, and so our assumption here is that our introduced species are at least physiologically compatible with the novel environment and this could certainly explain

the difference between what we observed and what the theory predicts. Going forward, we could conduct a similar experiment but impose some efficiency penalty (either directly to the *Energy* expenditure function, or perhaps as a tax on consumption) for the introduced species to the dissimilar environment, so as to model reduced climate-matching (e.g. suboptimal temperature or food resource composition). We could also extend this experiment to explore environment harshness (i.e. in our case the degree of spatially-dependent variation in resource abundance as another dimension), but this also creates a logistical problem due to computational constraints as this experiment already required enormous computational resources.

On the other hand, though we were able to control for genetic diversity of the introduced populations, it was impossible to know ahead of experimentation whether the range of genetic diversity we produced (i.e. from genetic entropy of zero to ~420 bits) would be enough to mediate different responses during establishment. We did observe differences in both short-term and long-term establishment, but were we able to increase the genetic diversity of the inoculations further (e.g. genetic entropy of 1000 bits), we may have observed a stronger response. On a related note – and this is a classical limitation in studies of this nature (Roman and Darling 2007; Wellband *et al.* 2018) – though we were able to control the degree of genetic diversity of the inocula, we cannot possibly know the degree to which the genetic diversity of the inocula resulted in functional diversity. It is practically impossible to quantify, even in our simulation, the degree to which diversity in genotypes yields diversity in phenotypes.

Another limitation was in the analysis of factors contributing to successful versus failed short-term establishment; as our inoculations occurred every 100 time-steps we were unable to analyze inoculations independently in cases where previous inocula remained intact in the introduced range, which is why we analyzed only "fresh" inoculations. The result was that once an introduced population established long-term, we were unable to analyze subsequent inoculations in this manner. This certainly led to a reduced sample size for that analysis; a better design for such analysis would be in which we performed every subsequent inoculation only when the previous inoculation ceased to exist, or created separate EcoSim simulations for independent inocula. Though, again, these designs bring logistical concerns in terms of our computational constraints. Lastly, another limitation was that we did only use ten simulations per treatment, again because of computational constraints. Thus, we were unable to say definitively that there was a relationship between long-term establishment and genetic diversity of the inocula; however, our study still provides sufficient reason to explore this further. This also has potential implications in our assessment of short-term establishment; we treated each "fresh" inoculum as an independent case though there may be a temporal dependency in the available genotypes from which we sampled the inocula (i.e., only ten simulations).

However, both the native range and novel range of the introduced individuals represent highly dynamic and evolving systems, thus we assumed each inoculum was independent and believe this treatment was valid here.

In conclusion, assuming that the introduced populations are physiologically capable of surviving in the novel environment, we found a positive relationship between genetic diversity and short-term establishment success when populations were introduced to a harsher environment with spatially-varying and lower resource abundance, but this relationship did not hold when the novel environment had uniform resource distribution. We also found evidence that genetic diversity of inocula affects long-term establishment of introduced populations, assuming multiple introductions from a single source, but that extreme diversity levels may be favourable for long-term establishment. Further, our study corroborates the evolutionary imbalance hypothesis and a mechanical explanation for it: introduced populations originating from EcoSim Niches outperformed those from standard EcoSim in both environments, and the difference in intensity of competition in the native ranges may explain why. Allee effects were found to be extremely important in the earliest stages of establishment as compactness was by far the most important factor in determining establishment success. Further, observed differences in compactness between introduced individuals and those in their respective native ranges, differences in costs associated with maintaining physical characteristics, and the associated differences in establishment success, highlighted the importance of reducing costs associated with guarding against Allee effects for establishing populations (i.e. increased intraspecific competition). We only transferred prey individuals in our simulations; it is possible that species of higher trophic levels may exhibit differences from what we observed here, and this warrants further research. Also, many real introductions have involved mixed communities (i.e. containing numerous species across different trophic levels; e.g. introductions via ballast water); it is also possible that different patterns may be observed when mixed communities are transferred. Future studies could make comparisons of the evolutionary trajectories taken by introduced populations receiving multiple introductions and those not. For us, this would require a set of EcoSim simulations in which admixture does not occur, and each simulated invasion is carried out over a long term such that we could observe evolution of each independent introduced population. Similarly, with multiple introductions, the degree to which admixture occurs in the introduced populations may be related to the genetic diversity of the inocula or in the established population, and this warrants further experimentation. Lastly, as called for by Dlugosch *et al.* (2015), admixture from multiple introductions could lead to different fitness responses based on genetic diversity of both the established populations and the inocula and further investigation is necessary on this front.

CHAPTER 5

Optimization and Performance Testing of a Sequence Processing Pipeline Applied to Detection of Nonindigenous Species

5.1 Introduction

Newly introduced populations that colonize novel ecosystems are usually small and inconspicuous (Leung, Drake, and Lodge 2004). Detection of small and geographically restricted populations is technically challenging, yet critically important to management of aquatic invasive species (AIS; Beric and MacIsaac 2015). Traditional early detection relies on techniques such as recruitment plates, video, scuba diving, trawling, and netting – which may require tremendous amounts of sampling effort (Hoffman *et al.* 2011) – typically followed by morphological identification. Furthermore, they may be ineffective if the introduced species is small, cryptic, or morphologically variable (Ficetola *et al.* 2008). These attributes characterize many AIS, rendering monitoring of underwater environments an especially challenging task. Generally, genetic approaches are promising in the early detection of AIS, circumventing numerous challenges of traditional surveillance (Smart *et al.* 2015).

When applied to complex communities, genetic detection of AIS or characterization of species composition typically involves sampling whole organisms (bulk sampling) or environmental DNA (eDNA) shed by them. In either case, a small 'barcode' region of the genome (Hebert *et al.* 2003) can be used to determine the taxonomic identity of mixed sequences (Cristescu 2014). There are two genetic approaches to detection of AIS. In the first, one must have a particular target (typically, a species) in mind (the "targeted" or "active" approach). Alternatively, metazoan metabarcoding (Fonseca *et al.* 2010) aims to recover a wide range of taxa in a community and passively discover AIS (the "passive" approach; Simmons *et al.* 2016). Metazoan metabarcoding typically involves the use of universal primers and PCR to amplify available genetic material aiming to recover all taxa from the captured sample. However, in reality, not all taxa are discovered with equal sensitivity due to primer design or choice, and consequently inconsistent amplification may occur (Creer *et al.* 2010; Xiong, Li, and Zhan 2016).

The metabarcoding process begins with a bulk sample, which often involves the use of specific nets to capture targets. Genomic DNA is then extracted and amplified using primers that are specifically designed or selected for the study. The amplified DNA is then sequenced, and once the sequences are obtained this data can be subjected to computational processing that might involve processes like filtration, denoising, or clustering. Processed sequences are then run against reference databases to determine

their taxonomic identity. Owing to the complex process of metabarcoding metazoan bulk samples (Figure 5.1, applied to detection of AIS), many potential sources of both false positive (type I) and false negative (type II) errors have been identified. A non-exhaustive list of potential sources of errors in this process includes primer design (Freeland 2017), PCR (Piggott 2016), next-generation sequencing (Fonseca *et al.* 2010), sequence processing (Flynn *et al.* 2015), reference library preparation (Zhan, He, *et al.* 2014), and taxonomic assignment inconsistencies, though it is difficult to quantify the impact of each (Xiong, Li, and Zhan 2016). Fortunately, by appropriately selecting parameters in computational sequence processing, the impact and frequency of errors can be reduced (Zhan, Xiong, *et al.* 2014; Brown *et al.* 2015; Flynn *et al.* 2015).

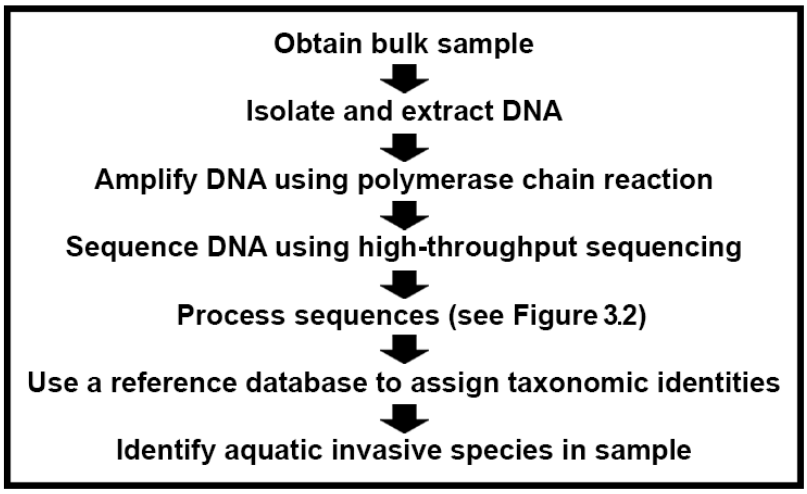


Fig. 5.1. Flowchart of the general metazoan metabarcoding process applied to bulk sampling in the context of aquatic invasions. In this study, we focus on the computational aspects of the process (sequence processing, BLAST, and identification of AIS).

Over the last decade, several sequence processing suites have been developed, including USEARCH (Edgar 2010), mothur (Schloss *et al.* 2009), QIIME (Caporaso *et al.* 2010), and RDP (Cole *et al.* 2014), each making simplifying assumptions that improve computational efficiency. Many of these suites share features, algorithms, or even programs. Intra-specific genetic variation within barcode regions can exist, so many programs allow users to cluster sequences into operational taxonomic units (OTUs) based upon genetic similarity (Schloss *et al.* 2009; Edgar 2013). OTUs are groups of sequences that share high similarity, typically at the species or genus level. UPARSE, which is built into the USEARCH program, can create clusters in order of decreasing sequence abundance after sequence dereplication (Edgar 2013). Although the most abundant sequence may not represent the true center of a species, this approach is computationally efficient and is more effective than other approaches (such as UCLUST or hierarchical clustering of mothur; Edgar 2013; Flynn *et al.* 2015). Other approaches to clustering – such as Bayesian (Hao *et al.* 2011), modularity-based (Wang *et al.* 2013), and

agglomerative clustering (Mahé *et al.* 2014) – may use different sequence identity definitions; that is, they penalize gaps in alignments differently. Several of these sequence processing suites have similar or shared features and algorithms, for example the clustering algorithms in QIIME are strictly third-party and some are closed-source (Caporaso *et al.* 2010). USEARCH is comprehensive and allows sequence trimming, minimum Phred score (Q) filtering, maximum expected error (MEE) filtering, clustering, denoising (Edgar 2016), and removal of sequences not meeting any arbitrary abundance threshold. These are all options that are regularly used in the related literature in some capacity, even in computational suites other than USEARCH (Bokulich *et al.* 2013; Pawlowski *et al.* 2014; Elbrecht and Leese 2015; Flynn *et al.* 2015; Brown *et al.* 2015; Brown *et al.* 2016; Chain *et al.* 2016; Hänfling *et al.* 2016; Port *et al.* 2016; Bista *et al.* 2017). USEARCH also has many other utilities for analysis after sequences have been processed, suh as computation of diversity indices and phylogenetic analysis.

The objective of sequence processing is to improve the integrity of results, but it may also be a source of error if performed poorly (Brown *et al.* 2015; Flynn *et al.* 2015; Xiong, Li, and Zhan 2016). Parameter selection in sequence processing involves a delicate balance between false positive and false negative error (Zhan *et al.* 2013). With overly stringent quality filtration, for example, sequences that identify truly present taxa in a sample may be removed, leading one to incorrectly infer absence of these taxa (false negative error). On the other hand, insufficient filtration can lead to false positive errors, because in downstream analyses, erroneous sequences could map to species not present in the sample. Filtering is discussed here for illustrative purposes; all other components of the pipeline (clustering, denoising, length cutoffs, abundance thresholds, *etc.*) similarly participate in this balance between false positives and false negatives and thus parameter selection is not straightforward. The optimal parameter sets (which minimize either or both types of error) depend on the aim of the study and are usually not known prior to processing. Currently, users have limited knowledge on which to base parameter selection.

Though computational processing of sequences is an essential part of taxonomic assignment for genetic sequence data, very few studies have attempted to rigorously address the problem of parameter selection (*e.g.* Bokulich *et al.* 2013). Instead, few (or single) aspects of sequence processing have been previously tested – often with low resolution (*i.e.* Pawlowski *et al.* 2014; Brown *et al.* 2015; Flynn *et al.* 2015; Brown *et al.* 2016) – though numerous processing steps and parameter values interact to produce the resultant set of sequences or OTUs. Parameter selection also depends on the goals and methods of the study (identification of AIS, species richness estimation, eDNA, bulk sampling, *etc.*). Thus, there is a need to test a wide range of processing steps and parameter values in concert and for different research scenarios.

We primarily sought to determine how users should select parameters when using a sequence processing pipeline (Figure 5.2) in a metazoan, bulk-sample, metabarcoding context. Simultaneously, we wanted to determine if and how research goals influence optimal parameter selection. Finally, we aimed to determine the performance of such a pipeline when parameters were appropriately selected given these research goals. Consequently, this study had two main investigations: *Optimization*, in which we searched for optimal parameter selection for the computational sequence processing pipeline, and *Performance Testing*, in which we performed simulations to assess the performance of selected 'most optimal parameter sets' in two ways: sensitivity and detectability (defined below under *Performance Testing*). In both parts of the study, we considered two common research applications of metabarcoding: accurate estimation of species richness and early detection of AIS. These research goals differ in how researchers will utilize sequence processing pipelines to shift the balance between protection against false positives and false negatives. Though it is always important to control for both types of errors, researchers estimating species richness via metabarcoding are typically concerned with minimizing both false positives and false negatives, while those involved in early detection of AIS are mainly concerned with minimizing false negatives.

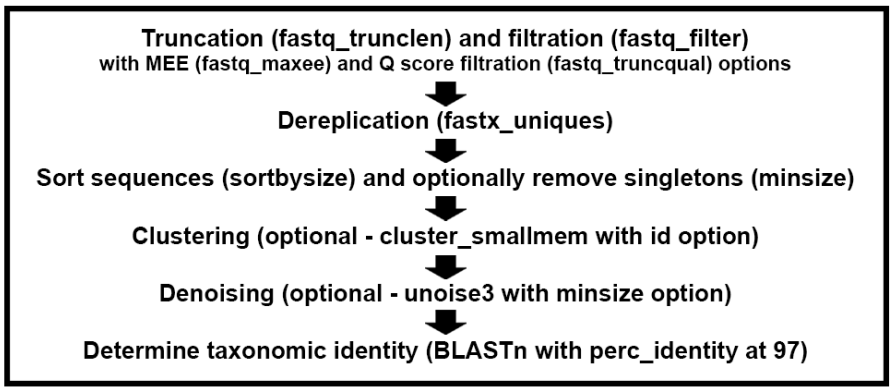


Fig. 5.2. Flowchart of the sequence processing pipeline used in this study. Relevant USEARCH commands and options used are shown in parentheses. The first step combines sequence trimming (truncation) and quality filtration (Phred score – Q, and maximum expected error – MEE). In the next step, sequences are dereplicated. Next, the sequences are sorted in terms of decreasing abundance (necessary for clustering and denoising) and singletons are removed. Clustering or denoising of the sequences may subsequently be performed. Finally, BLASTn is used to perform taxonomic assignment with a minimum identity threshold of 97% using BLASTn defaults.

5.2 Materials and Methods

Below, we give a brief overview of our study. We then describe our sequence processing pipeline, introduce our sequence datasets, explain the optimization process, and discuss our performance testing procedure.

First, we optimized a sequence processing and taxonomic assignment pipeline employing USEARCH v10.0.240_i86linux32 and BLASTn v2.6.0+ (Figure 5.2) using a mock (*i.e.* deliberately assembled) community of sequences from 20 AIS obtained via 454 pyrosequencing. We used the USEARCH package because it is comprehensive, fully automatable through scripting, and exhibits strong performance and efficiency (Edgar 2013). We optimized the pipeline separately for two common research goals: accurate estimation of species richness (which favors minimizing false negatives and false positives when sequences vary in abundance) and early detection of AIS (which favors sensitivity and minimizing false negatives, even for sequences of low abundance). This stage involved a search for parameter sets that generated OTUs that most accurately reflected the makeup of the mock community samples, which we described in detail below under the section "Optimization". Secondly, we took some of the most optimal parameter sets from the optimization phase and tested their performance through simulation. We tested performance using 20 different AIS, community samples from 10 ports, and the most effective 24 parameter sets (of 1050 total parameter sets tested), allowing us to observe dependencies between these factors. This allowed us to make recommendations for sequence processing parameter selection from a more general standpoint.

5.2.1 Sequence Processing

We defined a parameter set as a combination of sequence length, Q filter stringency, MEE filter stringency, clustering identity threshold (if clustering was used), denoising minimum sequence abundance (if denoising was used), and minimum sequence abundance after dereplication. The values we tested for each parameter can be found under *Optimization*. To elaborate, sequences shorter than the sequence length threshold were removed, while those longer than that length were trimmed accordingly. The Q filter we used was a minimum Q score filter, meaning that a sequence with any single base call with Q below the threshold was removed. The MEE filter computed the maximum number of expected errors across the entire sequence using Q scores of each base call. Sequences with an expected number of errors above the MEE threshold were removed. Clustering identity was the similarity threshold between an OTU's representative sequence and all other sequences in that OTU using UPARSE. Denoising in USEARCH (UNOISE3) considered sequence abundance and number of differences between sequences to predict whether a sequence was correct or not (Edgar 2016). In UNOISE3, the probability of incorrectness of a sequence was computed based on the abundance skew ratio (ratio of abundance) and number of differences between it and other sequences already deemed correct, and sequences were compared in order of decreasing abundance for efficiency (see Edgar 2016 for algorithm details). Denoising minimum abundance was the minimum abundance for a sequence to not be considered

noise, which also affected abundance skew ratio ratios for retained sequences. With any given denoising minimum abundance, retained (but noisy) low-abundance sequences counted towards abundances of their 'correct' counterparts. This could impact classification of sequences at the denoising step and could also influence downstream abundance-based analyses. Further, as the lower limit on sequence abundance was increased, remaining sequences could be classified as noisy or correct with greater confidence with the UNOISE3 algorithm (Edgar 2016). We left the other clustering and denoising parameters to their default values. Minimum sequence abundance after dereplication was simplified by either allowing or removing singletons.

We used the same sequence processing procedure in both optimization and performance testing (Figure 5.2). We used USEARCH for all sequence processing. This procedure took as input a single FASTQ file, though it could also be adapted for merged paired reads. In the first step, we truncated sequences, removed those not meeting the length requirement, and then filtered the sequences by quality. Next, we dereplicated and sorted sequences by abundance, which was necessary for the UPARSE clustering and UNOISE3 denoising algorithms built into USEARCH. In this step, if singletons were to be removed, only sequences with two or more replicates were retained. Whether clustering or denoising was performed or not was determined by the parameter set being tested (*i.e.*, the iteration of the optimization stage or the selected parameter sets in performance testing). We did not test combining clustering and denoising due to computational constraints. A chimera detection algorithm is embedded in the denoising algorithm of USEARCH that we used (UNOISE3), so chimera detection occurred if denoising was performed using the defaults for UNOISE3. Once sequence processing was complete, we checked the resultant set of sequences (or OTU representative sequences) against precomputed BLAST results (see Dataset Preparation below for BLAST precomputing procedure). All computing was performed on the Shared Hierarchical Academic Research Computing Network (SHARCNET).

5.2.2 Dataset Preparation

We acquired four published metabarcoding datasets of 18S V4 rDNA sequences. The amplified fragment length was ≥ 400bp for our target taxa. Primers for this marker effectively amplify a broad range of zooplankton taxa, making 18S a suitable marker for zooplankton metabarcoding studies (Zhan, Bailey *et al.* 2014). Conversely, the COI marker is highly variable for these taxa (sometimes, even in the primer binding sites) which may make it more suitable for studies taking the targeted approach than for metabarcoding highly divergent communities (Deagle *et al.* 2014; Zhan, Bailey *et al.* 2014; Hatzenbuhler *et al.* 2017). The drawback of 18S is that due to lower variability it

may be more difficult to assign identity at the species level. For each dataset, we obtained unprocessed sequences in FASTQ format.

