DEVELOPING GENOMIC RESOURCES USING STRIPED BASS (*MORONE SAXATILIS*): GENETIC STRUCTURE, ASSOCIATIONS, AND TEXT-MINING

by

Nathalie M LeBlanc

M.Sc., Acadia University, 2016
B.Sc. Honours, Acadia University, 2013

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

Doctor of Philosophy

in the Graduate Academic Unit of Biology

Supervisor: Scott Pavey, Ph.D. (Department of Biological Sciences, UNBSJ)

Examining Board: Jason Addison, Ph.D. (Department of Biology, UNBF), Chair
Donald Baird, Ph.D. (Department of Biology, UNBF)
Patricia Evans, Ph.D. (Faculty of Computer Science, UNBF)

External Examiner: Victoria Pritchard, Ph.D. (The UHI Rivers and Lochs Institute, University of the Highlands and Islands)

This dissertation is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

June, 2021

GENERAL ABSTRACT

The advent of next-generation sequencing technologies has revolutionized the field of molecular ecology, facilitating increasingly fine-scale detection of genetic differences among populations and adaptationally significant mutations. In this thesis, I use genomics to advance solve longstanding mysteries of Striped Bass genetics and lay the groundwork for future studies. In the first chapter, I characterize a group of Striped Bass that were thought to be extirpated in the Saint John River, but likely survive as a remnant population. In the second chapter, I investigate connectivity and relatedness of Striped Bass populations more widely across their native range on the North American Atlantic Coast. I found that Gulf of St. Lawrence, Shubenacadie River, and Saint John River populations were all very distinct from each other and from US populations. US Striped Bass, however, could be separated into three major regions: Hudson River-Kennebec River, Chesapeake Bay-Delaware River, and Roanoke River-Cape Fear River. Demonstrating that this work is useful for management, my SNP loci were able to assign 99% of Striped Bass to these six regions, the first time Roanoke River Striped Bass have been reliably distinguished from Chesapeake Bay bass. Additionally, the presence of apparent US-origin Striped Bass on the northeastern coast of Nova Scotia raises important questions about movement patterns of Striped Bass in this area and highlights the importance of further study. In the third chapter, I used computer modelling simulations to assess the performance of four recent techniques used to find associations between phenotypes and genotypes. I found that Random Forest algorithm with population correction performed similarly to a recent, complex model implemented in

confounder adjusted multiple testing. Finally, in chapter four I created 9 novel tools and used them to create an automated text-mining pipeline that can scan full-text articles and extract sentences that contain associations between genes and ecological variables. This pipeline is the first step toward improving genome annotations of non-model organisms such as Striped Bass. Together, these four chapters lay important groundwork for future genomic research both for Striped Bass and other ecologically important species.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Terms

Phenotype – Observable traits of an organism, including appearance, behavior, or physiology.

Genotype – An organism’s genetic code and the specific alleles possessed by an individual organism, particularly as they relate to phenotype.

NGS – Next-generation sequencing. A group of genetic sequencing technologies that allow for the genotyping of millions of DNA sequences at once.

RAD-seq - Restriction-site associated DNA sequencing. A type of sequencing that uses NGS technology. A genome is digested with restriction enzymes, and short fragments of DNA are amplified using adapters attached to the restriction enzyme cut sites and then sequenced.

ddRAD – A type of RAD-seq that uses two different restriction enzymes to cut DNA single-nucleotide polymorphisms (SNPs).

F_{ST} – A widely used measure of genetic divergence. Differences in allele frequencies between two populations are represented by a number from 0 to 1, where 0 denotes identical populations and 1 denotes populations with no alleles in common.

DAPC – Discriminant analysis of principal components. An analysis that combines principal components with discriminant analysis that attempts to maximize differences between groups of samples and then assigns samples to each group.

Genetic clustering – A type of analysis that iteratively creates a pre-defined number of genetic clusters based on sample genotypes, and then assigns individuals to those genetic clusters.