The first dataset, which we called D1, was a mock community of 20 AIS obtained from bulk zooplankton samples. All sequences of a given taxon were identically tagged by adding short sequences in the primers, unique to each taxon (Brown *et al.* 2015). This dataset was referred to as the "Tagged individual community" in the paper by Brown *et al.* (2015). The dataset originally contained 115902 sequences (unevenly distributed across the 20 taxa), with sequence length of approximately 400-600bp. This library was amplified using a primer pair developed by Zhan *et al.* (2013) and pyrosequenced using 454 GS-FLX Titanium platform (454 Life Sciences, Branford, CT, USA) by Genome Quebec (see Brown *et al.* 2015 for more details of library production). We removed all sequences of the invaders *Dreissena polymorpha* and *Ciona intestinalis* because preliminary analyses indicated that these samples were likely contaminated. We BLASTed all sequences of this dataset and found that roughly half of those from *D. polymorpha* and *C. intestinalis* aligned best with different taxa also found in this dataset. However, we also acquired 18S sequences of two other AIS: the green crab *Carcinus maenus*, which is a marine AIS of global importance, and the quagga mussel *Dreissena rostriformis bugensis*, which is a major problem in lakes in Europe and North America. We obtained green crab sequences from Brown *et al.* (2015), while those of quagga mussels were detected in bulk zooplankton samples (Chain *et al.* 2016). Both research groups used the same library production protocol and sequencing platform as described above, though the latter used a primer designed by Zhan, Bailey, *et al.* (2014). We refer to the dataset consisting of the sequences from the 18 taxa from the mock community, plus green crab and quagga mussel sequences, as D1. Therefore, D1 consisted of different abundances of sequences from 20 taxa with varied relatedness. This dataset was used for both the optimization and performance testing stages. For optimization and performance testing, we separated this dataset into 20 separate sequence sets, each consisting of sequences from a single taxon (Table 5.1). The amplified fragment for all 20 taxa was ≥ 400bp, and mean sequence length was 466bp.

Table 5.1: Dataset D1, with sequences grouped by taxon. Proportion of sequences kept at length 350 bp given a Phred score (Q) filter or MEE filter of varying strengths are shown as a proxy of dataset quality. Sequences ranged greatly in quality and abundance, with Brachionus and Mesocyclops yielding sequences of lowest quality. With a Phred score filter of 20, no sequences of Brachionus or Mesocyclops were retained.

Taxon	Sequences	Q = 10	Q = 20	MEE = 1
Artemia salina	2145	0.9920	0.0490	0.8015
Balanus crenatus	14732	0.9910	0.1310	0.8128
Brachionus calyciflorus	207	0.9950	0.0000	0.0483
Cancer sp.	1629	0.9940	0.1040	0.7185
Carcinus maenas	200	1.0000	0.1750	0.9400
Cercopagis pengoi	1222	0.9920	0.0110	0.7709
Corbicula fluminea	46915	0.9900	0.2980	0.8952

Daphnia mendotae	706	0.9750	0.0160	0.6232
Diacyclops thomasi	812	0.9900	0.0090	0.7106
Dreissena rostriformis bugensis	200	1.0000	0.1550	0.9450
Echinogammarus ischnus	7337	0.9820	0.2430	0.8327
Epischura lacustris	10002	0.9900	0.1400	0.8465
Leptodiaptomus ashlandi	5461	0.9890	0.0790	0.7539
Mesocyclops edax	1055	0.9910	0.0000	0.2812
Microsetella norvegica	814	0.9950	0.0530	0.8136
Oikopleura labradoriensis	3545	0.9940	0.1090	0.8434
Palaemonetes sp.	5170	0.9930	0.3630	0.9154
Pleuroxus denticulatus	644	0.9800	0.0080	0.6182
Senecella calanoides	348	0.9970	0.0140	0.4580
Themisto libellula	4269	0.9830	0.5000	0.9311

In performance testing, we also utilized a dataset that consisted of V4 18S rDNA derived from bulk zooplankton samples from ten Canadian ports (Chain *et al.* 2016). We kept each of these samples separated by port, and refer to this as D2 (Table 5.2). Sequences of D1 were computationally inoculated into samples from D2, as explained in more detail below under "Performance Testing". Primers and tags were removed from all sequences. In cases where, after sequencing, the primer or tag of a sequence did not match any original primers or tags, the sequence was removed.

Table 5.2: Dataset D2, containing sequences of ten Canadian ports sampled (see Chain *et al.* 2016). Number of sequences and proportion of sequences kept at length 350 bp given a Phred score (Q) filter of 10 and 20 or MEE filter of 1 are shown. Samples ranged greatly in quality and abundance. Churchill and Halifax yielded sequences of relatively low quality, whereas Hawkesbury, Sept Iles, and Thunder Bay yielded sequences of relatively high quality. With a Phred score filter of 20, no sequences of Churchill or Halifax are retained.

Location	Sequences	Q = 10	Q = 20	MEE = 1
Churchill	684163	0.2290	0.0000	0.0809
Halifax	877078	0.2480	0.0000	0.0477
Hamilton	686064	0.2660	0.0230	0.1750
Hawkesbury	444315	0.6370	0.1110	0.5076
Nanaimo	406215	0.6240	0.0200	0.4074
Nanticoke	480962	0.5820	0.0570	0.4305
Sept Iles	249663	0.9550	0.1900	0.8645
Thunder Bay	556984	0.6910	0.1170	0.5798
Vancouver	1008358	0.2670	0.0020	0.1359
Victoria	456391	0.5720	0.0310	0.3976

For optimization and performance testing, we needed to classify each sequence in D1 as correct, ambiguous, or incorrect. A correct sequence was one that aligned best with a reference sequence of its true identity, with identity $\geq 97\%$, whether alignments to other taxa were tied in similarity score or not. An ambiguous sequence was one that aligned

with a higher score to a reference sequence of a different taxon, though it still aligned to its correct taxon with identity ≥ 97%. An incorrect sequence aligned with a reference sequence of its true identity with identity < 97%.

When the pipeline was performed on a sample with a given parameter set, a set of OTUs was generated each of which had a representative sequence. These representative sequences were run against a reference database using an alignment search tool to determine their taxonomic identity. Basic Local Alignment Search Tool (BLAST - Altschul *et al.* 1990) is one such computational tool. We used the NCBI nucleotide database as our reference (retrieved June 2017) and BLASTn (BLAST for nucleotide sequences). We precomputed the class of each sequence so we could efficiently classify each OTU generated in optimization and performance testing based on its representative sequence. For optimization, this was necessary to evaluate the quality of each parameter set based on the OTUs it produced from the optimization samples. For performance testing, this was necessary to determine if an inoculated taxon could be correctly recovered from a sample.

We computed the list of all BLAST hits with ≥ 97% identity, which in BLASTn were sorted by decreasing hit similarity using the metrics E-value, bit-score, and identity. However, any number of hits may have had identical similarity scores using these metrics. Moreover, sequences (especially if they contained errors) may share more similarity with sequences of a different species than those of their own. Each of these situations made accurate identification of a sequence challenging. Worse yet, a sequence may not have aligned with sequences of its own species with sufficiently high alignment score, or the reference database may not have contained sequences of the queried species. Considering these complications, we classified each sequence in D1 as correct, incorrect, or ambiguous with the following definitions. A correct sequence was one that aligned best with a reference sequence of its true identity, with identity ≥ 97%, whether alignments to other taxa were tied in similarity score or not. An ambiguous sequence was one that aligned with a higher score to a reference sequence of a different taxon, though it still aligned to its correct taxon with identity ≥ 97%. An incorrect sequence aligned with a reference sequence of its true identity with identity < 97%.

To classify each sequence, we first had to establish a ground truth BLAST identity for each taxon using our reference database (Table 5.3). Of the 20 taxa, 11 BLAST identities matched their corresponding morphological identities to species and five matched to genus. Of the remainders, two taxa were assigned generic 18S metazoan identities, and two were assigned different identities altogether compared to their morphological identities. All samples in D1 obtained from Brown *et al.* (2016), where a sample is a set of sequences from a single taxon, were from specimens morphologically identified in that study. For these samples, if the majority of sequences aligned with

reference sequences of their morphological identity with ≥ 97% identity, the morphological identity was assumed correct. Otherwise, the taxon with the highest BLAST similarity score was assumed correct. In most cases, BLAST identity of sequences matched their morphological identity, though in some cases they did not, mainly because the morphological identity did not exist in the reference database. Sequences of *Dreissena* and *Carcinus* were identified through BLAST in Chain *et al.* (2016).

Table 5.3: Morphological and assumed BLAST identities of sequences from dataset D1, separated by taxon. Correct sequences were those that BLASTed to the assumed identity with rank 1 (using the BLASTn default sort method) and identity > 97%. Ambiguous sequences were those that BLASTed to the assumed identity with rank > 1 and identity > 97%. Incorrect sequences were those that did not BLAST to the assumed identity with identity > 97%. All taxa except Epischura, Mesocyclops, and Senecella had fewer than 1% incorrect sequences. For Mesocyclops, most sequences that were labelled incorrect did BLAST to Mesocyclops, but with an identity of less than 97%. Carcinus sequences exhibited high identity with many reference sequences.

Morphological Identity	Assumed BLAST Identity	Correct Sequences	Ambiguous Sequences	Incorrect Sequences	Correct + Ambiguous (%)	Incorrect (%)
Artemia salina	*Artemia salina*	2137	0	8	99.6	0.4
Balanus crenatus	*Balanus crenatus*	14724	0	8	99.9	0.1
Brachionus calyciflorus	*Brachionus calyciflorus*	207	0	0	100.0	0.0
Cancer sp.	*Cancer sp.*	1620	4	5	99.7	0.3
Carcinus maenas	*Carcinus maenas*	43	157	0	100.0	0.0
Cercopagis pengoi	*Cercopagis pengoi*	1217	0	5	99.6	0.4
Corbicula fluminea	*Corbicula fluminea*	46842	0	73	99.8	0.2
Daphnia mendotae	*Daphnia sp.*	694	11	1	99.9	0.1
Diacyclops thomasi	*Diacyclops bicuspidatus*	0	812	0	100.0	0.0
Dreissena rostriformis bugensis	*Dreissena rostriformis bugensis*	200	0	0	100.0	0.0
Echinogammarus ischnus	*Chaetogammarus ischnus*	7280	0	57	99.2	0.8
Epischura lacustris	*Eurytemora affinis*	9628	248	126	98.7	1.3
Leptodiaptomus ashlandi	*Leptodiaptomus ashlandi*	5414	25	21	99.6	0.4
Mesocyclops edax	*Mesocyclops pehpeiensis*	56	0	999	5.3	94.7
Microsetella norvegica	*Uncultured Metazoan Partial*	809	2	3	99.6	0.4
Oikopleura labradoriensis	*Uncultured Eukaryote 18S*	3543	0	2	99.9	0.1
Palaemonetes sp.	*Palaemonetes sp.*	5163	1	6	99.9	0.1
Pleuroxus denticulatus	*Pleuroxus denticulatus*	642	0	2	99.7	0.3
Senecella calanoides	*Euchirella sp.*	340	2	6	98.3	1.7
Themisto libellula	*Themisto libellula*	4246	1	11	99.5	0.3

For each taxon sample in D1, we generated five new samples by trimming the original sample to each of the lengths tested in this study (300bp, 325bp, 350bp, 375bp, and 400bp). We ran BLASTn on each of the trimmed samples for each taxon with a 97% identity cutoff. We then parsed all the BLAST results and classified each sequence according to the definitions above. To save these classifications, we generated three "BLAST cache" files for each taxon-length combination – one for correct, one for ambiguous, and one for incorrect sequences. In these files, we wrote the sequence labels for all sequences of the given class, for fast reference in the future. With a given OTU from optimization or performance testing, we could then search the cache files for the matching sequence label to determine its class.

5.2.3 Optimization

We tested 1050 parameter sets (see summary, Table 5.4). It is important to note that we tested 150 parameter sets without clustering or denoising, but tested 450 parameter sets with clustering and 450 with denoising because we explored three values for each clustering and denoising parameter. Testing fewer parameter sets without clustering or denoising implies that we explored a smaller space of possibilities for that method of processing, which can potentially lead to reduced observed optimality for this method. However, it was more important that, for each common parameter across the processing methods, we tested the same parameter values to keep the methods comparable. The parameters and values we tested were informed by related studies in the field and the characteristics of our sequence datasets.

Table 5.4: Synopsis and values used of the six sequence processing parameters tested in this study. In total, 1050 parameter sets were tested in the optimization stage. Clustering and denoising steps were optional and mutually exclusive.

Parameter	Synopsis	Values Tested
Sequence Length	Length of sequences – shorter sequences discarded, longer sequences trimmed	300, 325, 350, 375, 400
Minimum Phred Score (Q)	Minimum quality score per base call	10, 20, 30
Maximum Expected Error (MEE)	Sequence-wide expected error score	1.0, 1.5, 2.0, 2.5, 3.0
Clustering Identity Threshold (Optional)	Intraspecific genetic identity threshold	97%, 98%, 99%
Denoising Minimum Abundance Threshold (Optional)	Minimum abundance of a sequence to not be considered noise	2, 4, 8
Singletons	Do we keep unique sequences	Yes, no

Generally, we tested more lenient parameter values than those used in related studies (Bokulich *et al.* 2013; Pawlowski *et al.* 2014; Elbrecht and Leese 2015; Flynn *et al.* 2015; Brown *et al.* 2015; Brown *et al.* 2016; Chain *et al.* 2016; Hänfling *et al.* 2016;

Port *et al.* 2016; Bista *et al.* 2017) because our reads were comparatively long and variable in quality, and sequence quality decreases with sequence length. Aside from the studies by Brown *et al.* (2015) and Elbrecht and Leese (2015), the minimum length cutoff used in all related studies was less than 300bp. Q filters ranged from 20-30 depending on strategy (per base call, sliding window, mean across full sequence *etc.*) and sequence length. MEE filters ranged from 0.5-1.0 (Flynn *et al.* 2015; Brown *et al.* 2015; Port *et al.* 2016; Bista *et al.* 2017). Clustering identity thresholds ranged from 97%-99% across a variety of clustering algorithms (Flynn *et al.* 2015; Brown *et al.* 2015; Brown *et al.* 2016; Chain *et al.* 2016; Port *et al.* 2016; Bista *et al.* 2017; Clarke *et al.* 2017). Some studies discarding singletons while others kept them (Elbrecht and Leese 2015; Flynn *et al.* 2015; Brown *et al.* 2015; Brown *et al.* 2016; Chain *et al.* 2016; Port *et al.* 2016; Bista *et al.* 2017; Clarke *et al.* 2017). The UNOISE3 denoising algorithm (Edgar 2016) was more recently developed and thus none of the aforementioned studies used this algorithm.

Optimization consisted of two parts. In both parts, ranking of parameter sets was based on the number of correct, ambiguous, incorrect, and redundant OTUs generated by the pipeline. Redundant OTUs were defined, for a each taxon, as those for which a correct or ambiguous OTU was already found. For a given taxon, if we had a correct and an ambiguous OTU, the correct OTU took precedence and the ambiguous OTU was reclassified to redundant. Thus, a sample could yield at most 20 correct or ambiguous OTUs in total (one for each taxon), and any remaining correct or ambiguous OTUs were considered redundant. We considered the number of redundant OTUs in optimization for two reasons. First, it could take significantly more processing (manual work) to determine the identity of an OTU, particularly if it was correct but had multiple high-scoring hits in BLAST, or if it was ambiguous. Secondly, more computational time would have been necessary for downstream analysis if there were more OTUs with which to work. In each part of the optimization stage, to find optimal sets, we ranked the parameter sets in order of decreasing optimality based on the following criteria. Part I was designed to find parameter sets that most accurately estimated species richness (*i.e.* minimized false negatives and false positives with varied sequence abundances) from a bulk zooplankton sample (Figure 5.3a). Part II was designed to find parameter sets with high sensitivity (*i.e.* minimized false negatives with low sequence abundances, Figure 5.3b), which is more useful in the detection of AIS. In part I, we combined the samples from all 20 taxa from D1 to construct a single mock community sample. The number of sequences for each D1 taxon ranged from 200 to 46915. In part II of optimization, we generated 100 samples, each consisting of 1000 sequences. We generated these samples by randomly resampling D1, aggregating subsamples of 50 sequences from each taxon to form mock communities with low sequence abundance. Using only 50 sequences from each taxon forced the optimization process to favor more sensitive parameter sets – those

that could successfully recover taxa even with low sequence abundance – which was more appropriate when minimization of false negative error was vital. In both part I and part II, we then tested all 1050 parameter sets on all samples and computed the number of correct, ambiguous, incorrect, and redundant OTUs generated by the pipeline using the given parameter set across all samples. Finally, we ranked the parameter sets according to the optimization ranking scheme described below.

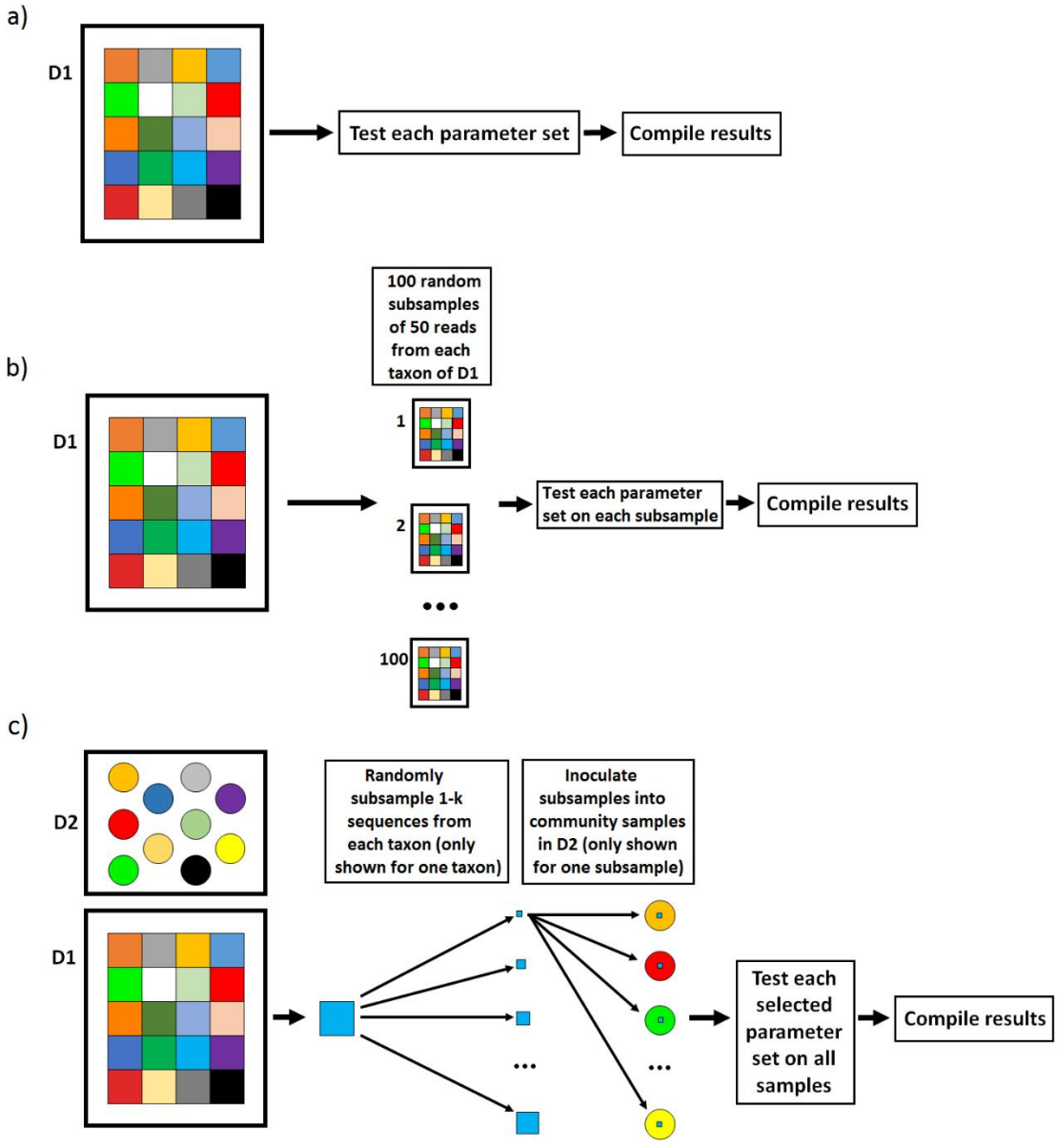


Fig. 5.3. The optimization method for accurate species richness estimates (a), early detection of AIS (b), and the performance testing method (c). Different colored boxes represent different taxa in dataset 1 (D1), and different colored circles represent different community samples in dataset 2 (D2). For performance testing of parameter sets optimized for accurate species richness estimates, k = 100. For performance testing of parameter sets optimized for early detection of AIS, k = 50. For a

given iteration i, where $1 \leq i \leq k$, different random subsamples with i sequences from a given taxon were used to inoculate each community.

In both parts of the optimization process, parameter set *a* was considered more optimal than parameter set *b* if the former's total number of correct and ambiguous OTUs was greater than the latter's. In the case of a tie, the parameter set with the greater number of correct OTUs was more optimal. Missing correct or ambiguous OTUs constitutes a false negative error, which is problematic in estimating species richness but potentially catastrophic in early detection of AIS. If two parameter sets were still tied, the parameter set with the fewest incorrect OTUs was considered more optimal. The more incorrect OTUs generated by a parameter set, the more likely a user could have been to commit a false positive error using that parameter set. If the number of incorrect OTUs was equal as well, the parameter set with fewer redundant OTUs was more optimal. If all OTU counts were equal, the parameter sets performed equally. For each part of the optimization process, we grouped the optimization results by parameter sets that performed clustering, denoising, or neither so that we could compare these three sequence processing methods.

To determine the concordance of parameter set rankings between the two research goals, we computed the Kendall rank correlation coefficient on the ranked parameter set lists for each sequence processing method. Furthermore, we determined the relative contribution to false negative and false positive errors of each of the parameters for six cases: three sequence processing methods across two research goals. In each case, we performed a multiple regression analysis using optimization results. The predictors were the parameter values and the response variables were the number of correct + ambiguous OTUs (which indicates increasing false negative error as it decreases from 20) and the number of incorrect OTUs (which indicates increasing false positive error as it increases from zero). We standardized parameter values for each regression, which allowed us to use the magnitude of the regression coefficients to rank parameters by their relative contributions. In each case, we reported the regression coefficients (to indicate relative contribution) and associated *p* values (to indicate significance of their contributions).

5.2.4 Performance Testing

We ran a series of simulations to test performance of the pipeline in detecting target sequences that were computationally inoculated into real bulk zooplankton samples using 24 selected parameter sets from optimization (Figure 5.3c). The "target" sequences were a subset of sequences all belonging to a single AIS from D1. We chose 12 parameter sets from optimization part I and 12 from part II. We did not simply choose the top 12 parameter sets from each part of optimization because many of the top parameter sets were quite similar. For both parts of the optimization stage, we chose four parameter sets

for each processing method – clustering, denoising, and neither clustering nor denoising. We always chose the top parameter set for a processing method, and subsequently selected parameter sets that performed the next best but were at least two parameters different from any other previously selected parameter set until we had a total of four parameter sets for that category. We conducted performance testing in two parts, mirroring the two parts of optimization. In part I, a simulation consisted of inoculating each port sample in D2 with target sequences, iterating from 1-100 randomly selected sequences of a target taxon from D1. We did this for every taxon in D1. We then ran the pipeline with all selected parameter sets from optimization part I on the simulated data. For each combination of target taxon, port, and parameter set, we performed 25 simulations. For each simulation, we recorded if the target was detected with up to 100 sequences inoculated into the sample and, if so, how many sequences were required to detect it. Therefore, we defined two measures of performance: detectability and sensitivity. Detectability was defined as the ratio of simulations in which the target was found given some number of target sequences inoculated into a community sample. Sensitivity was defined as the number of sequences required to detect the target. Sensitivity was not recorded if the target was not detected. Part II was identical to part one, except we used selected parameter sets from optimization part II and inoculated only up to 50 sequences of the target into the sample because the parameter sets from optimization part II were expected to be far more sensitive. We inoculated up to 50 sequences of the target due to computational constraints and because we found in preliminary work that if the target was not found with 50 sequences in the sample, it was likely undetectable.

5.3 Results

5.3.1 Optimization

Classification of sequences prior to optimization revealed that D1 could yield, at most, 1484 incorrect OTUs and trimming alone could be responsible for false negative error. D1 ranged from 749 incorrect sequences at length 400bp to 1484 at length 325bp, with the remaining sequences classified as correct or ambiguous. At most, only 19 taxa could possibly be recovered (18 correct and one ambiguous) using sequence lengths of 300bp and 325bp, whereas at lengths 350bp, 375bp, and 400bp all 20 taxa could be recovered (19 correct and one ambiguous). The samples we used to optimize for early detection of AIS ranged from a mean of 53.9 (SD = 2.1) incorrect sequences at length 325bp to 25.9 (SD = 4.0) at length 400bp, with the remaining sequences classified as correct or ambiguous. The mean total number of taxa that could be recovered was 19.00 (SD = 0.00) at lengths 300bp and 325bp (18 correct and one ambiguous), 19.97 (SD = 0.17) at

length 350bp (18.97 correct and one ambiguous) and 20 (SD = 0) at lengths 375bp, and 400bp.

When optimizing for species richness estimation without clustering or denoising, incorrect OTUs ranged from 19 to 613 for 27 parameter sets that recovered all taxa (Figure 5.4a). With clustering, 18 parameter sets that recovered all taxa yielded from four to 184 incorrect OTUs (Figure 5.4b). Only the top 22 parameter sets using denoising recovered all 20 taxa without any incorrect OTUs (Figure 5.4c). With denoising, 46 parameter sets recovered all taxa with a maximum of 8 incorrect OTUs. When optimizing for early detection of AIS, no parameter set recovered all taxa without allowing some incorrect OTUs to pass through. Without clustering or denoising, 13 parameter sets recovered all taxa however they yielded a mean of 11.1 incorrect OTUs as well (Figure 5.5a). Further, 43 parameter sets without clustering or denoising recovered at least 19 taxa on average. No parameter set involving clustering recovered all taxa in all replicates; however, 15 recovered a mean of over 19 taxa and yielded a mean of 3.5 incorrect OTUs with a maximum of 9.1 incorrect OTUs (Figure 5.5b). With denoising, the top four parameter sets recovered all taxa while only yielding at most a mean of 0.25 incorrect OTUs (Figure 5.5c).

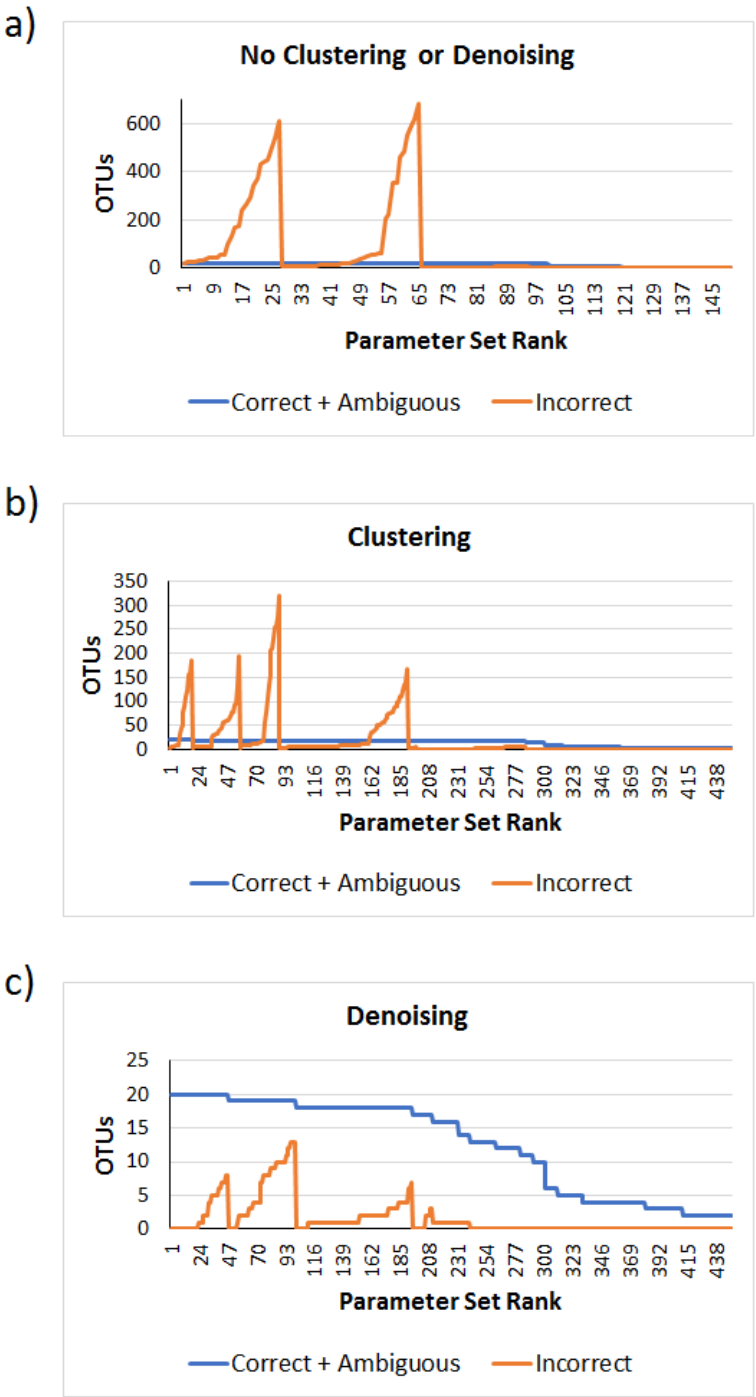


Fig. 5.4. Number of correct + ambiguous and incorrect OTUs for parameter sets optimized for accurate estimates of species richness using no clustering or denoising (a), clustering (b), and denoising (c), by optimization rank. Note the difference in scale on the y axis and that we tested fewer parameter sets using no clustering or denoising.

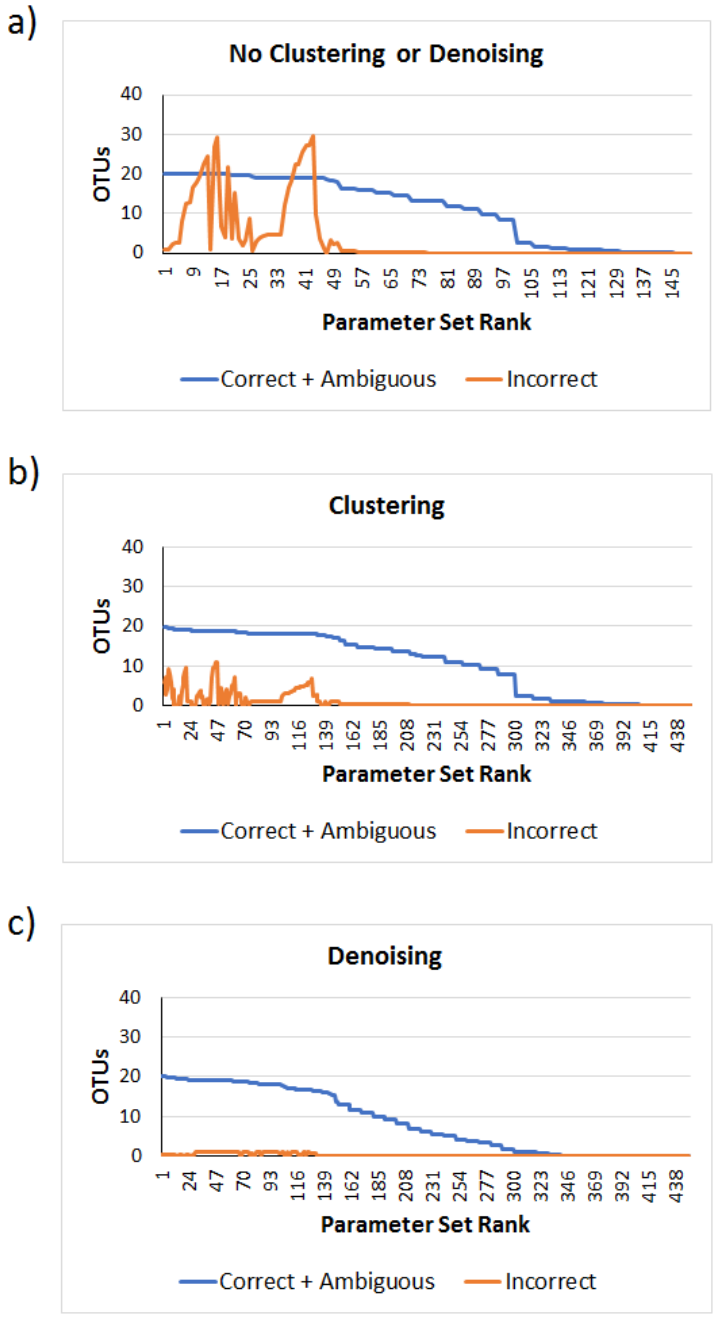


Fig. 5.5. Number of correct + ambiguous and incorrect OTUs for parameter sets optimized for high sensitivity using no clustering or denoising (a), clustering (b), and denoising (c), by optimization rank. Note that we tested fewer parameter sets using no clustering or denoising.

The most optimal parameter sets favored longer sequences with relatively weak filtering. For example, of the top 20 parameter sets from each category (clustering, denoising, or neither, for estimation of species richness or early detection of AIS – 120 parameter sets in total), 106/120 (88.3%) trimmed sequences at length ≥ 375bp. Trimming at shorter lengths was only viable if no clustering or denoising was performed,

and even then it was suboptimal. No top 20 parameter set in any category used a Q filter with strength > 10. Of top ten parameter sets from each category, the mean MEE filter was 2.23, which was relaxed with respect to the range tested and relative to the literature (Flynn *et al.* 2015; Brown *et al.* 2015; Port *et al.* 2016; Bista *et al.* 2017). When aiming to optimize species richness estimation, the MEE filter had a mean of 2.12 (Table 5.5, selected parameter sets), whereas for early detection of AIS it was 2.33 (Table 5.6, selected parameter sets). When denoising, the MEE filter in top ten parameter sets was more relaxed, particularly for early detection of AIS (mean MEE = 2.60). The top 12 parameter sets for accurate species richness estimation for pipelines without clustering or denoising all discarded singletons, as did the top five optimized for early detection of AIS. For pipelines involving clustering, the top eight parameter sets discarded singletons when seeking to optimize species richness estimation. Conversely, the top nine parameter sets with clustering kept singletons when optimizing for early detection of AIS. For denoising, keeping or discarding singletons did not matter because the minimum denoising abundance threshold tested was two. Using clustering, the top 18 parameter sets for accurate species richness estimation used an identity threshold of 99%, whereas the top 24 parameter sets for early detection of AIS also used an identity of 99%. For denoising, the top 14 parameter sets for species richness estimation used a minimum abundance threshold of eight, whereas the top 12 parameter sets for early detection of AIS used a threshold of two sequences.

Table 5.5: Selected high-ranking parameter sets from optimization using all sequences in D1 in a single sample. This optimization aimed to determine parameter sets that most accurately reconstruct a community with low false positive error. We selected four parameter sets per processing method (clustering, denoising, or neither) by selecting the best one and subsequently selecting those that had at least two parameters different from any previously selected (to reduce redundancy for performance testing). Correct operational taxonomic units (OTUs) were those that BLASTed to the assumed identity with rank 1 (using the BLASTn default sort method) and identity > 97%. Ambiguous OTUs were those that BLASTed to the assumed identity with rank > 1 and identity > 97%. Incorrect OTUs were those that did not BLAST to the assumed identity with identity > 97%. Correct and ambiguous OTUs were combined here because the correct identity of ambiguous OTUs could be determined downstream. Trim length is the length of sequences in base pairs (bp), Q filter is the strength of the Phred score filter, MEE filter is the strength of the maximum expected errors filter, and clustering ID is the clustering identity threshold. For processing methods, C represents Clustering, D represents Denoising, and NCOD represents No Clustering Or Denoising. These parameter sets were used for performance testing. Overall, 17/1050 parameter sets recovered all 20 AIS with only 50 sequences from each taxon in all 100 replicates.

Optimized for Accurate Species Richness Estimates

Trim Length (bp)	Q Filter	MEE Filter	Processing Method	Clustering ID (%) or Denoising Minimum Abundance	Singletons	Correct + Ambiguous OTUs	Incorrect OTUs	Redundant OTUs
375	10	1.5	D	8	No	20	0	15

375	10	3.0	D	8	Yes	20	0	16
400	10	2.0	D	8	No	20	0	24
400	10	2.5	D	8	Yes	20	0	25
400	10	1.5	C	99	No	20	4	59
375	10	2.0	C	99	No	20	8	110
400	10	1.5	NCOD	N/A	No	20	19	4250
400	10	1.0	C	99	Yes	20	32	562
375	10	2.0	NCOD	N/A	No	20	36	4345
350	10	2.5	NCOD	N/A	No	20	54	4299
375	10	1.5	C	99	Yes	20	78	1826
400	10	1.0	NCOD	N/A	Yes	20	95	14155

Table 5.6: Selected high-ranking parameter sets from optimization with 50 sequences of each taxon from D1 per sample, with 100 replicates. This optimization aimed to determine parameter sets that yielded high sensitivity. We selected four parameter sets per processing method (clustering, denoising, or neither) by selecting the best one and subsequently selecting those that had at least two parameters different from any previously selected (to reduce redundancy for performance testing). Correct operational taxonomic units (OTUs) were those that BLASTed to the assumed identity with rank 1 (using the BLASTn default sort method) and identity > 97%. Ambiguous OTUs were those that BLASTed to the assumed identity with rank > 1 and identity > 97%. Incorrect OTUs were those that did not BLAST to the assumed identity with identity > 97%. Correct and ambiguous OTUs were combined here because the correct identity of ambiguous OTUs could be determined downstream. Trim length is the length of sequences in base pairs (bp), Q filter is the strength of the Phred score filter, MEE filter is the strength of the maximum expected errors filter, and clustering ID is the clustering identity threshold. For processing methods, C represents Clustering, D represents Denoising, and NCOD represents No Clustering Or Denoising. These parameter sets were used for performance testing. Overall, 17/1050 parameter sets recovered all 20 AIS with only 50 sequences from each taxon in all 100 replicates.

Optimized for Early Detection of AIS

Trim Length (bp)	Q Filter	MEE Filter	Processing Method	Clustering ID (%) or Denoising Minimum Abundance	Singletons	Correct + Ambiguous OTUs	Incorrect OTUs	Redundant OTUs
400	10	3	D	2	No	20	0.2	21.3
400	10	2.5	D	2	Yes	20	0.3	21.1
400	10	2.5	NCOD	N/A	No	20	1.0	59.2
375	10	2	NCOD	N/A	No	20	2.3	50.6
400	10	1.5	NCOD	N/A	Yes	20	8.2	349.1
375	10	2.5	NCOD	N/A	Yes	20	22.6	409.1
400	10	2.5	C	99	Yes	19.8	5.8	37.0

375	10	3	D	2	Yes	19.75	0.4	18.8
375	10	2.5	D	2	No	19.73	0.4	18.7
375	10	3	C	99	Yes	19.6	9.1	61.9
400	10	3	C	99	No	19.3	0.0	1.4
375	10	2.5	C	99	No	19.0	0.3	2.6

We observed concordance of parameter set rankings determined by optimization for the two research goals. When clustering was used, the Kendall tau was 0.80, signifying strong concordance ($p < 0.001$). The Kendall tau was 0.79 when denoising was used and 0.77 when no clustering nor denoising was used ($p < 0.001$ in each case). Multiple regression analysis determined that parameter selection accounted for less variation in the number of correct + ambiguous OTUs recovered when determining species richness (80%, 89%, and 80%, when clustering, denoising, or neither, respectively; see Table 5.7 for summary of multiple regression results) than when aiming for early detection of AIS (95%, 94%, and 95% respectively). Conversely, given either research goal, parameter selection accounted for comparable amounts of variation in the number of incorrect OTUs recovered (41%, 51%, and 47% for estimation of species richness, 48%, 55%, and 50% for early detection of AIS).

Table 5.7: Coefficients and p values for multiple regression given standardized parameter values to predict the number of correct + ambiguous OTUs and the number of incorrect OTUs for each sequence processing method and for each research goal. Coefficient magnitude signifies importance of the corresponding parameter in determining the predicted value, and p value indicates significance of impact. "Q" denotes Q filter, "Length" denotes sequence length, "Singletons" denotes whether singletons were kept or discarded, "MEE" denotes maximum expected error filter, "ID" denotes clustering identity threshold, and "DMA" denotes denoising minimum abundance.