DU – Designatable Unit. Groups of individuals or populations within a species that are considered geographically or genetically distinct by the Committee on the Status of Endangered Wildlife in Canada (COSEWIC).

GSI - Genetic stock identification. The use of genetic markers for determining the proportion of fish from different origin populations present in fishery stocks that contain a mixture of fish from multiple origins. Also called genetic mixed stock analysis.

GATE - General Architecture for Text Engineering. An open-source development environment for text-mining applications.

ANNIE - A Nearly-New Information Extraction System. An open-source collection of basic text-mining annotation tools that break text files into words, space, and sentences and identifies parts of speech.

JAPE - Java Annotation Patterns Engine. A finite state transducer implemented in the GATE environment that consists of a series of statements (“rules”) that identify patterns within text and then output new annotations based on those patterns.

GENERAL INTRODUCTION

Molecular ecology and genetics

The field of molecular ecology uses DNA or RNA to examine how organisms interact with each other in the wild and change over time (Freeland et al. 2011, de la Harpe et al. 2017). Every individual organism has a genome, which can be several hundred thousand to billions of nucleotide base pairs long, and contain genes that code for the breadth of observable biodiversity. Analyzing genomes can convey information about the genetic basis for visible morphological, physiological, and behavioral differences among individuals and populations (Freeland et al. 2011).

By examining patterns in how molecular variation is distributed within and between populations, we can reveal the level of relatedness and dispersal patterns of individuals within those populations (Freeland et al. 2011). This is because genetic mutations develop and spread through populations in predictable ways in response to different levels of connectivity and population sizes. Isolated populations will independently accumulate mutations that may become widespread due to random chance (genetic drift) or due to natural selection favoring that mutation in that population. Populations that have high amounts of migration between them as individuals move back and forth, in contrast, will remain genetically similar to each other. Together, these make up the four main forces that determine how genetically different two populations will be: mutation, genetic drift, natural selection, and migration. Mutations are the source of genetic differences between individuals, genetic drift and natural selection act to increase differences between populations, and migration acts to increase similarities between the same. When there is not enough migration between populations to counteract the effects of genetic drift and natural selection, the two populations will gradually diverge from each other at a speed that is directly correlated to the size of that population: smaller populations will diverge more rapidly while larger populations will diverge more slowly (Freeland et al. 2011).

Not every group of organisms represents a single breeding population. Migratory organisms in particular often gather together on feeding grounds, mixing together with other breeding populations. In these situations, genetics can be used as a tool to determine the origin of individuals within these mixed groups. In fisheries science, this is often called genetic stock identification (GSI) or genetic mixed stock analysis. Phenotypic traits often used for mixed stock analysis can be influenced by environmental factors the fish has been exposed to throughout its life, while a fish's genome reflects only the individual's genetic inheritance (Waldman et al. 1997). The ideal genetic markers for GSI are cost-efficient to genotype in large numbers, can be easily combined with other genetic markers and with other datasets, and give consistent and unambiguous results when used in multiple different laboratory settings.

The methods used to capture genetic variability to assess genetic structure, adaptation and performing GSI have changed over time as new methods offer improvements in power, coverage, or cost efficiency. The different kinds of genetic snapshots provided by these methods are commonly called genetic markers. Until recently microsatellites, which are composed of tandem repeated motifs of 2 to 6 nucleotides, were the genetic marker of choice for answering most questions in molecular ecology due to high variation in how many repeats are found within different individuals and among populations (Chambers and MacAvoy 2000, Defaveri et al. 2013). However, the most common methods of genotyping microsatellites, by estimating allele size using electrophoresis, are extremely difficult to standardize with data from other time periods, laboratories, or instruments as genotypes are subject to shifts in apparent allele size based on variables such as laboratory temperature during electrophoresis (Davison and Chiba 2003, Morin et al. 2009a). Combining microsatellite datasets requires one or more individuals to be present in both datasets to act as controls (Davison and Chiba 2003) or careful examination of allele size distributions by instrument, lab, or date (Morin et al. 2009a). Additionally, because of the large variation in how many alleles are present at a microsatellite locus, measures of genetic diversity calculated from different sets of microsatellite loci cannot easily be compared to each other (Seeb et al. 2007).