			Correct + Ambiguous		Incorrect	
Research Goal	**Processing Method (adj. r-squared correct + ambiguous, adj. r-squared incorrect)**	**Parameter**	**Coefficient**	**p value**	**Coefficient**	**p value**
Species Richness	Clustering (0.80, 0.41)	Length	-0.474	<0.001	-4.29	0.007
		Q	-5.746	<0.001	-21.21	<0.001
		MEE	0.035	0.797	6.47	<0.001
		ID	0.163	0.225	6.46	<0.001
		Singletons	-0.42	0.002	-15.46	<0.001
	Denoising (0.89, 0.51)	Length	-0.5	<0.001	-0.494	<0.001
		Q	-6.294	<0.001	-1.5749	<0.001
		MEE	0.091	0.399	0.1542	0.082
		DMA	-0.995	<0.001	-0.9596	<0.001
		Singletons	<0.001	1	<0.001	1
	Neither (0.80, 0.47)	Length	-0.44	0.085	-12.97	0.165
		Q	-6.12	<0.001	-86.02	<0.001
		MEE	0.028	0.911	26.11	0.006
		Singletons	-0.421	0.099	-59.2	<0.001

Early Detection of AIS	Clustering (0.95, 0.48)	Length	-0.5478	<0.001	-0.2049	0.001
		Q	-7.1928	<0.001	-1.0415	<0.001
		MEE	0.065	0.404	0.2399	<0.001
		ID	0.107	0.169	0.2183	0.001
		Singletons	-0.8472	0	-0.664	0
	Denoising (0.94, 0.55)	Length	-0.6302	0	-0.1134	0
		Q	-7.3123	0	-0.2579	0
		MEE	0.1327	0.13	0.024	0.049
		DMA	-1.4029	0	-0.0425	0.001
		Singletons	0	1	0	1
	Neither (0.95, 0.50)	Length	-0.52	0.001	-0.762	0.077
		Q	-7.627	0	-4.331	0
		MEE	0.065	0.661	1.149	0.008
		Singletons	-0.873	0	-2.643	0

We found that, regardless of research goal or processing method, Q filter strength most strongly determined both the number of correct + ambiguous OTUs recovered and the number of incorrect OTUs recovered ($p < 0.001$ in each case; coefficient and p values, Table 5.7; ranking of parameter importance, Table 5.8). Generally, MEE filtration had little contribution to correct + ambiguous OTU counts, and was most significant ($p = 0.13$) when denoising was used for early detection of AIS. Conversely, MEE filtration was generally important in reducing the number of incorrect OTUs ($p < 0.05$ in all cases, except when denoising for species richness estimates), always ranking third except when denoising was performed (in which case it ranked fourth). Sequence length was generally important in determining correct + ambiguous OTUs ($p < 0.05$ except when no clustering nor denoising was used for species richness estimation), with a mean rank of three. On the other hand, sequence length generally had a weaker contribution to the number of incorrect OTUs (mean rank = 3.8), and was insignificant when neither clustering nor denoising was used for either research goal ($p>0.05$ in both cases). Keeping or discarding singletons was insignificant in determining either OTU count (correct + ambiguous or incorrect) when denoising was used, for either research goal. Otherwise, its mean rank was 2.5 for recovering correct + ambiguous OTUs and 2.0 in all cases for recovering incorrect OTUs ($p < 0.05$ in all cases except when no clustering or denoising was used for estimation of species richness). When clustering was used, identity threshold ranked fourth for each research goal and OTU count, and was not significant in determining the number of correct + ambiguous OTUs (otherwise, $p < 0.05$). Conversely, clustering identity threshold strongly impacted the number of incorrect OTUs ($p < 0.05$ for each research goal). When denoising was used, the denoising minimum abundance had a significant impact in all cases ($p < 0.05$) with a mean rank of 2.3.

Table 5.8: Parameter rankings (denoted as "Rank") for each goal (estimation of species richness or early detection of AIS) and for each sequence processing method (clustering, denoising, or neither), in terms of relative impact on the two optimization criteria (correct + ambiguous OTUs, incorrect OTUs). "Q" denotes Q filter, "Length" denotes sequence length

cutoff, "Singletons" denotes whether singletons were kept or discarded, "MEE" denotes maximum expected error filter, "ID" denotes clustering identity threshold, and "DMA" denotes denoising minimum abundance. See Table 5.7 for coefficients and p values related to parameter impacts, determined by standardized multiple regression. Asterisk denotes significant impact at α = 0.05.

Species Richness

	Rank	Correct + Ambiguous	Incorrect
Clustering	1	Q*	Q*
	2	Length*	Singletons*
	3	Singletons*	MEE*
	4	ID	ID*
	5	MEE	Length*
Denoising	1	Q*	Q*
	2	DMA*	DMA*
	3	Length*	Length*
	4	MEE	MEE
	5	Singletons	Singletons
Neither	1	Q*	Q*
	2	Length	Singletons*
	3	Singletons	MEE*
	4	MEE	Length

Early Detection of AIS

	Rank	Correct + Ambiguous	Incorrect
Clustering	1	Q*	Q*
	2	Singletons*	Singletons*
	3	Length*	MEE*
	4	ID	ID*
	5	MEE	Length*
Denoising	1	Q*	Q*
	2	DMA*	Length*
	3	Length*	DMA*
	4	MEE	MEE*
	5	Singletons	Singletons
Neither	1	Q*	Q*
	2	Singletons*	Singletons*
	3	Length*	MEE*
	4	MEE	Length

5.3.2 Performance Testing

Distributions of the number of sequences necessary to detect targets varied by parameter set and exhibited positive skewness (*i.e.* parameter sets optimized for early detection of AIS without clustering or denoising, Figure 5.6 – long tails above the mean and fewer samples below). No distribution for any single taxon, port, or parameter set (optimized for either research goal) was normal (Kolmogorov-Smirnov test for normality, $p < 0.05$ in all cases), yielding generally high variance.

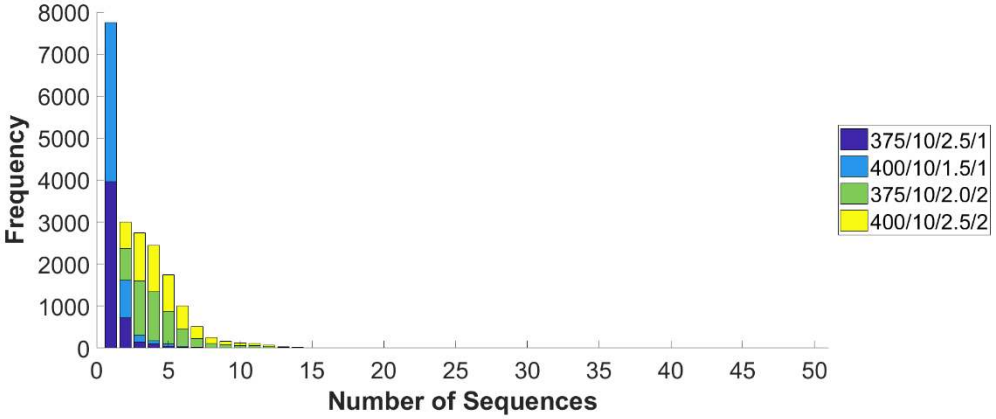


Fig. 5.6. Distributions of number of sequences required (x-axis) to detect target taxa in community samples for parameter sets optimized for the early detection of AIS using no clustering or denoising. Frequency was the number of simulations in which the target was detected at a given number of sequences inoculated. Each color represents a selected parameter set. Parameter sets are in the format "length/Q filter/MEE filter/minimum abundance".

For parameter sets optimized for species richness estimation, detectability with 10 target sequences inoculated into the port sample was nearly perfect without clustering or denoising for all taxa aside from *Brachionus*, *Dreissena*, and *Mesocyclops* (Figure 5.7a, left column). The latter species detectability was poor owing to the low quality of their sequences relative to those for other taxa. A similar pattern was observed with clustering, though several ports (*e.g.* Hamilton, Nanticoke, and Thunder Bay) yielded low detectability for several taxa (Figure 5.7b, left column). Detectability across all combinations of port and taxon was very poor when denoising was used (Figure 5.7c, left column).

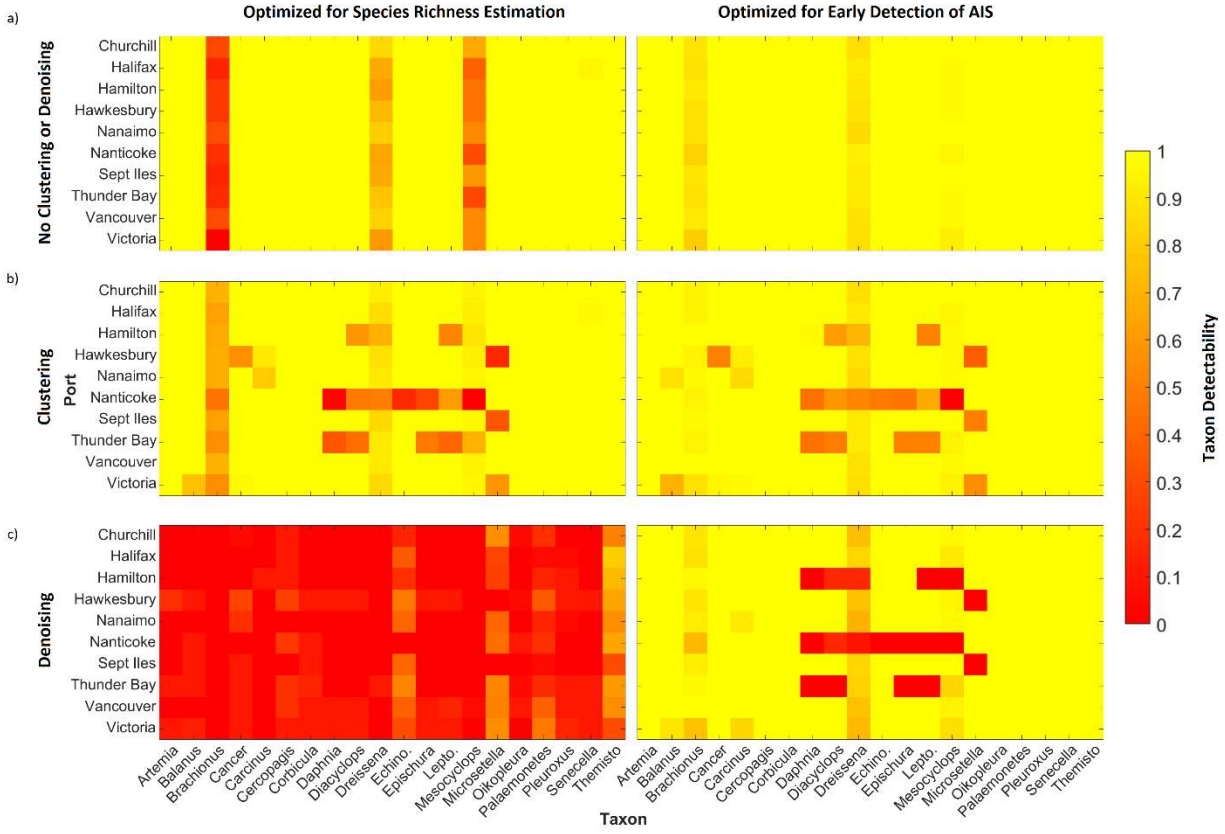


Fig. 5.7. Detectability of taxa in mock samples, given as a value between 0 (no detectability of the target taxon at the port; red) and 1 (perfect detectability of the target taxon at the port; yellow) for parameter sets optimized for estimation of species richness (left column) and parameter sets optimized for early detection of AIS (right column) using no clustering or denoising (a), clustering (b), and denoising (c), across all ports and taxa, with 10 sequences of each taxon inoculated into the original port sample. Detectability for a given port and taxon was computed using all replicates involving the given port and taxon (that is, across all parameter sets tested). See Table 5.3 for species names.

A similar but slightly improved detectability pattern was observed for parameter sets optimized for early detection of AIS not using clustering or denoising, when

compared to those optimized for species richness (Figure 5.7a). Detectability of *Brachionus* and *Mesocyclops* were significantly improved across ports for parameter sets using clustering optimized for early detection of AIS when compared to those optimized for estimation of species richness ($p < 0.001$), otherwise there were no significant differences in detection for any port or taxon ($p > 0.05$). A similar detectability pattern was observed for clustering using parameter sets optimized for early detection of AIS as compared to those optimized for estimation of species richness (Figure 5.7b), though a slight overall improvement was observed (only *Brachionus* detectability was significantly improved; $p < 0.001$). Overall, we observed high variation in recovery ratio across ports and target when clustering or denoising was performed with parameter sets optimized for early detection of AIS (Figure 5.7b and 5.7c, right column). For example, the freshwater ports of Nanticoke, Thunder Bay, and Hamilton yielded low detectability, as recovery ratios were only 0.806, 0.887, and 0.939, respectively, when clustering was used. When denoising, the respective recovery ratios were even lower, only 0.648, 0.782, and 0.765. We observed no cases where a taxon could not be detected if 10 target sequences were present in the sample when clustering was optimized for early detection of AIS. Though the pattern for denoising was similar to that of clustering, many combinations of taxon and port yielded no detectability (Figure 5.7c, right column). Nevertheless, denoising parameter sets optimized for early detection of AIS yielded a significant improvement in detectability over those optimized for estimation of species richness for all taxa and all ports ($p < 0.05$).

Using parameter sets optimized for species richness estimation, detectability confidence reached 90% and 95% with the fewest sequences required using pipelines without clustering or denoising (Figure 5.8a). On average, 6.3 and 8.5 sequences were required to detect the target in 90% and 95% of replicates, respectively, when neither clustering nor denoising were used. With clustering, these values rose to 8.6 and 16.6 sequences, respectively. Denoising performed much worse, requiring 69.8 target sequences to reach 90% detectability while 95% detectability was unattainable. Detectability confidence was maximized in parameter sets optimized for early detection of AIS when clustering and denoising were not performed (Figure 5.8b). Without clustering or denoising, only 5.3 and 6.6 sequences were required for 90% and 95% detectability respectively, 15.2% and 22.6% lower than when parameter sets were optimized for species richness estimates. These values rose to 6.8 and 11.3 target sequences, respectively, when clustering was used (11.2% and 31.8% lower than parameter sets optimized for species richness estimates, respectively), and 10.6 and 43.4 target sequences when denoising was used (84.9% fewer sequences for the 90% interval than parameter sets optimized for species richness estimates).

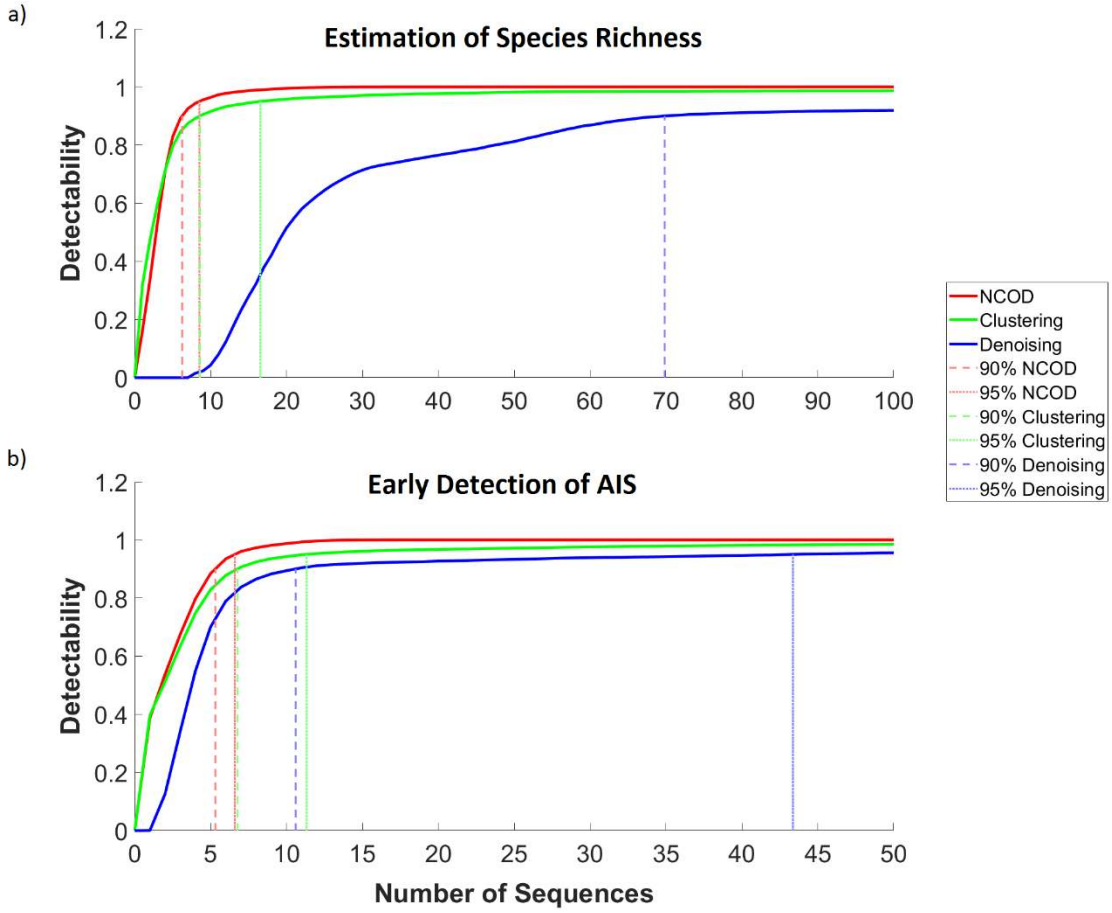


Fig. 5.8. Overall detection probability of taxa for parameter sets optimized for estimation of species richness (a) and early detection of AIS (b) using no clustering or denoising ("NCOD" – red), clustering (green), and denoising (blue), per number of target sequences inserted into the original sample. Detection probability was computed using all combinations of taxon, port, and parameter sets, across 25 replicates. Shown as dotted lines are 90% and 95% detection for each sequence processing method. Note the difference in x-axis labels. For estimation of species richness using denoising, 95% detection was not achieved.

With parameter sets optimized for species richness estimation, sensitivity was far worse if denoising was used than if clustering or neither clustering nor denoising were used. Without clustering or denoising, only 3.9 (SD = 3.1) sequences were required to detect the target. This increased to 4.5 (SD = 7.0) sequences when clustering, and to 25.3 (SD = 16.4) when denoising. As expected, sensitivity improved with the top parameter sets that had been optimized for early detection of AIS. We found that 3.6 (SD = 4.9) reads were required to detect AIS (when they were detectable) using clustering, whereas denoising required 5.5 (SD = 5.8) reads. Without clustering or denoising, the pipeline was very sensitive, requiring only 2.9 (SD = 2.2) sequences. With clustering, we detected the AIS in only 98.5% of cases with 50 sequences inoculated. In contrast, denoising and neither clustering nor denoising detected the AIS in 95.5% and 100% of cases, respectively.

For parameter sets optimized for early detection of AIS, four taxa (*Daphnia*, *Diacyclops*, *Dreissena*, and *Leptodiaptomus*; Figure 5.9a – sensitivity for parameter sets optimized for early detection of AIS across taxa) required more than five sequences to be detected if clustering was used. This value rose to nine taxa if denoising was used, with the highly invasive *Dreissena* requiring the most sequences (mean 10.3; SD = 6.0). Without either clustering or denoising, only one taxon (*Brachionus*) required more than five sequences for detection (5.8; SD = 3.5). Variance in sensitivity was greater in taxa that yielded reduced sensitivity.

Using parameter sets optimized for early detection of AIS, we found that sensitivity varied little across ports (Figure 5.9b; sensitivity for parameter sets optimized for early detection of AIS across ports) except for Nanticoke when clustering (sequences required = 6.3; SD = 9.3) or denoising (sequences required = 11.2; SD = 12.9) was performed. Hamilton and Thunder Bay also yielded relatively lower sensitivity when clustering was done, requiring 3.9 (SD = 5.4) and 5.2 (SD = 8.1) sequences, respectively, or 6.6 (SD = 7.8) and 6.5 (SD = 7.5) sequences when denoising was performed. Sensitivity across ports was very consistent with or without clustering and denoising.

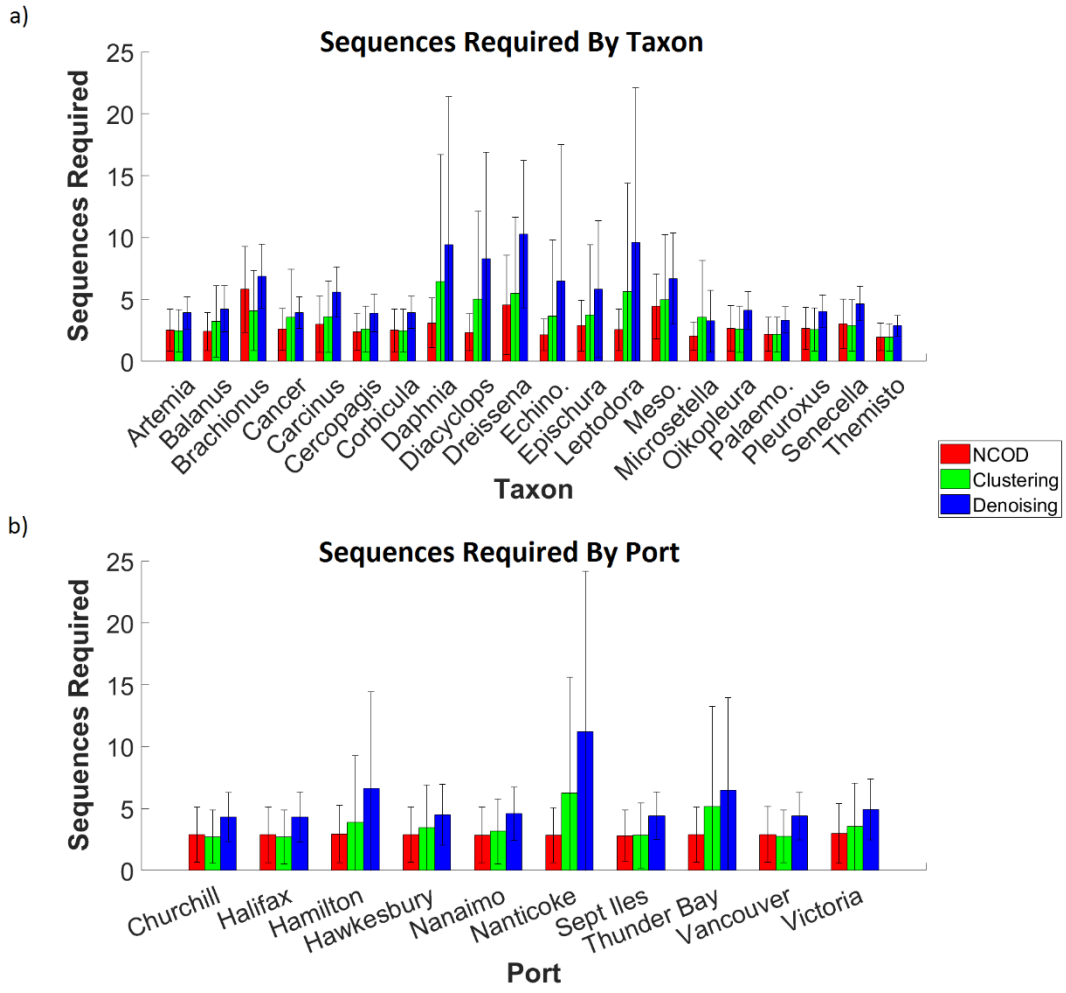


Fig. 5.9. The sensitivity per taxon across all ports (a) and per port across all taxa (b), for parameter sets optimized for early detection of AIS using no clustering or denoising ("NCOD" – red), clustering (green), and denoising (blue). Error bars show standard deviation from the mean. See Table 5.3 for species names.

5.4 Discussion

In this study, we sought to assist users to optimally select processing steps and parameter values for sequence processing pipelines during metabarcoding of bulk zooplankton samples for the 18S marker on the 454 platform. Generally, we observed that trimming sequences to 375-400bp was most favorable when a 400-600bp fragment was sequenced, and mild sequence quality filtration ($1.5 \leq MEE \leq 3.0$, $Q = 10$) worked best when overall sequence quality varied across samples (see summary of our findings on optimal parameter selection in Table 5.9). In optimization, denoising outperformed pipelines using clustering or neither clustering nor denoising regardless of the research objective. However, performance testing revealed that sequences – particularly at low abundance – of some taxa could wrongly be classified as noise during denoising, which resulted in false negative errors (see Figure 5.7). Denoising pipelines also yielded very different

distributions for sensitivity when compared to those that used clustering or neither clustering nor denoising. Denoising could drastically reduce sensitivity, particularly if the minimum abundance threshold for denoising was high (eight sequences). However, a high denoising minimum abundance threshold did reduce false positive errors, which indicated that it was useful for species richness estimates but not for early detection of AIS (when sensitivity and detectability are imperative). Naturally, without clustering or denoising, the pipeline was most sensitive and yielded highest detectability, as the AIS targets were detected in every case. Both clustering and denoising reduced false positive errors in optimization, however these errors could be mitigated with further processing, so skipping clustering and denoising proved the best way to process metabarcoding sequences for the early detection of AIS.

Table 5.9: Summary of optimal sequence processing pipeline parameter selection for zooplankton 18S metabarcoding, given two research goals: estimation of species richness and early detection of AIS. "Q" denotes Q filter, "Length" denotes sequence length cutoff, "Singletons" denotes whether singletons should be kept or discarded, "MEE" denotes maximum expected error filter, and NCOD denotes "No Clustering or Denoising". Note that keeping or discarding singletons in the early detection of AIS depends on whether the user will be clustering the data or not.

Parameter/Option	Estimation of Species Richness	Early Detection of AIS
Length	375-400 bp	375-400 bp
Q	10	10
MEE	1.5-2.5	2-3
Singletons	Discard	Depends on processing method. NCOD? Discard. Clustering? Keep.
Clustering Identity	99%	99%
Denoising Minimum Abundance	8	2
Processing Method	Denoising	No Clustering or Denoising

Our study is the first to optimize such a sequence processing pipeline for metazoan bulk sample metabarcoding. In addition, we tested 1050 parameter combinations for two different research objectives (*i.e.* estimating species richness and early detection of AIS). Other studies have focused on a single aspect of the sequence processing pipeline (Zhan, Xiong, *et al.* 2014; Brown *et al.* 2015), tested relatively few combinations of parameters (Flynn *et al.* 2015), tested the ordering of processing steps (May *et al.* 2014), or tested bulk sample processing prior to sequencing, with mostly fixed sequence processing parameters (Zhan *et al.* 2013; Piggott 2016). Brown *et al.* (2015) focused specifically on clustering sequence identity, and found that a 97% identity threshold was sufficient in UPARSE to recover most taxa. Testing many parameter combinations also allowed us to explore interdependency between parameters and processing methods, even though it was computationally intensive. For example, even with high parallelization (~200 concurrent runs) of optimization and performance testing,

the computational time required for this project was approximately two months on a high-performance computing network (with CPU speeds of 2.2-2.7 GHz).

Further, our study is novel in that we tested the performance of optimized pipelines by computationally inoculating sequences of 20 species into real community samples to determine what can be expected for sensitivity and detectability given different combinations of community structure and ecosystem. In related work, Zhan, Xiong, *et al.* (2014) spiked biomass of two AIS into two community samples. They found that relationships between false negative errors and exclusion of singletons, doubletons, and tripletons with varied Phred score filters and biomass of target species spiked into real community samples. With strong filtering (Q = 30), spiked biomass of the marine scallop *Argopecten irradians* could not be detected in a real freshwater sample collected at Nanticoke, Lake Erie, though doubletons were usually recovered, provided relatively weak filtering was done (Q ≤ 20) and sufficient biomass was present (Zhan, Xiong, *et al.* 2014). Flynn *et al.* (2015) tested the ability of a similar pipeline to determine species richness of a mock zooplankton community using relaxed (length 250-600bp, average Q ≤ 20) and stringent (length ≥ 400bp, MEE ≤ 0.5) filtering methods, in combination with three different clustering algorithms with fixed clustering identity (97%). They concluded that UPARSE creates clusters most precisely and that stringent filtering was needed to accurately describe species richness. With a deeper optimization of this pipeline, we have corroborated their suggestions with respect to sequence length; however, our findings indicate that filtration can be more relaxed than they suggested. They also speculated that relaxed filtering might be necessary to recover rare taxa or sequences (*i.e.* in detection of AIS), a finding we explicitly tested and confirmed in this study.

Here, optimization of the pipeline revealed that keeping singletons generally did not reduce false negative errors except when using clustering in the context of early detection of AIS (in which the best nine parameter sets all kept singletons). Otherwise, removing singletons was a simple and uncostly means of reducing false positive errors. Generally, during optimization, retaining singletons increased redundancy and false positive errors without decreasing corresponding false negative errors. Though singletons could represent extremely rare taxa (see Zhan *et al.* 2013; Brown *et al.* 2015), they were more likely to be artifacts (Edgar 2013; Flynn *et al.* 2015). Owing to the high sensitivity of the pipeline despite removal of singletons, we recommend that the advantages of reduced redundancy and false positive errors outweigh the disadvantage of slightly reduced sensitivity. Thus, singletons can generally be removed with little negative impact.

Previous studies covering different taxa, amplified fragments, and applications have utilized sequence processing strategies that included more stringent Q filtering, typically between 20 and 30 (Bista *et al.* 2017; Elbrecht and Leese 2015; Hänfling *et al.*

2016). In our study, with moderate filtering (Q ≥ 20, MEE ≤ 1.5), and especially at longer sequence lengths, all sequences of some species (particularly *Brachionus* and *Mesocyclops*) were removed, resulting in false negative errors whether the aim was to estimate species richness or to maximize sensitivity. This finding corroborated that of Zhan, He, *et al.* (2014), who noted that rare taxa were more likely to be lost with increasing Q filter strength and informational sequences (those that represented otherwise undetected taxa) were removed at any stringency. Relaxed filtration allowed longer sequences to be analyzed downstream, as sequence quality generally decreased with sequence length. This is important because longer sequences generally provided greater taxonomic resolution and accuracy, allowing more appropriate definition of clusters (if clustering is used), more appropriate classification of a read as noisy (if denoising is used), and more accurate taxonomic assignment during BLAST. The downside of relaxed filtration was that it can increase false positive error.

We found that the most optimal parameter sets for estimating species richness allowed slightly more stringent filtration, which corroborated findings of Flynn *et al.* (2015). If the aim of the study is to accurately estimate species richness, sacrificing sensitivity and detectability (*i.e.* increasing false negative error) to decrease false positive error is justifiable. However, users should not increase the stringency of the Q filter as it is extremely sensitive and will remove a sequence if it has a single low-quality base call. Conversely, if the objective is the early detection of AIS, false negative error is typically more costly than false positive error (a false positive error can potentially be mitigated downstream *e.g.* when identifying sequences in BLAST), so filtration should be relaxed. Therefore, with respect to filtration and sequence length, we recommend mild MEE filtration (1.5-2.5 for species richness estimation, 2.0-3.0 for early detection of AIS), relaxed Q filtration (10 at most), and trimming sequences ≥ 375bp. The upper bound on MEE and lower bound on Q filtration holds regardless of sequencing platform, as we used 454 pyrosequencing but cutting-edge sequencers may improve read quality. The lower bound on MEE filtration could be reduced with newer sequencing technology, but Q filtration strength should not be increased for the reasons outlined above. The optimal sequence length depends on the amplified fragment and the length of sequences in the sample (which depends on sampling method and sequencing technology). Our amplified fragment was at least ~400bp in target taxa – and 98% of target sequences were ≥400bp – because we used 454 pyrosequencing of DNA extracted from bulk samples (eDNA sequences will likely be shorter due to degradation). Hence, it is sensible that our optimal sequence length (375-400bp) was close to the minimum amplified fragment length in our taxa; taxonomic resolution was maximized while very few sequences were wrongfully excluded due to failing to meet the length cutoff. In studies where most sequences reach the minimum amplified fragment length in target taxa, we recommend using a length

cutoff of approximately 90-100% of the minimum amplified fragment length in target taxa.

We found that both clustering and denoising were useful in reducing false positives in the estimation of species richness. However, both should be avoided in the context of early detection of AIS because both sensitivity and detectability were reduced. We also found that a 99% clustering identity threshold was more optimal than the commonly-used 97% identity threshold for bulk zooplankton 18S metabarcoding for either research goal, and a denoising minimum abundance threshold of 8 was best for estimation species richness.

Typically, 97% is considered a standard for clustering identity thresholds (see Edgar 2013). Through optimization, we found that 99% clustering identity performed better for zooplankton using the 18S V4 fragment. There were two highly related taxa, *Carcinus* and *Cancer,* which impacted the optimality of parameter sets using clustering. In practice, this situation may occur where two distinct species in a sample share very high identity (> 97%). Users of such sequence processing pipelines will not know in advance what the appropriate clustering threshold is, as it depends on the relatedness of taxa in their sample, and incorrect assignment of the threshold can be a source of errors that have dire consequences (particularly for early detection of AIS; Brown *et al.* 2015). In our study, sensitivity and detectability were reduced when clustering because clusters form that contain sequences from both the community sample and AIS sequences that we introduced to the sample. In real applications of this pipeline, clustering may create clusters with sequences from more than one species, hiding sequences of taxa and inadvertently rendering them undetectable downstream. Fortunately, with modern computers, clustering when a reference database exists is often unnecessary; one could use a parallel computing strategy (*e.g.* in BLAST) that could reduce computational time and keep processing of metabarcoding data tractable. Our findings suggest that clustering reduces taxonomic resolution and removes potentially informational sequences from the dataset prior to taxonomic assignment. Thus, we support suggestions of Brown *et al.* (2016) and Chain *et al.* (2016) to avoid clustering altogether if early detection of AIS is the project goal. However, false positive errors were reduced by clustering spurious reads with their correct counterparts, which is especially beneficial when estimating species richness. Thus, if clustering is necessary or desired, we recommend using a higher similarity threshold than what has been classically used – 99% instead of 97% – to reduce false negatives while simultaneously reducing spurious OTUs.

With respect to denoising, the current version of USEARCH uses UNOISE3, which is a relatively new algorithm and is technically a form of clustering itself. Its likeness to clustering was evident in its performance. In terms of detectability, the combinations of taxa and ports that clustering struggled with were nearly the same as

denoising, though the latter performed slightly worse. For problematic combinations of taxa and ports, denoising usually fared worse than clustering. Further, denoising yielded slightly lower sensitivity than clustering. For early detection of AIS, these characteristics are potentially problematic. False negative errors could occur by incorrectly flagging valid sequences as noise. On the other hand, with respect to species richness estimates, incorrect sequences were removed more effectively through denoising than through any other processing method. The default minimum abundance threshold was 8, which we found worked very well for conducting species richness estimates. However, we also found that this default threshold was not viable when aiming for early detection of AIS. Thus, our recommendation for denoising is like that of clustering. If this pipeline is being used for early detection of AIS, either avoid denoising or use a conservative minimum abundance threshold (for instance, minimum abundance of 2-4 rather than 8). If, on the other hand, the pipeline is being used to estimate species richness, denoising with a minimum abundance threshold of eight will remove a high proportion of spurious reads and serve to reduce false positives. The denoising algorithm of USEARCH does allow users to save all OTUs (including those flagged as chimeric or noisy). Thus, an alternative is to denoise but be cognizant that some sequences may be wrongly flagged as chimeric or noisy. Then, further analysis could then reduce false negatives even after denoising (*e.g.* by running BLAST with the chimeric or noisy sequences, aligning them against denoised OTUs, or applying an evolutionary model to the denoised OTUs and chimeric or noisy sequences).

Application of next-generation sequencing in surveillance of AIS requires careful consideration of many options including sequencing technology, genetic marker, and computational pipeline. Choice in sequencing technology has complex implications, manifested primarily in differences in sequence quality and length. We used 454 pyrosequencing in our study, though newer sequencing platforms could reduce sequencing errors. When this pipeline is used to determine species richness, one can potentially utilize more stringent filtering, though two or three base call errors in a sequence of length $\geq$ 375bp is unlikely to cause a serious problem. Regardless, longer sequences improve taxonomic resolution and weaker filtration allows rare (and potentially otherwise undetectable) taxa to be discovered, thus care must be taken to not filter too strongly in the context of surveillance for AIS.

With respect to marker choice, we used 18S in our study but COI has shown higher sequence variability and improved taxonomic assignment (Tang *et al.* 2012; Zhan, Bailey, *et al*. 2014; Hatzenbuhler *et al.* 2017). This variability can be a double-edged sword; as it is apparent even in primer binding sites, COI can have issues with primer generality (Ficetola *et al.* 2010; Deagle *et al.* 2014; Zhan, Bailey *et al.* 2014; Hatzenbuhler *et al.* 2017). Consequently, false negative errors may be more likely to

occur because of inconsistent amplification which would be particularly detrimental to early detection of AIS. In the metabarcoding context, the variability of COI relative to 18S may impact sequence clustering, denoising, and taxonomic assignment (*e.g.* through BLAST). With a higher-resolution marker, sequences of different species will be more likely to be split into different OTUs during clustering (given some arbitrary identity threshold) and some sequences when denoising may be less likely to be considered noise because of increased sequence divergence. Downstream, taxonomic assignment in BLAST may be more confident for some taxa when using COI. Therefore, higher-resolution markers could increase sensitivity and reduce false negatives whether clustering or denoising are used (because of the aforementioned advantages in sequence processing). However, even with a higher-resolution marker (for example COI), we do not recommend clustering or denoising when conducting early detection of AIS for the reasons mentioned above. Many computational sequence processing suites offer similar (if not identical) features or algorithms for trimming, filtering, clustering, and denoising (Schloss *et al.* 2009; Edgar 2010; Caporaso *et al.* 2010; Cole *et al.* 2014). Consequently, many of our findings are generalizable to different sequence processing suites.

Regardless of marker and despite advancements in next-generation sequencing technologies, sequence quality and processing are, and will continue to be, important issues (van Dijk *et al.* 2014; O'Rawe, Ferson, and Lyon 2015). Benchmarking and optimizing computational pipelines for experiments that use different markers and target aquatic taxa will be helpful for refining metabarcoding analytical guidelines. Testing with different markers may yield different recommendations in terms of sequence length – as it depends on marker length and variability of target regions – and quality filtration – as it depends on sequence length. Testing with different taxa may yield different results across the entire pipeline, depending on the marker used. Because of the prevalence of metabarcoding in current research (and accordingly, the prevalence of computational sequence processing), there is a need for more studies that deeply explore and optimize sequence processing pipelines for different applications. We advise users conducting biological invasions research with metabarcoding to test multiple parameter sets when processing data and, when possible, skip clustering or denoising. One can obtain a consensus from multiple runs with different parameters, improving confidence and gaining different perspectives of the data. In the context of early detection of AIS and across the range of parameters tested, we observed no situation where a parameter did not contribute to either false positive or false negative error in a significant manner (aside from singletons when denoising was used, Table 5.7). Thus, all parameters should be carefully considered in this context.

One important implication of our study is that, in metabarcoding, there will almost always be some false positive error and some false negative error. To fully

eliminate false negative error – especially with low sequence abundance for some taxa, as is ideal in the context of early detection of AIS – there will almost surely be some false positive error and it can become a serious issue. Given the potential difficulty in balancing false positive and false negative errors in this context, does metabarcoding have a place in the early detection of AIS? We believe it does, though it may be difficult to confirm that a target AIS is in a sampled waterbody using metabarcoding (or a single marker) alone. A more effective strategy for conservation or AIS management applications would be to first use metabarcoding with the sequence processing strategy that we suggested, followed by a targeted genetic approach using highly species-specific markers and primers (*e.g.* using COI) or traditional sampling methods to confirm the presence of the species with greater confidence. For a given combination of marker, target taxon, and sampling method, until a deep optimization is performed, analyzing sequence retention given length and filtering strength can provide some information with which to start a small search for good parameters.