With the advent of next-generation sequencing (NGS) technologies, single-nucleotide polymorphisms (SNPs) have become the dominant genetic marker in almost every field in ecological genetics (Putman and Carbone 2014). A SNP is a single nucleotide polymorphism where one nucleotide base has mutated into another. A “SNP dataset” consists of genotypes of individuals for multiple SNP loci. (Defaveri et al. 2013, de la Harpe et al. 2017). Alone, a single SNP conveys relatively little information about genetic variation (Morin et al. 2009b); however current NGS sequencing allows us to amplify and sequence hundreds to thousands of SNPs, representing regions throughout an organism’s entire genome (Li and Wang 2017). Because of the ease of genotyping large numbers of SNPs, these markers have demonstrated high power for detecting very fine-scale population structure (Feder and Mitchell-Olds 2003, Ozerov et al. 2013, Putman and Carbone 2014). Additionally, because SNPs are identified by the presence of one of four nucleotides at a genetic position, there is no ambiguity in identifying alleles among samples in different datasets, making them an appealing choice for any tool intended to be used multiple times.

Next-generation sequencing has also impacted the opportunities that exist within studies of adaptation in the wild (Morin et al. 2004; Santure and Garant 2018). In the absence of natural selection, drift and migration act on all areas of the genome equally, creating very similar patterns of differentiation (Freeland et al. 2011). By identifying loci that show atypical patterns of differentiation relative to other SNPs in that species, scientists can investigate possible genetic signs of natural selection and local adaptation to specific populations (Renaut et al. 2011). SNPs allow scientists to search for these adaptations even in non-model species with few genomic resources (Morin et al. 2004). In addition to population-level patterns of adaptation, scientists have been investigating the connection between individual phenotypes such as color or body size with the underlying causal genetics for years, using several generations of carefully controlled breeding. The use of next-generation sequencing techniques to discover large numbers of SNPs in one study has allowed scientists to draw these connections between genotypes and phenotypes without this controlled breeding (Bailey-Wilson and Wilson 2011), resulting in the emergence of genome-wide association studies (GWAS). For the first time, association studies could be performed on species that could not easily be crossed in a laboratory setting, expanding the field beyond a small number of well-studied model species. Even SNP datasets originally created to answer questions about genetic structure can potentially be used to help fill in knowledge gaps surrounding adaptation and sex determination in fish species (Gamble 2016). Due to the staggering array of information these new sequencing techniques can generate, next-generation sequencing is often credited with revolutionizing the field of molecular ecology and is becoming essential tools in any molecular ecologist's toolkit (Kumar and Kocour 2017).

Setting the stage for my thesis

This thesis was one of the first projects launched in a newly created genomics lab at UNB Saint John and was one of two projects seeking to establish a reliable and efficient next-generation sequencing pipeline that would allow genomics research on freshwater and marine systems in Atlantic Canada. To this end, we created an international team spanning Atlantic Canada and the east coast of USA to assess current research questions that could be answered with more precise genetic tools, as well as available resources for answering these questions. We sought to advance our understanding of Striped Bass *Morone saxatilis* (Walbaum, 1792) in several areas, including 1) assessing fine-scale genetic structure to better understand movements and spawning patterns, particularly in the Saint John River system and in the densely populated Chesapeake Bay system, 2) producing a panel of SNPs for reliable stock identification on a finer scale than previously possible, 3) identifying genetic factors underpinning phenotypes and life history traits, 4) improving annotation of the existing Striped Bass genome by developing an automated text-mining pipeline that can extract ecological associations corresponding with known genes and unknown sequences.