CHAPTER 6

Summary, Conclusion, and Future Work

6.1 Summary

In summary, Chapter 2 of this dissertation introduced simulations and artificial life, which are large and interdisciplinary fields that allow us to develop, analyze, describe, and conduct experiments on simplified systems that emulate real systems. It then discussed the basics of biological invasions, their importance, prediction, and management. It discussed individual-based models as a useful tool in their prediction, not only practically but also theoretically. Most practical applications of individual-based models involve conservation of a species that interacts directly with an invader, the testing or outcome predictions of management schemes, or the prediction of spread of the invasive species mapped to real time and space. Theoretical studies, on the other hand, touch a wide variety of aspects of biological invasions but are mostly studies of selection pressures faced by populations early in introduction and expansion. We discussed the importance of early detection when the inevitable occurs and species are introduced and highlighted the use of DNA metabarcoding to identify species in samples (eDNA or bulk). Chapter 3 introduced EcoSim, the individual-based model used as an experimental platform in Chapter 4, using the ODD protocol and provided some results from the standard variant of EcoSim. Many of the results produced by the standard EcoSim variant corroborated observations made in the real world. EcoSim has been shown to produce interesting and realistic predictions concerning eco-evolutionary theories, and the latest rendering of EcoSim shows promise for future experimentation.

In Chapter 4, EcoSim Niches was introduced as a new subvariant of EcoSim. EcoSim Niches was interesting in that it produced an environment with greater complexity, which we correctly hypothesized would produce greater genetic diversity and corresponding diversity in adaptive strategies. We also produced a variant called EcoSim Invasions, which was used to simulate a scenario of multiple introductions of prey populations across two environments reciprocally, and over a gradient of genetic diversity ranging from zero (completely clonal inocula) to an empirically-determined maximum level of diversity. EcoSim Invasions was the first iteration of an EcoSim variant designed to simulate biological invasions occurring across ecosystems evolving in time; the simulation showed great promise in allowing us to study theoretical eco-evolutionary phenomena in the context of biological invasions. Finally, in Chapter 5, we optimized parameter sets for a sequence processing pipeline used in the context of estimation of species richness or early detection of aquatic invasive species. We conducted simulation experiments with real 18S bulk zooplankton metabarcoded community sequence datasets, involving computationally inoculating the bulk samples

with sequences from known invaders and attempting to recover them via a common sequence processing pipeline using parameterization determined to be optimal earlier in the study.

6.2 Conclusion

The field of ALife has blossomed from basic cellular automata into the vision put forth by von Neumann and Langton, in which simulated living, reproducing, and evolving systems would complement studies of real living systems. EcoSim, one of the most advanced individual-based ALife models, yielded many insights in Chapters 3 and 4 which are parallel to observations in nature. These parallel observations further cemented its viability in usage as an experimental platform in evolutionary ecology, along with the previous works involving the simulation (e.g. Golestani, Gras, and Cristescu 2012; Mashayekhi and Gras 2012; Gras *et al.* 2015; Khater, Murariu, and Gras 2015). It also provided novel insights in Chapter 4; most importantly, we found circumstantial evidence that genetic diversity aids in the short-term establishment of introduced populations. Unexpectedly, we found that the degree of difference between native and introduced range for populations did not affect observed establishment success given genetic diversity as previously theorized (Hufbauer *et al.* 2012; Fridley and Sax 2014; Rius and Darling 2014; Estoup *et al.* 2016). Instead, the abundance and variation in spatial distribution of resources in the environment yielded stronger relationships between diversity and establishment success, corroborating studies by Hufbauer *et al.* (2013) and Szűcs *et al.* (2017) while contrasting a study by Szűcs *et al.* (2014). Further, low-diversity introduced populations were found to potentially benefit disproportionately from multiple introductions in the long term, and multiple introductions have long been proposed as a means of rescuing low-diversity populations from genetic bottleneck (Kolbe *et al.* 2004; Præbel *et al.* 2013). Introduced populations that minimized Allee effects were more successful than those that could not, which was expected based on prior theory and studies (Sakai *et al.* 2001; Kanarek and Webb 2010; Bock *et al.* 2015), and the evolutionary imbalance hypothesis (Fridley and Sax (2014) was strongly corroborated in this dissertation.

In Chapter 5, with respect to genetic sequence processing we found differences in the optimal parameter sets that users should select given the application, as proposed by Flynn *et al.* (2015); stringent filtration was required to get accurate species richness counts, while relaxed filtration was necessary to maintain the sensitivity required in order detect aquatic invasive species in the context of early detection. Furthermore, clustering and denoising processes were not viable in the context of early detection, but they were found to useful in reducing false positive errors in estimations of species richness, which supported suggestions of Brown *et al.* (2016) and Chain *et al.* (2016) regarding the

preprocessing of sequence data for these tasks. Through our simulation experiments, we observed differences in the number of inoculated species recovered by the pipelines given differences in parameter selection. Parameter sets optimized for early detection performed better at recovering taxa, as expected. However, when using the metabarcoding approach for early detection, in sequence processing an extremely delicate balance between type I and type II errors exists (Xiong, Li, and Zhan 2016). Consequently, in the current state of the associated technology, we recommend usage of the metabarcoding only as a first pass using extremely weak filtration, no clustering, and no denoising. Suspected detections should be, where possible, followed up with a targeted barcoding approach to confirm the presence of the suspected invader.

As mentioned previously, biological invasions are incredibly important to study (Colautti *et al.* 2006; Vilà *et al.* 2011; Simberloff *et al.* 2013) and the number of translocated species is not reaching saturation (Seebens *et al.* 2017). Many practical and theoretical questions in biological invasions lend themselves to being studied with modelling and simulations approaches (e.g. Travis and Dytham 2002; Klopfstein, Currat, and Excoffier 2006; Travis *et al.* 2009; Fronhofer, Poethke, and Dieckmann 2015; Henriques-Silva *et al.* 2015; Van Petegem *et al.* 2016; Yoann *et al.* 2016; Anderson and Dragićević 2018; Bonte and Bafort 2018; Day *et al.* 2018). High-performance computational systems are becoming more ubiquitous and more powerful; simulation software are taking advantage of this by becoming more sophisticated (e.g. through the use of real GIS data), more complex (by incorporating more submodels, and more computationally complex ones), and larger (handling more individuals and interactions simultaneously, and over longer time periods). The studies presented in this dissertation, for instance, used a total CPU uptime of well over ten years parallelized over thousands of jobs. In conclusion, the study of biological invasions, particularly using computer simulations, is an extremely important field that is constantly expanding and evolving with technology. Indeed, it was technology that has allowed such translocation of species so far from their native regions; now technology must be used to help us keep species where they ought to be.

6.3 Future Work

With respect to the relationship between genetic diversity and establishment success, there are numerous opportunities to be explored. As we transferred just prey individuals, species from higher trophic levels may exhibit differences in how genetic diversity aids in establishment in novel regions. Further, real introductions often contain mixed communities (e.g. ballast water; Briski *et al.* 2014); some of the relationships we observed might change if mixed communities were instead transferred between environments. We now have data for simulated invasions involving multiple

introductions; a similar study, which is called for in the literature, can be conducted in which the introductions are entirely independent in order to comparatively determine how multiple introductions affect the evolutionary trajectories of the established populations (Kolbe *et al.* 2004; Bock *et al.* 2015; Dlugosch *et al.* 2015). We can also use the data generated in this dissertation to explore differences in genetic admixture over the gradient of genetic diversity of the introduced inocula (Bock *et al.* 2015; Dlugosch *et al.* 2015); do low-diversity inocula lead to populations that retain alleles from a greater number of sources in the long term and over multiple introductions? This work represents only the first attempts at simulating biological invasions with EcoSim, and there are many eco-evolutionary theoretical questions that EcoSim can potentially help to answer.

Advancements to the technologies associated with barcoding and metabarcoding (e.g. sequencing, computational processing pipelines, etc.) will only improve the usefulness of these approaches in early detection (van Dijk *et al.* 2014; O'Rawe, Ferson, and Lyon 2015; Xiong, Li, and Zhan 2016). The related work presented in this dissertation certainly advances our understanding of the current usefulness of the approach, but in the future the dynamics of the situation might change such that metabarcoding can be trusted on its own in the context of early detection (Cristescu and Hebert 2018). As we showed, the barcoding (i.e. targeted or active) approach should be performed alongside early screening conducted in a passive manner (i.e. metabarcoding) as the balance between type I and type II error is currently too fine. Further, our study optimized and tested the computational pipelines on zooplankton bulk 18S samples that were pyrosequenced; this did allow us to garner some general insights, but it is a relatively specific case in the space of possibilities. Undoubtedly, the future of early detection of aquatic invaders lies in the usage of environmental DNA (eDNA), which presents its own unique challenges (e.g. degradation in the environment; Xiong, Li, and Zhan 2016; Cristescu and Hebert 2018). Further, other markers (i.e. COI) are of different lengths and exhibit varying degrees of diversity across taxa (Deagle *et al.* 2014; Zhan, Bailey *et al.* 2014; Hatzenbuhler *et al.* 2017). Our study involved the detection zooplankton, but similar techniques can be certainly applied to fish or any other taxa. Lastly, we used pyrosequencing, which is of the most primitive high-throughput sequencing technologies; this allowed us to present our results as a "worst-case" scenario, that is, the performance should only improve from what we observed. That all being said, there is a large space of possibilities yet to be explored using a similar optimization-via-simulation approach.

REFERENCES

Abbot P *et al.* (2010) Inclusive fitness theory and eusociality. *Nature*. doi:10.1038/nature09831

Aguilar W, Santamaría-Bonfil G, Froese T, Gershenson C (2014) The past, present, and future of artificial life. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2014.00008

Altschul SF *et al.* (1990) Basic local alignment search tool. *Journal of Molecular Biology*, 215:403-410

Anderson T, Dragicevic S (2018) Network-agent based model for simulating the dynamic spatial network structure of complex ecological systems. *Ecological Modelling*, 389:19-32

Andersson MB (1994) *Sexual selection*. Princeton University Press, Princeton

Arnold KE (2000) Group mobbing behaviour and nest defence in a cooperatively breeding Australian bird. *Ethology*, 106:385-393. doi:10.1046/j.1439-0310.2000.00545.x

Aspinall A, Gras R (2010) K-Means clustering as a speciation method within an individual-based evolving predator-prey ecosystem simulation, *6th International Conference on Active media technology*, Springer-Verlag Berlin, Heidelberg, pp. 318-329

Augusiak J, Van den Brink PJ, Grimm V (2014) Merging validation and evaluation of ecological models to 'evaludation': A review of terminology and a practical approach. *Ecological Modelling*, 280:117-128

Augustine DJ, McNaughton SJ (1998) Ungulate effects on the functional species composition of plant communities: herbivore selectivity and plant tolerance. *Journal of Wildlife Management,* 62:1165-1183

Aukema JE *et al.* (2011) Economic impacts of non-native forest insects in the continental United States. *PLoS ONE*, 6:24587. https://doi.org/10.1371/journal.pone.0024587

Banks J, Carson J, Nelson B, Nicol D (2001) *Discrete-Event System Simulation*. Prentice Hall.

Barbosa M *et al.* (2012) Fitness consequences of female multiple mating: A direct test of indirect benefits. *BMC Evolutionary Biology*, 12:185. https://doi.org/10.1186/1471-2148-12-185

Bardgett RD, Wardle DA, Yeates GW (1998) Linking above-ground and below-ground interactions: how plant responses to foliar herbivory influence soil organisms. *Soil Biology and Biochemistry*, 30:1867-1878

Bardgett RD, Streeter T, Bol R (2003) Soil microbes compete effectively with plants for organic nitrogen inputs to temperate grasslands. *Ecology*, 84:1277-1287

Barnes MA *et al.* (2014) Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science & Technology*, 48:1819-1827. doi:10.1021/es404734p

Barrett SC (2015) Foundations of invasion genetics: the Baker and Stebbins legacy. *Molecular Ecology*, 24:1927-1941. doi:10.1111/mec.13014

Bateson P (1983) *Mate choice*. Cambridge University Press, Cambridge

Beric B, MacIsaac HJ (2015) Determinants of rapid response success for alien invasive species in aquatic ecosystems. *Biological Invasions*, 17:3327-3335

Berryman AA (1992) The origins and evolution of predator-prey theory. *Ecology*, 73:1530-1535

Bertolino S (2009) Animal trade and non-indigenous species introduction: the world-wide spread of squirrels. *Diversity and Distributions*, 15:701-708. doi:10.1111/j.1472-4642.2009.00574.x

Bista I *et al.* (2017) Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. *Nature Communications*. doi:10.1038/ncomms14087

Blackburn TM *et al.* (2011) A proposed unified framework for biological invasions. *Trends in Ecology and Evolution*, 26:333-339

Blaxter KL (1989) *Energy metabolism in animals and man*. Cambridge University Press, Cambridge

Bocedi G *et al.* (2014) RangeShifter: a platform for modelling spatial eco-evolutionary dynamics and species' responses to environmental changes. *Methods in Ecology and Evolution*, 5:388-396. doi:10.1111/2041-210X.12162

Bock DG *et al.* (2015) What we still don't know about invasion genetics. *Molecular Ecology*, 24:2277-2297. doi:10.1111/mec.13032

Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*, 10:57-59

Bollache L *et al.* (2006) Spines and behaviour as defences against fish predators in an invasive freshwater amphipod. *Animal Behaviour*, 72:627-633

Bonesi L, Rushton SP, Macdonald DW (2007) Trapping for mink control and water vole survival: Identifying key criteria using a spatially explicit individual based model. *Biological Conservation*, 136:636-650. https://doi.org/10.1016/j.biocon.2007.01.008

Bonte D, Bafort Q (2019) The importance and adaptive value of life-history evolution for metapopulation dynamics. *Journal of Animal Ecology*, 88:24- 34. https://doi.org/10.1111/1365-2656.12928

Botkin D, Janak J, Wallis J (1972) Some ecological consequences of a computer model of forest growth. *Journal of Ecology*, 60:849-872. doi:10.2307/2258570

Botta-Dukát Z, Czúcz B (2016) Testing the ability of functional diversity indices to detect trait convergence and divergence using individual-based simulation. *Methods in Ecology and Evololution*, 7:114-126. doi:10.1111/2041-210X.12450

Box GEP (1979) Robustness in the strategy of scientific model building. In *Robustness in Statistics*. Academic Press, pp. 201-236

Bovy HC *et al.* (2015) Predicting the predatory impacts of the "demon shrimp" *Dikerogammarus haemobaphes*, on native and previously introduced species. *Biological Invasions*, 17:597. https://doi.org/10.1007/s10530-014-0751-9

Brännström A, Sumpter DJT (2005) The role of competition and clustering in population dynamics. *Proceeds of the Royal Society of London*. 272:2065-2072

Breiman L (2001) Random Forests. *Machine Learning*. 45:5-32. doi:10.1023/A:1010933404324

Briski E, Bailey SA, Cristescu ME, MacIsaac HJ (2010) Efficacy of 'saltwater flushing' in protecting the Great Lakes from biological invasions by invertebrate eggs in ships' ballast sediment. *Freshwater Biology*, 55:2414-2424. doi:10.1111/j.1365-2427.2010.02449.x

Briski E, Chan F, MacIsaac HJ, Bailey SA (2014) A conceptual model of community dynamics during the transport stage of the invasion process: a case study of ships' ballast. *Diversity and Distributions*, 20:236-44

Briski E *et al.* (2018) Beyond propagule pressure: importance of selection during the transport stage of biological invasion. *Frontiers in Ecology and the Environment*, 16:347-353.

Britten GL *et al.* (2014) Predator decline leads to decreased stability in a coastal fish community. *Ecology Letters*, 17:1518-1525

Brodie III ED, Brodie Jr ED (1999) Predator-prey arms races: asymmetrical selection on predators and prey may be reduced when prey are dangerous. *Bioscience*, 49:557-568

Brodie Jr ED, Ridenhour BJ, Brodie III ED (2002) The evolutionary response of predators to dangerous prey: hotspots and coldspots in the geographic mosaic of coevolution between garter snakes and newts. *Evolution*, 56:2067-2082

Broennimann O, Guisan A (2008) Predicting current and future biological invasions: both native and invaded ranges matter. *Biology Letters*, 4:585-589. doi:10.1098/rsbl.2008.0254

Brown EA *et al.* (2015) Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? *Ecology and Evolution, 5*:2234-2251

Brown EA *et al.* (2016) Early detection of aquatic invaders using metabarcoding reveals a high number of non-indigenous species in Canadian ports. *Diversity and Distributions*, 22:1045-

1059
Buckley YM, Briese DT, Rees M (2003) Demography and management of the invasive plant species *Hypericum perforatum*. II. Construction and use of an individual-based model to predict population dynamics and the effects of management strategies. *Journal of Applied Ecology*, 40:494-507. doi:10.1046/j.1365-2664.2003.00822.x
Bürger R (2000). *The mathematical theory of selection, recombination, and mutation*. Wiley, Chichester
Burton OJ, Travis JMJ (2008) The frequency of fitness peak shifts is increased at expanding range margins due to mutation surfing. *Genetics*, 179:941-950. https://doi.org/10.1534/genetics.108.087890
Burton OJ, Phillips BL, Travis JM (2010) Trade-offs and the evolution of life-histories during range expansion. *Ecology Letters*, 13:1210-1220. doi:10.1111/j.1461-0248.2010.01505.x
Butler PJ, Green JA, Boyd IL, Speakman JR (2004) Measuring metabolic rate in the field: the pros and cons of the doubly labelled water and heart rate methods. *Functional Ecology*, 18:168-183
Caporaso JG *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data, *Nature Methods*, 7:335-336
Carroll SP *et al.* (2005) And the beak shall inherit - evolution in response to invasion. *Ecology Letters*, 8:944-951. doi:10.1111/j.1461-0248.2005.00800.x
Chain FJJ, Brown EA, MacIsaac HJ, Cristescu ME (2016) Metabarcoding reveals strong spatial structure and temporal turnover of zooplankton communities among marine and freshwater ports. *Diversity and Distributions*, 22:493-504
Chapman JL, Reiss MJ (1999) *Ecology: principles and applications*. Cambridge University Press, Cambridge
Chen Y *et al.* (2018) Rapid microevolution during recent range expansion to harsh environments. *BMC Evolutionary Biology*, 18:87. doi:10.1186/s12862-018-1311-1
Chivers C, Drake DAR, Leung B (2017) Economic effects and the efficacy of intervention: exploring unintended effects of management and policy on the spread of non-indigenous species. *Biological Invasions,* 19:1795. https://doi.org/10.1007/s10530-017-1391-7
Christie MR *et al.* (2016) A single generation of domestication heritably alters the expression of hundreds of genes. *Nature Communications*. doi:10.1038/ncomms10676
Clark M, Galef B (1995) Prenatal influences on reproductive life history strategies. *Trends in Ecology & Evolution*, 10:151-153
Clarke LJ, Beard JM, Swadling KM, Deagle BE (2017) Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecology and Evolution*, 7:873-883
Clune J *et al.* (2008) Natural selection fails to optimize mutation rates for long-term adaptation on rugged fitness landscapes. *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1000187
Clune J, Goldsby HJ, Ofria C, Pennock RT (2011) Selective pressures for accurate altruism targeting: evidence from digital evolution for difficult-to-test aspects of inclusive fitness theory. *Proceeds of the Royal Society of London*, 278:666-674
Colautti RI, MacIsaac HJ (2004) A neutral terminology to define 'invasive' species. *Diversity and Distributions*, 10:135-141. doi:10.1111/j.1366-9516.2004.00061.x
Colautti RI, Grigorovich IA, MacIsaac HJ (2006) Propagule Pressure: A null model for biological invasions. *Biological Invasions*, 8:1023. https://doi.org/10.1007/s10530-005-3735-y
Colautti RI *et al.* (2006) Characterised and projected costs of nonindigenous species in Canada. *Biological Invasions*, 8:45. https://doi.org/10.1007/s10530-005-0236-y
Colautti RI, Lau JA (2015) Contemporary evolution during invasion: evidence for differentiation, natural selection, and local adaptation. *Molecular Ecology*, 24:1999-2017. doi:10.1111/mec.13162

Cole JR *et al.* (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 42:633-642
Creer S *et al.* (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology*, 19:4-20
Cristescu ME (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, 29:566-571
Cristescu ME (2015) Genetic reconstructions of invasion history. *Molecular Ecology*, 24:2212-2225. doi:10.1111/mec.13117
Cristescu ME and Hebert PDN (2018) Uses and misuses of environmental DNA in biodiversity science and conservation. *Annual Reviews of Ecology, Evolution, and Systematics*, 49:209-230
Currat M, Ray N, Excoffier L (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes*, 4:139-142. doi:10.1046/j.1471-8286.2003.00582.x
Davies TJ *et al.* (2004) Environmental energy and evolutionary rates in flowering plants. *Proceeds of the Royal Society of London*, 271:2195-2200
Day CC *et al.* (2018) Using simulation modeling to inform management of invasive species: A case study of eastern brook trout suppression and eradication. *Biological Conservation*, 221:10-22
Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters,* 10. doi:10.1098/rsbl.2014.0562
DeAngelis DL, Grimm V (2014) Individual-based models in ecology after four decades. *F1000Prime Rep*, 6:39
de Jager M, Bartumeus F, Kölzsch A *et al.* (2013) How superdiffusion gets arrested: ecological encounters explain shift from Lévy to Brownian movement. *Proceeds of the Royal Society of London*. doi:10.1098/rspb.2013.2605
de Los Santos CB, Neuparth T, Torres T *et al.* (2015) Ecological modelling and toxicity data coupled to assess population recovery of marine amphipod Gammarus locusta: Application to disturbance by chronic exposure to aniline. *Aquatic Toxicology*, 163:60-70
Devaurs D, Gras R (2010) Species abundance patterns in an ecosystem simulation studied through Fisher's logseries. *Simulation Modelling Practice and Theory*, 18:100-123
De Ventura L, Kopp K, Seppälä K, Jokela J (2017) Tracing the quagga mussel invasion along the Rhine river system using eDNA markers: early detection and surveillance of invasive zebra and quagga mussels. *Management of Biological Invasions*, 8:101-112. https://doi.org/10.3391/mbi.2017.8.1.10
Dirzo R, Raven PH (2003) Global state of biodiversity and loss. *Annual Review of Environment and Resources*, 28:137-167. doi:10.1146/annurev.energy.28.050302.105532
Dlugosch KM, Parker IM (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology*, 17:431-449. doi:10.1111/j.1365-294X.2007.03538.x
Dlugosch KM *et al.* (2015) The devil is in the details: genetic variation in introduced populations and its contributions to invasion. *Molecular Ecology*, 24:2095-2111. doi:10.1111/mec.13183
Dodd JA *et al.* (2014) Predicting the ecological impacts of a new freshwater invader: Functional responses and prey selectivity of the 'killer shrimp', *Dikerogammarus villosus*, compared to the native *Gammarus pulex*. *Freshwater Biology*, 59:337-352. https://doi.org/10.1111/fwb.12268
Drent RH, Van der Wal R (1999) Cyclic grazing in vertebrates and the manipulation of the food resource. In: Olff H, Brown VK, Drent R H (eds), *Herbivores: between plants and predators*. Blackwell, London, pp. 271-299

Duggan IC, Rixon CAM, MacIsaac HJ (2006) Popularity and propagule pressure: Determinants of introduction and establishment of aquarium fish. *Biological Invasions*, 8:377. https://doi.org/10.1007/s10530-004-2310-2
Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26:2460-2461
Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, 10:996-998
Edgar RC (2016) UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. Retrieved from *bioRxiv* 081257. doi:10.1101/081257
Eklöf J, Šuba J, Petersons G, Rydell J. (2014) Visual acuity and eye size in five European bat species in relation to foraging and migration strategies. *Environmental Experimental Biology*, 12:1-6
Elbrecht V, Leese F (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS One*. doi:10.1371/journal.pone.0130324
Estoup A *et al.* (2016) Is there a genetic paradox of biological invasion?. *Annual Review of Ecology, Evolution, and Systematics*, 47:51-72. http://dx.doi.org/10.1146/annurev-ecolsys-121415-032116
Falk D (1990) Brain evolution in Homo: The 'radiator' theory. *Behavioural and Brain Sciences*, 13:333-344
Ficetola GF, Miaud C, Pompanon F, Taberlet P (2008) Species detection using environmental DNA from water samples. *Biology Letters*, 4:423-425
Ficetola GF *et al.* (2010) An in silico approach for the evaluation of DNA barcodes. *BMC Genomics*, 11. doi:10.1186/1471-2164-11-434
Flynn JM *et al.* (2015) Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, 5:2252-2266
Fonseca VG *et al.* (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*. doi:10.1038/ncomms1095
Forsman A (2014) Effects of genotypic and phenotypic variation on establishment are important for conservation, invasion, and infection biology. *Proceedings of the National Academy of Sciences*, 111:302-307. doi:10.1073/pnas.1317745111
Fortuna MA, Zaman L, Wagner AP, Ofria C (2013) Evolving digital ecological networks. *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1002928
Frank BM, Baret PV (2013) Simulating brown trout demogenetics in a river/nursery brook system: The individual-based model DemGenTrout. *Ecological Modelling*, 248:184-202
Frank D, Evans R (1997) Effects of native grazers on N cycling in a north-temperate grassland ecosystem: Yellowstone National Park. *Ecology*, 78:2238-2249
Frank D, Groffman P (1998) Ungulate vs. landscape control of soil C and N processes in grasslands of Yellowstone National Park. *Ecology*, 79:2229-2241
Freeland JR (2017) The importance of molecular markers and primer design when characterizing biodiversity from environmental DNA. *Genome*, 60:358-374. doi:10.1139/gen-2016-0100
Fridley JD, Sax DF (2014) Darwinian framework for invasion biology. *Global Ecology and Biogeography*, 23:1157-1166. doi:10.1111/geb.12221
Friman VP, Hiltunen T, Laakso J, Kaitala V (2008) Availability of prey resources drives evolution of predator-prey interaction. *Proceeds of the Royal Society of London*, 275:1625-1633
Fronhofer EA, Poethke HJ, Dieckmann U (2015) Evolution of dispersal distance: Maternal investment leads to bimodal dispersal kernels. *Journal of Theoretical Biology*, 365:270-279
Garamszegi LZ, Møller AP, Erritzøe J (2002) Coevolving avian eye size and brain size in relation to prey capture and nocturnality. *Proceeds of the Royal Society of London*, 269:961-967

Gardner M (1970) Mathematical Games: The fantastic combinations of John Conway's new solitaire game "Life". *Scientific American*. 223:120-123.
Genovesi P (2005) Eradications of invasive alien species in Europe: a review. *Biological Invasions*, 7:127. https://doi.org/10.1007/s10530-004-9642-9
Gillooly JF, Allen AP, West GB, Brown JH (2005) The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proceedings of the National Academy of Sciences*, 102:140-145
Goldsby HJ, Knoester DB, Ofria C, Kerr B (2014) The evolutionary origin of somatic cells under the dirty work hypothesis. *PLoS ONE*. doi:10.1371/journal.pbio.1001858
Goldsmit J *et al.* (2018) Projecting present and future habitat suitability of ship-mediated aquatic invasive species in the Canadian Arctic. *Biological Invasions*, 20:501. https://doi.org/10.1007/s10530-017-1553-7
Golestani A, Gras R (2010) Regularity analysis of an individual-based ecosystem simulation. *Chaos*, 20:3120
Golestani A, Gras R (2011) Multifractal phenomena in EcoSim, a large scale individual-based ecosystem simulation, *International Conference on Artificial Intelligence*, Las Vegas, 991-999
Golestani A, Gras R (2012) Identifying origin of self-similarity in EcoSim, an individual-based ecosystem simulation, using wavelet-based multifractal analysis. *Proceedings of the world congress on engineering and computer science 2012 (WCECS 2012)*, San Francisco, pp. 1275-1282
Golestani A, Gras R, Cristescu M (2012) Speciation with gene flow in a heterogeneous virtual world: can physical obstacles accelerate speciation? *Proceeds of the Royal Society of London*, 279:3055-3064
Goslee SC, Peters DPC, Beck KG (2006) Spatial prediction of invasion success across heterogeneous landscapes using an individual-based model. *Biological Invasions*, 8:193. https://doi.org/10.1007/s10530-004-2954-y
Grant PR, Grant BR (2006) Evolution of character displacement in Darwin's finches. *Science*, 313:224-226
Gras R, Devaurs D, Wozniak A, Aspinall A (2009) An individual-based evolving predator-prey ecosystem simulation using a fuzzy cognitive map as the behavior model. *Artificial Life*, 15:423-463
Gras R, Golestani A, Hendry AP, Cristescu ME (2015) Speciation without pre-defined fitness functions. *PLoS ONE*. doi:10.1371/journal.pone.0137838
Grimm V *et al.* (2005) Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, 310:987-991. doi:10.1126/science.1116681
Grimm V *et al.* (2006) A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, 198:115-126
Grimm V *et al.* (2010) The ODD protocol: A review and first update. *Ecological Modelling*, 221:2760-2768
Grimm V, Railsback SF (2013) *Individual-based modeling and ecology*. Princeton University Press, Princeton, New Jersey
Grimm V *et al.* (2014) Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecological Modelling*, 280:129-139
Hazlerigg CRE *et al.* (2014) Population relevance of toxicant mediated changes in sex ratio in fish: An assessment using an individual-based zebrafish (*Danio rerio*) model. *Ecological Modelling*, 280:76-88
Hamilton E, Frank D (2001) Can plants stimulate soil microbes and their own nutrient supply? Evidence from a grazing tolerant grass. *Ecology*, 82:2397-2402
Hänfling B *et al.* (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, 25:3101-3119

Hao X, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* 27:611-618
Hargreaves AL, Eckert CG (2014) Evolution of dispersal and mating systems along geographic gradients: implications for shifting ranges. *Functional Ecology*, 28:5-21. doi:10.1111/1365-2435.12170
Hartl DL, Jones EW (2004) *Genetics: analysis of genes and genomes*. Jones & Bartlett Publishers, Burlington
Harvey CT, Qureshi SA, MacIsaac HJ (2009) Detection of a colonizing, aquatic, non-indigenous species. *Diversity and Distributions*, 15:429-437. doi:10.1111/j.1472-4642.2008.00550.x
Hatzenbuhler C *et al.* (2017). Sensitivity and accuracy of high-throughput metabarcoding methods for early detection of invasive fish species. *Scientific Reports*, 7:46393. doi:10.1038/srep46393
Hebert PD, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270:313-321
Hemmingsen AM (1960) Energy metabolism as related to body size and respiratory surfaces, and its evolution. *Reports of the Steno Memorial Hospital and Nordisk Insulin Laboratorium* 9:1-110
Henriques-Silva R *et al.* (2015) On the evolution of dispersal via heterogeneity in spatial connectivity. *Proceedings of the Royal Society B: Biological Sciences*, 282:20142879. doi:10.1098/rspb.2014.2879
Herborg L, O'Hara P, Therriault TW (2009) Forecasting the potential distribution of the invasive tunicate *Didemnum vexillum*. *Journal of Applied Ecology*, 46:64-72. doi:10.1111/j.1365-2664.2008.01568.x
Higgins SI, Richardson DM, Cowling RM (1996) Modeling invasive plant spread: the role of plant-environment interactions and model structure. *Ecology*, 77:2043-2054. doi:10.2307/2265699
Higgins SI, Richardson DM (1998) Pine invasions in the southern hemisphere: modelling interactions between organism, environment and disturbance. *Plant Ecology*, 135:79. https://doi.org/10.1023/A:1009760512895
Higgins S, Richardson D (1999) Predicting plant migration rates in a changing world: The role of long-distance dispersal. *The American Naturalist*, 153:464-475. doi:10.1086/303193
Higgins SI, Richardson DM, Cowling RM (2000) Using a dynamic landscape model for planning the management of alien plant invasions. *Ecological Applications*, 10:1833-1848
Hik DS, Jefferies RL (1990) Increases in the net aboveground primary production of a salt-marsh forage grass: a test of the predictions of the herbivore-optimization model. *Journal of Ecology*, 78:180-195
Hiltunen T, Ayan GB, Becks L (2015) Environmental fluctuations restrict eco-evolutionary dynamics in predator-prey system. *Proceeds of the Royal Society of London*. doi:10.1098/rspb.2015.0013
Hobbs NT (1996) Modification of ecosystems by ungulates. *Journal of Wildlife Management*, 60:695-713
Hoffman JC *et al.* (2011) Effort and potential efficiencies for aquatic non-native species early detection. *Canadian Journal of Fisheries and Aquatic Sciences*, 68:2064-2079
Holland JH (1992) Genetic Algorithms. *Scientific American*, 267:66-73
Holland J (1998) Emergence: from chaos to order. Addison-Wesley, Redwood City
Hornby GS, Globus A, Linden DS, Lohn JD (2006). Automated antenna design with evolutionary algorithms. *American Institute of Aeronautics and Astronautic*s.
Hoskin CJ, Higgie M, McDonald KR, Moritz C (2005) Reinforcement drives rapid allopatric speciation. *Nature*, 437:1353
Hraber PT, Jones T, Forrest S (1997) The ecology of Echo. *Artificial Life*, 3:165-190

Hufbauer RA *et al.* (2012) Anthropogenically induced adaptation to invade (AIAI): Contemporary adaptation to human-altered habitats within the native range can promote invasions. *Evolutionary Applications*, 5:89-101
Hufbauer RA *et al.* (2013) Role of propagule pressure in colonization success: disentangling the relative importance of demographic, genetic and habitat effects. *Journal of Evolutionary Biology*, 26:1691-99
Hulme PE (2009) Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, 46:10-18. doi:10.1111/j.1365-2664.2008.01600.x
Humair F, Humair L, Kuhn F, Kueffer C (2015) E-commerce trade in invasive plants. *Conservation Biology*, 29:1658-1665. doi:10.1111/cobi.12579
Jasienska G (2003) Energy metabolism and the evolution of reproductive suppression in the human female. *Acta Biotheoretica*, 51:1-8
Jerde CL, Mahon AR, Chadderton WL, Lodge DM (2011) "Sight-unseen" detection of rare aquatic species using environmental DNA. Conservation Letters, 4:150-157. doi:10.1111/j.1755-263X.2010.00158.x
Kanarek AR, Webb CT (2010) Allee effects, adaptive evolution, and invasion success. *Evolutionary Applications*, 3:122-135. doi:10.1111/j.1752-4571.2009.00112.x
Kantz H, Schreiber T (1997) *Nonlinear time series analysis*. Cambridge University Press, Cambridge
Karim Pour M, Bandehbahman S, Gras R, Cristescu M (2017) An individual-based modeling approach to investigate sympatric speciation via specialized resource usage. *Open Journal of Ecology*, 7:222-269
Keith JM, Spring D (2013) Agent-based Bayesian approach to monitoring the progress of invasive species eradication programs. *Proceedings of the National Academy of Sciences*, 110:13428-13433
Khater M, Murariu D, Gras R (2014) Contemporary evolution and genetic change of prey as a response to predator removal. *Ecological Informatics*, 22:13-22
Khater M, Salehi E, Gras R (2011) Correlation between genetic diversity and fitness in a predator-prey ecosystem simulation, *24th Australian Joint Conference on Artificial Intelligence,* Perth, Australia, pp. 422-431.
Khater M, Gras R. (2012) Adaptation and genomic evolution in EcoSim. In: Ziemke T, Balkenius C, Hallam J (eds) *From Animals to Animats 12, Proceedings of the 12th International Conference on Simulation of Adaptive Behavior*, SAB 2012, Odense, Denmark, pp. 219-229
Kiltie RA (2000) Scaling of visual acuity with body size in mammals and birds. *Functional Ecology*, 14:226-234
Kimura M (1965) A stochastic model concerning the maintenance of genetic variability in quantitative characters. *Proceedings of the National Academy of Sciences*, 54:731-736.
Kleiber M (1932) Body size and metabolism. Hilgardia, 6:315-353. doi:10.3733/hilg.v06n11p315
Kleiber M (1961) *The fire of life. An introduction to animal energetics*. Wiley, New York
Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion, *Molecular Biology and Evolution*, 23:482-490. https://doi.org/10.1093/molbev/msj057
Kolar CS, Lodge DM (2001) Progress in invasion biology: predicting invaders. *Trends in Ecology & Evolution*, 16:199-204. doi:https://doi.org/10.1016/S0169-5347(01)02101-2
Kolbe JJ *et al.* (2004) Genetic variation increases during biological invasion by a Cuban lizard. *Nature*, 431:177-181
Kosko B (1986) Fuzzy cognitive maps. *International Journal of Man-Machine Studies*, 24:65-75
Krams I, Krama T, Igaune K (2006) Mobbing behaviour: reciprocity-based co-operation in breeding Pied Flycatchers *Ficedula hypoleuca*. *Ibis*, 148:50-54
Krams I, Krama T, Igaune K, Mänd R. Experimental evidence of reciprocal altruism in the pied flycatcher. *Behavioral Ecology and Sociobiology*, 62:599-605

Krause KE, Dinh KV, Nielsen TG (2017) Increased tolerance to oil exposure by the cosmopolitan marine copepod *Acartia tonsa*. *Science of the Total Environment*, 607/608: 87-94
Krebs J, Davies N (1997) *Behavioural ecology: an evolutionary approach*, 4th edn. Blackwell Publishers, Oxford
Kvam P *et al.* (2013) Computational Evolution of Decision-Making Strategies. In: Noelle DC, Dale R, Warlaumont AS *et al.* (eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, Austin, TX, pp. 1225-1230
LaBar T, Hintze A, Adami C (2016) Evolvability tradeoffs in emergent digital replicators. *Artificial Life*, 22:483-498
Landguth EL, Cushman SA (2010) CDPOP: A spatially explicit cost distance population genetics program. *Molecular Ecological Resources*, 10:156-161
Landguth EL, Bearlin A, Day CC, Dunham J (2017) CDMetaPOP: an individual-based, eco-evolutionary model for spatially explicit simulation of landscape demogenetics. *Methods in Ecology and Evolution*, 8:4-11
Laverty C *et al.* (2015) Differential ecological impacts of invader and native predatory freshwater amphipods under environmental change are revealed by comparative functional responses. *Biological Invasions,* 17:1761. https://doi.org/10.1007/s10530-014-0832-9
Lawson Handley LJ *et al.* (2011) Ecological genetics of invasive alien species. *BioControl*, 56:409-428. https://doi.org/10.1007/s10526-011-9386-2
Leighton PA *et al.* (2012) Predicting the speed of tick invasion: an empirical model of range expansion for the Lyme disease vector *Ixodes scapularis* in Canada. *Journal of Applied Ecology*, 49:457-464. doi:10.1111/j.1365-2664.2012.02112.x
Lenski RE, Ofria C, Collier TC, Adami C (1999) Genome complexity, robustness and genetic interactions in digital organisms. *Nature*, 400:661-664
Lenski RE, Ofria C, Pennock RT, Adami C (2003) The evolutionary origin of complex features. *Nature*, 423:139-144
Leung B, Drake JM, Lodge DM (2004) Predicting invasions: Propagule pressure and the gravity of Allee effects. *Ecology*, 85:1651-1660
Lewis RJ, Kappeler PM (2005) Seasonality, body condition, and timing of reproduction in *Propithecus verreauxi verreauxi* in the Kirindy Forest. American Journal of *Primatology*, 67:347-364
Li Y, Brose U, Meyer K, Rall BC (2017) How patch size and refuge availability change interaction strength and population dynamics: a combined individual- and population-based modeling experiment. *PeerJ*. doi:10.7717/peerj.2993
Leonard WR, Ulijaszek SJ (2002) Energetics and evolution: an emerging research domain. *American Journal of Human Biology*, 14:547-550
Lodge DM (1993) Biological invasions: Lessons for ecology. *Trends in Ecology & Evolution*, 8:133-137. https://doi.org/10.1016/0169-5347(93)90025-K
Lustenhouwer N, Williams JL, Levine JM (2019) Evolution during population spread affects plant performance in stressful environments. *Journal of Ecology*, 107:396-406. https://doi.org/10.1111/1365-2745.13045
MacPherson B, Mashayekhi M, Scott R, Gras R (2017) Exploring the connection between emergent animal personality and fitness using a novel individual-based model and decision tree approach. *Ecological Informatics*, 40:81-92.
McKinney ML, Lockwood JL (1999) Biotic homogenization: a few winners replacing many losers in the next mass extinction. *Trends in Ecology & Evolution*, 14:450-453. doi:https://doi.org/10.1016/S0169-5347(99)01679-1
MacPherson B, Gras R (2016) Individual-based ecological models: Adjunctive tools or experimental systems? *Ecological Modelling*, 323:106-114
Mahé F *et al.* (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2:e593