In this thesis, I develop a double-digest restriction-site associated DNA protocol used to discover 2 to 7 thousand SNP loci in 3 different datasets. I use these to describe fine-scale genetic structure of Striped Bass in the Bay of Fundy as well as in all major regions of the species' native range on the North American Atlantic Coast. I also conduct initial investigations into the suitability of my SNP dataset in detecting causal loci underlying phenotypic traits using computer simulations, and develop an initial proof of concept pipeline that extracts ecological associations present in text articles. While Striped Bass genetic structure has been investigated with a number of genetic markers in the past, including mitochondrial DNA markers (Wirgin et al. 1997a) and microsatellites (Bentzen and Paterson 2008, Gauthier et al. 2013, Wirgin et al. 2020), it has rarely been examined along its entire range and never before with SNPs. Initial investigations in chapter 3 and 4 pave the way for future investigations into genotype-phenotype associations using our newly developed SNP libraries, and in an ambitious project to develop badly needed text-mining tools for molecular ecology in the next-generation sequencing era.

Overview of thesis chapters

In this general introduction, I will detail the background and study design of the four studies that make up my thesis, each touching on an aspect of molecular ecology in the age of next-generation sequencing. In chapter 1 and 2 of this thesis, I genotype the first large group of juveniles seen in the Saint John River since the population disappeared in the 1970s and report fine-scale genetic structure of all major Striped bass populations along their native range, successfully discriminating between 6 genetically distinct regions. Chapter 3 details initial investigations into the suitability of our current SNP pipeline in detecting causal loci underlying phenotypic traits, using the newly released simulation package naturalGWAS. Finally, chapter 4 presents the first step in a powerful text-mining pipeline that can scan full text articles and extract sentences containing ecological associations with known genes.

Characterizing juvenile Striped Bass in the Saint John River

The question of whether native Striped Bass still survive in the Saint John River has dogged research and management of this population for decades (Andrews et al. 2017). Striped Bass in the Bay of Fundy were listed as endangered in 2012, due to loss of recruitment in 2 of the 3 historical spawning populations within the bay (COSEWIC 2012). One of these populations was located in the Saint John River, which supported a directed Striped Bass fishery until its closure in 1978 following evidence of failed spawning and the disappearance of the local population (Andrews et al. 2017). Then, in 2008 a genetic study of adult Striped Bass in the Saint John River using microsatellites discovered a group of genetically distinct individuals thought to belong to one or two age cohorts hatched in the mid-1990s, although it was not possible to confirm that these individuals were not migrants from an unsampled population outside of the Saint John

River (Bentzen and Paterson 2008). In 2014, large numbers of juvenile Striped Bass were discovered in the Saint John River for the first time since 1979, prompting questions about their origin and the status of spawning within the river. By analyzing their genetic profile, we can determine whether these juveniles appear to be descended from migrant US or Shubenacadie Striped Bass that frequent the Saint John River, or if they are genetically divergent from other Striped Bass populations such as the group discovered by Bentzen and Paterson (2008).

The initial genetic characterization of juveniles captured in the Saint John River was analyzed and published for rapid dissemination in light of the ongoing discussion surrounding the Mactaquac Dam (Bradford et al. 2015, Andrews et al. 2017), and is detailed in chapter 1 of this thesis. DNA was extracted from juveniles collected in 2014 and 2015, as well as from Striped Bass taken from Shubenacadie River, Hudson River, and Chesapeake Bay. Using this DNA, I measured genetic differences among the four groups of Striped Bass to determine where these juvenile Striped Bass likely came from. We found that Saint John River juveniles were markedly distinct from all other groups, supporting Bentzen and Paterson's findings of a genetically distinct group of Striped Bass in the river, and providing evidence of a surviving local spawning population. Additionally, hybrid individuals showing mixed descent from the Saint John River genetic group and the two other groups indicate migrant spawning within the river.

Findings detailed in this chapter represent one part of an extensive investigation into Striped Bass movements in the Saint John River, described in full in Andrews, 2019, as part of the Mactaquac Aquatic Ecosystem Study (MAES) to inform an upcoming decision regarding the fate of the failing Mactaquac Generating Station on the Saint John River. I conducted genetic analyses for several subsequent studies related to this project; these are included in publications detailing abundance of juvenile Striped Bass of different age cohorts between 2014 and 2019 (Andrews et al. 2019), as well as movement data from acoustically tagged juveniles and sub-adults (Andrews et al. 2020b) and acoustically tagged adult Striped Bass (Andrews et al. 2020a).

Contrasting high gene flow US populations with isolated Canadian populations

An important role of genetic markers in fisheries management is stock discrimination and mixed stock analysis, where fish of unknown origin can be traced back to their source population. Accurate stock discrimination helps fisheries managers understand the distribution of individuals from different spawning populations in a stock from year to year and thereby better predict population sizes from year to year as the health of origin populations varies (Cuéllar-Pinzón et al. 2016). Mixed stock analysis can also help to track shifts in distribution of a species over time, particularly as climate change results in northward range shifts across the globe (Sunday et al. 2012, Spies et al. 2020).

Chesapeake Bay is the largest producer of Striped Bass in their native range, and thought to be the origin of the majority of Striped Bass found along the coast between North Carolina and the Bay of Fundy (Wirgin et al. 1997b, Richards and Rago 1999). Stock discrimination attempting to quantify the proportion of Striped Bass from different spawning grounds in different areas have been complicated by pronounced genetic similarities among all migratory Striped bass populations from North Carolina to Maine (Waldman et al. 2012). Spawning populations in this area are typically separated into three or four management groups: Hudson River, Roanoke River, Chesapeake Bay, and occasionally Delaware River. Typically, existing genetic markers cannot reliably differentiate between Chesapeake Bay and Roanoke River groups, and so these are combined in stock discrimination studies (Waldman and Fabrizio 1994, Waldman et al. 2012). Genetic markers have also failed to consistently differentiate between Delaware River and Chesapeake Bay, and among rivers within Chesapeake Bay (Brown et al. 2005), despite frequent significant FST values indicating these rivers contain discrete populations (Brown et al. 2005, Gautier et al. 2013, Wirgin et al. 2020). The application of SNPs, which have the potential for very fine-scale discrimination of genetic groups (Bourret et al. 2013, Baetscher et al. 2017), may overcome challenges in stock discrimination due to the genetic similarity seen in this area of the Striped Bass range.

To this end, in Chapter 2 of this thesis I used 1256 single nucleotide polymorphism (SNP) loci and samples from 477 Striped Bass to investigate genetic structure present among 15 locations along the Striped Bass native range. These locations spanned the major migratory populations in Hudson River, Chesapeake Bay, and North Carolina, including samples from six tributaries within the Chesapeake Bay, as well as samples from the recently restored Kennebec River in Maine and several known Canadian spawning populations. The Mira River, located on the northeastern coast of Nova Scotia, represents an area of the Canadian Striped Bass range currently being investigated in depth for the first time (Buhariwalla 2018). My objective in the chapter as a whole was twofold: 1) conduct a general overview of population genetic structure along the Striped Bass range, and 2) determine whether the SNP dataset assembled in this study has greater power to discriminate between Striped Bass groups than previously examined genetic markers.

To assess the performance of our SNPs in stock discrimination, we compared it to a previous study that attempted to use microsatellites to discriminate between populations spanning Hudson River to the Santee-Cooper system in South Carolina (Gauthier et al. 2013). My SNP dataset successfully assigned 99% of Striped Bass samples back to one of six distinct genetic groups. Importantly, Striped bass from Chesapeake Bay and Roanoke River were easily distinguished from each other. Broad patterns of genetic structure included greater divergence among Canadian populations than US populations, extremely small genetic differences among Chesapeake Bay tributaries, individuals showing both US origin and Gulf of St. Lawrence origin in the Mira River, and genetic similarity between Kennebec River juveniles and Hudson River Striped Bass.