Mallet J (1995) A species definition for the modern synthesis. *Trends in Ecology & Evolution* 10:294-299
Marcot BG, Singleton PH, Schumaker NH (2015) Analysis of sensitivity and uncertainty in an individual-based model of a threatened wildlife species. *Natural Resource Modeling*, 28:37-58. doi:10.1111/nrm.12056
Marini G *et al.* (2017) The effect of interspecific competition on the temporal dynamics of *Aedes albopictus* and *Culex pipiens*. Parasites & Vectors, 10:102
Marshall JA (2016) What is inclusive fitness theory, and what is it for? *Current Opinion in Behavioral Sciences*, 12:103-108
Mashayekhi M, Gras R (2012) Investigating the effect of spatial distribution and spatiotemporal information on speciation using individual-based ecosystem simulation, *GSTF Journal on Computing*, 2:98-103
Mashayekhi M, MacPherson B, Gras R (2014) Species-area relationship and a tentative interpretation of the function coefficients in an ecosystem simulation. *Ecological Complexity*, 19:84-95
Mashayekhi M, MacPherson B, Gras R (2014) A machine learning approach to investigate the reasons behind species extinction. *Ecological Informatics*, 20:58-66
May A *et al.* (2014) Unraveling the outcome of 16S rDNA-based taxonomy analysis through mock data and simulations. *Bioinformatics*, 30:1530-1538
McNab BK (2002) Minimizing energy expenditure facilitates vertebrate persistence on oceanic islands, *Ecology Letters*, 5:693-704
McNaughton S, Banyikwa FF, McNaughton MM (1997) Promotion of the cycling of diet-enhancing nutrients by African grazers. *Science*, 278:1798-1800
Mech SG, Zollner PA (2002) Using body size to predict perceptual range. *Oikos*, 98:47-52
Mergeay J, Verschuren D, De Meester L (2006) Invasion of an asexual American water flea clone throughout Africa and rapid displacement of a native sibling species. *Proceedings of the Royal Society B: Biological Sciences*, 273:2839-2844. http://doi.org/10.1098/rspb.2006.3661
Møller AP (2009) Basal metabolic rate and risk-taking behaviour in birds. *Journal of Evolutionary Biology*, 22:2420-2429
Molvar EM, Bowyer RT, Van Ballenberghe V (1993) Moose herbivory, browse quality, and nutrient cycling in an Alaskan treeline community. *Oecologia*, 94:473-479
Mönkkönen M, Forsman JT, Bokma F (2006) Energy availability, abundance, energy-use and species richness in forest bird communities: a test of the species-energy theory. *Global Ecology and Biogeography*, 15:290-302
Mueller P, Diamond J (2001) Metabolic rate and environmental productivity: Well-provisioned animals evolved to run and idle fast. *Proceedings of the National Academy of Sciences*, 98:12550-12554
Muirhead JR, MacIsaac HJ (2005) Development of inland lakes as hubs in an invasion network. *Journal of Applied Ecology*, 42:80-90. doi:10.1111/j.1365-2664.2004.00988.x
Murphy JT, Voisin M, Johnson M, Viard, F (2016) Abundance and recruitment data for *Undaria pinnatifida* in Brest harbour, France: Model versus field results. *Data in Brief*, 7:540-545. doi:10.1016/j.dib.2016.02.075
Nagy KA (2005) Field metabolic rate and body size. *Journal of Experimental Biology*, 208:1621-1625
Navarrete A, van Schaik CP, Isler K (2011) Energetics and the evolution of human brain size. *Nature*, 480:91
Nguyen *et al.* (2011) On weather affecting to brown plant hopper invasion using an agent-based model. *MEDES '11: International ACM Conference on Management of Emergent Digital EcoSystems*, San Francisco, CA, USA, November 21-24, 2011

Niklas KJ, Enquist BJ (2001) Invariant scaling relationships for interspecific plant biomass production rates and body size. *Proceedings of the National Academy of Sciences*, 98:2922-2927
Niven JE, Laughlin SB (2008) Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology*, 211:1792-1804
Nowak MA, Tarnita CE, Wilson EO (2010) The evolution of eusociality. *Nature*, 466:1057-1062
O'Rawe JA, Ferson S, Lyon G (2015) Accounting for uncertainty in DNA sequencing data. *Trends in Genetics*, 31:61-66
Ofria C, Wilke CO (2004) Avida: a software platform for research in computational evolutionary biology. *Artificial Life*, 10:191-229
Olff H, Ritchie ME (1998) Effects of herbivores on grassland plant diversity. *Trends in Ecology & Evolution*, 13:261-265
Olson RS *et al.* (2013) Predator confusion is sufficient to evolve swarming behavior. *Journal of the Royal Society Interface*. doi:10.1098/rsif.2013.0305
Ostrowski EA, Ofria C, Lenski RE (2015) Genetically integrated traits and rugged adaptive landscapes in digital organisms. *BMC Ecology*. doi:10.1186/s12862-015-0361-x
Pafilis P, Meiri S, Foufopoulos J, Valakos E (2009) Intraspecific competition and high food availability are associated with insular gigantism in a lizard. *Naturwissenschaften*, 96:1107-13
Pawlowski J *et al.* (2014) Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Molecular Ecology Resources*, 14:1129-1140. doi:10.1111/1755-0998.12261
Pedley TJ (1977) Scale effects in animal locomotion. *Quarterly Reviews of Biology*, 53:473-474
Peischl S, Excoffier L (2015) Expansion load: recessive mutations and the role of standing genetic variation. *Molecular Ecology*, 24:2084-2094. doi:10.1111/mec.13154
Peters RH (1986) *The ecological implications of body size*. Cambridge University Press, Cambridge
Pethybridge H *et al.* (2013) Responses of European anchovy vital rates and population growth to environmental fluctuations: An individual-based modeling approach. *Ecological Modelling*, 250:370-383
Phan CH, Huynh HX, Drogoul A (2010) An agent-based approach to the simulation of brown plant hopper (BPH) invasions in the Mekong Delta. *RIVF 2010: IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future*, Hanoi, Vietnam, November 1-4, 2010
Phillips BL, Brown GP, Webb JK, Shine R (2006) Invasion and the evolution of speed in toads. *Nature*, 439:803. doi:10.1038/439803a
Phillips BL, Brown GP, Shine R (2010) Life-history evolution in range-shifting populations. *Ecology*, 91:1617-1627. doi:10.1890/09-0910.1
Piana PA, Gomes LC, Agostinho AA (2006) Comparison of predator-prey interaction models for fish assemblages from the neotropical region. *Ecological Modelling*, 192:259-270
Piggott MP (2016) Evaluating the effects of laboratory protocols on eDNA detection probability for an endangered freshwater fish. *Ecology and Evolution*, 6:2739-2750
Pluess T *et al.* (2012) Which factors affect the success or failure of eradication campaigns against alien species?. *PLoS ONE*, 7;e48157. doi:10.1371/journal.pone.0048157
Pochon X *et al.* (2017) Wanted dead or alive? Using metabarcoding of environmental DNA and RNA to distinguish living assemblages for biosecurity applications. *PLoS ONE*, 12:e0187636. https://doi.org/10.1371/journal.pone.0187636
Port JA *et al.* (2016) Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, 25:527-541
Potier S *et al.* (2016) Visual abilities in two raptors with different ecology. *Journal of Experimental Biology*, 219:2639-2649

Præbel K, Gjelland KØ, Salonen E, Amundsen PA (2013) Invasion genetics of vendace (*Coregonus albula (L.)*) in the Inari-Pasvik watercourse: revealing the origin and expansion pattern of a rapid colonization event. *Ecology and evolution*, 3:1400-1412. doi:10.1002/ece3.552

Prothero JW (1979) Maximal oxygen consumption in various animals and plants. *Comparative Biochemistry and Physiology - Part A: Molecular & Integrative Physiology*, 64:463-466

Pyšek P, Richardson DM (2010) Invasive species, environmental change and management, and health. *Annual Review of Environment and Resources*, 35:25-55

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: The barcode index number (BIN) system. *PLoS ONE*, 8:e66213. https://doi.org/10.1371/journal.pone.0066213

Ray TS (1991) An approach to the synthesis of life. In: Langton C, Taylor C, Farmer JD, Rasmussen S (eds) *Proceedings of Artificial Life II*, Addison-Wesley, Redwood City, pp. 371-408

Ricciardi A, MacIsaac HJ (2000) Recent mass invasion of the North American Great Lakes by Ponto Caspian species. *Trends in Ecology & Evolution*, 15:62-65

Richards CL *et al.* (2006) Jack of all trades, master of some? On the role of phenotypic plasticity in plant invasions. *Ecology Letters*, 9:981-993

Ricotta C (2000) From theoretical ecology to statistical physics and back: self-similar landscape metrics as a synthesis of ecological diversity and geometrical complexity. *Ecological Modelling*, 125:245-253

Rius M, Darling JA (2014) How important is intraspecific genetic admixture to the success of colonising populations? *Trends in Ecology & Evolution*, 29:233-242. https://doi.org/10.1016/j.tree.2014.02.003

Rodda GH, Jarnevich CS, Reed RN (2008) What parts of the US mainland are climatically suitable for invasive alien pythons spreading from Everglades National Park? *Biological Invasions*, 11:241-252

Roman J, Darling JA (2007) Paradox lost: genetic diversity and the success of aquatic invasions. *Trends in Ecology & Evolution*, 22:454-464. doi:https://doi.org/10.1016/j.tree.2007.07.002

Rutowski RL, Gislén L, Warrant EJ (2009) Visual acuity and sensitivity increase allometrically with body size in butterflies. *Arthropod Structure & Development*, 38:91-100

Sakai A *et al.* (2001) The population biology of invasive species. *Annual Review of Ecology and Systematics* 32:305-332

Safi K, Seid MA, Dechmann DKN (2005) Bigger is not always better: when brains get smaller. *Biology Letters*, 1:283-286

Samson E *et al.* (2017) Early engagement of stakeholders with individual-based modelling can inform research for improving invasive species management: the round goby as a case study. *Frontiers in Ecology and Evolution*, 5:149. doi:10.3389/fevo.2017.00149

Santini L *et al.* (2016) A trait-based approach for predicting species responses to environmental change from sparse data: how well might terrestrial mammals track climate change? *Global Change Biology*, 22:2415-2424

Sax DF *et al.* (2007) Ecological and evolutionary insights from species invasions. *Trends in Ecology & Evolution*, 22:465-471. doi:https://doi.org/10.1016/j.tree.2007.06.009

Schloss PD *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75:7537-7541

Schmidt-Nielsen K (1984) *Scaling: Why is animal size so important?* Cambridge University Press, Cambridge

Schmolke A, Thorbek P, DeAngelis DL, Grimm V (2010) Ecological models supporting environmental decision making: a strategy for the future. *Trends in Ecology & Evolution* 25:479-486

Schultz MP, Bendick JA, Holm ER, Hertel WM (2011) Economic impact of biofouling on a naval surface ship, *Biofouling*, 27:87-98, doi:10.1080/08927014.2010.542809
Scott R, MacPherson B, Gras R (2018) A comparison of stable and fluctuating resources with respect to evolutionary adaptation and life-history traits using individual-based modeling and machine learning. *Journal of Theoretical Biology*, 459:52-66.
Seebens H *et al.* (2017) No saturation in the accumulation of alien species worldwide. *Nature Communications*, 8:14435. doi:10.1038/ncomms14435
Seuront L *et al.* (1996) Multifractal analysis of phytoplankton biomass and temperature in the ocean. *Geophysical Research Letters*, 23:3591-3594
Shanmuganathan T *et al.* (2010) Biological control of the cane toad in Australia: a review. *Animal Conservation*, 13:16-23. doi:10.1111/j.1469-1795.2009.00319.x
Shaw AK, Kokko H, Neubert MG (2018) Sex difference and Allee effects shape the dynamics of sex-structured invasions. *Journal of Animal Ecology*, 87:36-46. https://doi.org/10.1111/1365-2656.12658
Shepherd GM (1994) *Neurobiology*. Oxford University Press, Oxford
Shochat E *et al.* (2010) Invasion, competition, and biodiversity loss in urban ecosystems. *BioScience*, 60:199-208, https://doi.org/10.1525/bio.2010.60.3.6
Siemann E, Rogers WE (2001) Genetic differences in growth of an invasive tree species. *Ecology Letters*, 4:514-518. doi:10.1046/j.1461-0248.2001.00274.x
Simberloff D (2009) The role of propagule pressure in biological invasions. *Annual Review of Ecology, Evolution, and Systematics*, 40:81-102. https://doi.org/10.1146/annurev.ecolsys.110308.120304
Simberloff D *et al.* (2013) Impacts of biological invasions: what's what and the way forward. *Trends in Ecology & Evolution*, 28:58-66. doi:https://doi.org/10.1016/j.tree.2012.07.013
Simmons AD, Thomas CD (2004) Changes in dispersal during species' range expansions. *The American Naturalist*, 164:378-395
Simmons M *et al.* (2016) Active and passive environmental DNA surveillance of aquatic invasive species. *Canadian Journal of Fisheries and Aquatic Sciences*, 73:76-83.doi:10.1139/cjfas-2015-0262
Smart AS *et al.* (2015) Environmental DNA sampling is more sensitive than a traditional survey technique for detecting an aquatic invader. *Ecological Applications*, 25:1944-1952
Stahl WRR (1965) Organ weights in primates and other mammals. *Science*, 150:1039-1042
Stahl WRR (1967) Scaling of respiratory variables in mammals. *Journal of Applied Physiology*, 22:453-460
Stephens D, Krebs J (1986) *Foraging theory*. Princeton University Press, Princeton
Strauss SY, Lau JA, Carroll SP (2006) Evolutionary responses of natives to introduced species: what do introductions tell us about natural communities? *Ecology Letters*, 9:357-374
Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, 8:25. doi:10.1186/1471-2105-8-25
Suarez AV, Holway DA, Tsutsui ND (2008) Genetics and behavior of a colonizing species: The invasive Argentine ant. *The American Naturalist*, 172:72-84
Sutter M, Kawecki TJ (2009) Influence of learning on range expansion and adaptation to novel habitats. *Journal of Evolutionary Biology*, 22:2201-2214. doi:10.1111/j.1420-9101.2009.01836.x
Svanbäck R, Bolnick DI (2007) Intraspecific competition drives increased resource use diversity within a natural population. *Proceeds of the Royal Society of London*, 274:839-844
Svanbäck R, Eklöv P, Fransson R, Holmgren K (2008) Intraspecific competition drives multiple species resource polymorphism in fish communities. *Oikos*, 117:114-124
Szűcs M, Melbourne BA, Tuff T, Hufbauer RA (2014) The roles of demography and genetics in the early stages of colonization. *Proceedings of the Royal Society of London*, 281:1-8

Szűcs M, Melbourne BA, Tuff T, Weiss-Lehman C, Hufbauer RA (2017) Genetic and demographic founder effects have long-term fitness consequences for colonizing populations. *Ecology Letters*, 20:436-444

Tang CQ *et al.* (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences*, 109:16208-16212

Taylor CM, Hastings A (2005) Allee effects in biological invasions. *Ecology Letters*, 8:895-908. doi:10.1111/j.1461-0248.2005.00787.x

Thearling K, Ray T (1994) Evolving multi-cellular artificial life. In: Brooks RA, Maes P (eds) *Proceedings of Artificial Life IV*, MIT Press, Cambridge, pp. 283-288

The HDF Group (2000) Hierarchical data format version 5. http://www.hdfgroup.org/HDF5. Accessed Feb 2014

Thuiller W *et al.* (2005) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, 11:2234-2250. doi:10.1111/j.1365-2486.2005.001018.x

Travis JMJ, Dytham C (2002) Dispersal evolution during invasions. *Evolutionary Ecology Research*, 4:1119-1129

Travis JMJ *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. https://doi.org/10.1093/molbev/msm167

Travis JMJ, Mustin K, Benton TG, Dytham C (2009) Accelerating invasion rates result from the evolution of density-dependent dispersal. *Journal of Theoretical Biology*, 259:151-158. doi:10.1016/j.jtbi.2009.03.008

Travis JM, Harris CM, Park KJ, Bullock JM (2011) Improving prediction and management of range expansions by combining analytical and individual-based modelling approaches. *Methods in Ecology and Evolution*, 2:477-488. doi:10.1111/j.2041-210X.2011.00104.x

Uchmański J (2016) Individual variability and metapopulation dynamics: An individual-based model. *Ecological Modelling*, 334:8-18

van Dijk EL, Auger H, Jaszcyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. *Trends in Genetics*, 30, 418-426

Van Petegem KHP, Boeye J, Stoks R, Bonte D (2016) Spatial selection and local adaptation jointly shape life-history evolution during range expansion. *The American Naturalist*, 188:485-498

Vilà M *et al.* (2011) Ecological impacts of invasive alien plants: a meta-analysis of their effects on species, communities and ecosystems. *Ecology Letters*, 14:702-708. doi:10.1111/j.1461-0248.2011.01628.x

von Neumann J, Burks AW (1966) Theory of Self-Reproducing Automata. University of Illinois Press.

Wang X, Yao J, Sun Y, Mai V. (2013) M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. *BMC Bioinformatics*, 14:43

Wares JP, Hughes AR, Grosberg RK (2005) Mechanisms that drive evolutionary change: Insights from species introductions and invasions. In *Species invasions: insights into ecology, evolution, and biogeography*. Oxford University Press, Oxford, pp. 229-257

Wellband KW, Pettitt-Wade H, Fisk AT, Heath DD (2018) Standing genetic diversity and selection at functional gene loci are associated with differential invasion success in two non-native fish species. *Molecular Ecology*, 27:1572-1585. https://doi.org/10.1111/mec.14557

Williams S, Yaeger L (2017) Evolution of neural dynamics in an ecological model. *Geosciences*, 7:49

Xiao S *et al.* (2016) Modeling the relative importance of ecological factors in exotic invasion: The origin of competitors matters, but disturbance in the non-native range tips the balance. *Ecological Modelling*, 335:39-47

Xiong W, Li H, Zhan A (2016) Early detection of invasive species in marine ecosystems using high-throughput sequencing: technical challenges and possible solutions. *Marine Biology*. doi:10.1007/s00227-016-2911-1

Yaeger L (1994) Computational genetics, physiology, metabolism, neural systems, learning, vision, and behavior or PolyWorld: life in a new context. In Proc. Artificial Life III, Santa Fe Institute Studies in the Sciences of Complexity, vol. 17, Addison-Wesley, Redwood City, p 263-298

Yaeger LS (2013) Identifying neural network topologies that foster dynamical complexity. *Advances in Complex Systems.* doi:10.1142/S021952591350032X

Yoder J, Yaeger L (2014) Evaluating topological models of neuromodulation in Polyworld. *Artificial Life* 14:916-923. doi:10.7551/978-0-262-32621-6-ch149

Yoann T *et al.* (2016) Global change and climate-driven invasion of the Pacific oyster (Crassostrea gigas) along European coasts: a bioenergetics modelling approach. *Journal of Biogeography*, 43:568-579.

Zayed A, Constantin ŞA, Packer L (2007) Successful biological invasion despite a severe genetic load. *PLoS ONE*, 2:e868. https://doi.org/10.1371/journal.pone.0000868

Zhan A *et al.* (2013) High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. *Methods in Ecology and Evolution*, 4:558-565

Zhan A *et al.* (2014) Reproducibility of pyrosequencing data for biodiversity assessment in complex communities. *Methods in Ecology and Evolution*, 5:881-890

Zhan A, Xiong W, Song H, MacIsaac HJ (2014) Influence of artifact removal on rare species recovery in natural complex communities using high-throughput sequencing. *PLoS ONE.* doi:10.1371/journal.pone.0096928

Zhan A, Bailey SA, Heath DD, MacIsaac HJ (2014) Performance comparison of genetic markers for high-throughput sequencing-based biodiversity assessment in complex communities. *Molecular Ecology Resources*, 14:1049-1059

Zhang Z *et al.* (2019) Evolution of increased intraspecific competitive ability following introduction: The importance of relatedness among genotypes. *Journal of Ecology*, 107:387-395. https://doi.org/10.1111/1365-2745.13016

VITA AUCTORIS

NAME:	Ryan D. Scott
PLACE OF BIRTH:	Windsor, ON
YEAR OF BIRTH:	1987
EDUCATION:	University of Windsor, B.Sc., Windsor, ON, 2010 University of Windsor, BCS., Windsor, ON, 2013