Genotype-phenotype associations

The development of a SNP library for Striped Bass opens the possibility of investigating phenotype-genotype associations for the first time, and the question of which association method is best used for this purpose. In a species with strong genetic structure such as that seen among Canadian populations of Striped Bass, it is essential to ensure that an association test is able to effectively take neutral genetic structure into account while also retaining power to detect true causal loci (Zhao et al. 2007). In particular, Random Forest machine learning methods have seen growing popularity in association studies due to their ability to simultaneously measure the importance of large numbers of loci in predicting a given phenotype (Brieuc et al. 2018); however there is not yet a ‘standard’ method of correcting for population structure when using Random Forest. I wanted to assess the performance of a number of popular association tests and in particular the performance of an easily implemented method of Random Forest population structure described in Zhao et al. (2012). This method has been evaluated in simulations containing small numbers (1 to 5) of causal loci, but has not been assessed in detecting causal loci of polygenic traits.

Chapter 3 details the most complete of my explorations into genotype-phenotype connections within Striped Bass. A large number of samples collected over the course of my PhD contained length and weight information that can be combined to give an individual’s condition factor: essentially how big the fish is controlled for length. While examining possible avenues for analyzing this data, I discovered a new R package known as *naturalGWAS*. Released and recommended in 2018 (Santure and Garant 2018, François and Caye 2018), adoption of this new tool has been slow. NaturalGWAS uses empirical genotypes as the basis of its simulations in order to create a set of quantitative phenotypes whose numerical values are tied to a given number of loci in the dataset. Because the quantitative phenotypes produced by naturalGWAS closely resemble my existing condition factor data, I decided to use this package to assess the suitability of possible association tests for use on Striped Bass.

Using 171 Striped Bass that had length and weight measurements available for condition factor calculations, I assembled a library of 7319 SNPs that passed quality filtering. With this dataset and naturalGWAS, I assessed the performance of 4 different association tests in detecting the causal loci underlying quantitative phenotypes, compared to the performance of a theoretical "ideal" mixed model association test. Using empirical genotypes, I created 100 sets of simulated phenotypes determined by 2, 5, 10, 20, or 30 causal loci at different effect sizes. Each effect size-casual loci combination was repeated five times to create five replicate phenotypes. I evaluated each association test's ability to correctly identify these causal loci, and calculated power (number of detected causal loci/total number of causal loci), number of false positives, and false discovery rate (number of false positives/total number of significant loci). I also used three different thresholds when correcting for multiple testing to see if different tests performed better when more liberal or more conservative methods were used.

Across simulations, two association methods performed better on our dataset than others: confounder adjusted multiple testing (CATE) and regression-corrected Random Forest described by Zhao (2012) referred to here as Zhao's Random Forest. Zhao's Random Forest, notably, produced remarkably few false positives relative to other tests, even at the most liberal multiple test correction threshold. We ran association tests on our empirical condition factor phenotypes using both CATE and Zhao’s Random Forest, and found no evidence of casual loci in our SNP dataset.

Ecological annotations in the genomic age

In a world of exponentially increasing genomic data and literature publication, fields such as biomedicine have long relied on information extraction, databases, and curated ontologies to ensure knowledge of gene functions and disease associations are readily accessible to researchers (Ashburner et al. 2000, Spasic et al. 2005, Fleuren and Alkema 2015). At the same time, the rise of high-throughput next-generation sequencing techniques in molecular ecology has resulted in an increasing number of genes with information about ecological traits to which that gene helps mediate reactions, which is largely spread out in unconnected literature and supporting documents (Pavey et al. 2012). Associating ecological contexts to genes is known as ecological annotation, and it complements the more traditional functional annotation data curated by medical research (Landry and Aubin-Horth 2007).

In order to access and connect the growing body of research surrounding ecological association, there is a need for similar tools as those used in biomedicine (Pavey et al. 2012). The ability for researchers to quickly find existing ecological associations enables molecular ecologists both to place functional information into an ecological context, and identify ecological patterns in genes of unknown function that can shed light on its role in an ecosystem (Pavey et al. 2012, Andrew et al. 2013). In particular, 1) the ability to automatically parse the natural language present in peer-reviewed research and extract existing association data from published literature in regards to genes and also un-annotated sequences, 2) comprehensive ontologies that both consolidate information about ecological associations with genes and standardize terminology, 3) ultimately, a tool allowing automatic population of ontologies with existing association data, that can be curated downstream.

As the first step toward these objectives, I assessed the state of text-mining tools for ecological annotation and explored possible avenues toward bridging the gap between current tools and a potential tool for automatically extracting ecological associations for use in an ontology. The result of this investigation is presented in chapter 4, where I detail an initial proof of concept pipeline created using rule-based co-occurrence annotation, that scans full text articles for sentences containing ecological association information and extracts those sentences into a csv file. I assembled a corpus of 137 manually curated full text articles to train and test this pipeline and determined that it was able to successfully extract 88% of manually identified sentences with association data contained in them, and 36% of all extracted sentences contained complete association information. The vocabulary and rule sets created for this pipeline will be made available for future text-mining projects in molecular ecology, both within and outside of the Pavey Lab. I also outline next steps both for my pipeline, and for additional tools required to expand ecological text-mining to detect associations with unnamed sequences.

Statement of authorship

Each chapter in this thesis represents an individual study intended for peer-review and publication. Co-authors of each article are listed on the corresponding title pages of each chapter, and sources of all tissue used in chapters 1 to 3 are listed in the corresponding acknowledgements. For all chapters, I drafted the first version of the manuscript and incorporated all recommendations and edits from co-authors and peer reviewers in subsequent versions. I am the first and corresponding author in all published chapters.

Chapter 1 has been published in the North American Journal of Fisheries Management. It was accepted on September 28, 2018 and released in volume 38 in December 2018. This chapter was a collaborative effort on the part of myself and several co-authors, listed in the chapter, particularly Samuel Andrews who identified the juveniles under investigation in the Saint John River and provided “on the ground” context for the ongoing question of a surviving native population of Striped Bass in this river. I was not responsible for tissue collection of samples in this chapter, although I was involved in locating and negotiating a source for Shubenacadie River tissue samples. I did all laboratory work and library preparation, as well as designing and conducting all downstream data analysis.

Chapter 2 has been published in the Evolutionary Applications journal. It was accepted April 15, 2020 and released in volume 6 in July 2020. In this chapter, I conducted the majority of laboratory work, and also supervised DNA extraction of a subset of samples and library preparation of two libraries. I did all laboratory work and library preparation, designing and conducting downstream data analysis, including refining quality filtration procedures from chapter 1.

Chapter 3 was submitted to Molecular Ecology Resources on November 27, 2020 and is under review. This chapter used DNA samples from chapter 2 as well as additional samples extracted and prepared by me. I decided on the final topic of this chapter, designed and carried out the simulations presented therein, and interpreted the results.

Chapter 4 remains to be submitted for publication; several journals have been reviewed as possible submission targets but a decision has not yet been made. In this chapter, I began with a pipeline framework previously used to test a toxicology text-mining pipeline, and updated the code as necessary to incorporate the tools developed in the chapter. All new gazetteers and JAPE rules were designed, researched, and written by me. Manual curation of articles was begun by me and completed by several lab technicians under my direction and supervision. Test runs of the pipeline were run by me, and analysis of the resulting extracted sentences for calculation of recall and precision was also done by me.

Literature Cited

Andrew, R. L., L. Bernatchez, A. Bonin, C. A. Buerkle, B. C. Carstens, B. C. Emerson, D. Garant, T. Giraud, N. C. Kane, S. M. Rogers, J. Slate, H. Smith, V. L. Sork, G. N. Stone, T. H. Vines, L. Waits, A. Widmer, and L. H. Rieseberg. 2013. A road map for molecular ecology. Molecular Ecology 22:2605–2626.

Andrews, S. N., T. Linnansaari, R. A. Curry, and M. J. Dadswell. 2017. The misunderstood Striped Bass of the Saint John River, New Brunswick: Past, present, and future. North American Journal of Fisheries Management 37:235–254.

Andrews, S. N. 2019. Restoration potential for reproduction by Striped Bass (*Morone saxatilis*) in the Saint John River, New Brunswick. PhD Thesis. University of New Brunswick.

Andrews, S. N., T. Linnansaari, N. Leblanc, S. Pavey, and R. A. Curry. 2019. Interannual variation in spawning success of Striped Bass (*Morone saxatilis*) in the Saint John River, New Brunswick. River Research and Applications 36:13–24.

Andrews, S. N., T. Linnansaari, R. A. Curry, N. M. Leblanc, and S. A. Pavey. 2020a. Winter ecology of Striped Bass (*Morone saxatilis*) near its northern limit of distribution in the Saint John River, New Brunswick. Environmental Biology of Fishes 103:1343–1358.

Andrews, S. N., T. Linnansaari, N. Leblanc, S. Pavey, and R. Curry. 2020b. Movements of juvenile and sub-adult Striped Bass *Morone saxatilis* in the Saint John River, New Brunswick, Canada. Endangered Species Research 43:281–289.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, and J. T. Eppig. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25:25–29.

Baetscher, D. S., D. J. Hasselman, K. Reid, E. P. Palkovacs, and J. C. Garza. 2017. Discovery and characterization of single nucleotide polymorphisms in two anadromous alosine fishes of conservation concern. Ecology and Evolution 7:6638–

6648.

Bailey-Wilson, J. E., and A. F. Wilson. 2011. Linkage analysis in the next-generation sequencing era. Human Heredity 72:228–236.

Bentzen, P., and I. G. Paterson. 2008. Report: genetic analysis of Striped Bass collected by Kingsclear First Nation in the Saint John River, New Brunswick. Report to the Department of Fisheries and Oceans, Dartmouth, Nova Scotia. p. 1-22.

Bourret, V., M. P. Kent, C. R. Primmer, A. Vasemägi, S. Karlsson, K. Hindar, P. McGinnity, E. Verspoor, L. Bernatchez, and S. Lien. 2013. SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic Salmon (*Salmo salar*). Molecular Ecology 22:532–551.

Bradford, R. G., E. A. Halfyard, T. Hayman, and P. Leblanc. 2015. Overview of 2013 Bay of Fundy Striped Bass biology and general status. Canadian Science Advisory Secretariat Research Document.

Brieuc, M. S. O., C. D. Waters, D. P. Drinan, and K. A. Naish. 2018. A practical introduction to Random Forest for genetic association studies in ecology and evolution. Molecular Ecology Resources 18:755–766.

Brown, K. M., G. A. Baltazar, and M. B. Hamilton. 2005. Reconciling nuclear microsatellite and mitochondrial marker estimates of population structure: breeding population structure of Chesapeake Bay Striped Bass (*Morone saxatilis*). Heredity 94:606–15.

Buhariwalla, C. 2018. Documenting aspects of the ecology of Striped Bass *Morone saxatilis* (Walbaum, 1792) in northeastern Nova Scotia (Master’s thesis). Acadia University, Wolfville, NS, Canada.

Chambers, G. K., and E. S. MacAvoy. 2000. Microsatellites: consensus and controversy. Comparative biochemistry and physiology. Part B, Biochemistry & molecular biology 126:455–76.

COSEWIC. 2012. COSEWIC Assessment and Status Report Striped Bass.

Cuéllar-Pinzón, J., P. Presa, S. J. Hawkins, and A. Pita. 2016. Genetic markers in marine fisheries: Types, tasks and trends. Fisheries Research 173:194–205.

Davison, A., and S. Chiba. 2003. Laboratory temperature variation is a previously unrecognized source of genotyping error during capillary electrophoresis. Molecular Ecology Notes 3:321–323.

de la Harpe, M., M. Paris, D. N. Karger, J. Rolland, M. Kessler, N. Salamin, and C. Lexer. 2017. Molecular ecology studies of species radiations: Current research gaps, opportunities and challenges. Molecular Ecology:2608–2622.

Defaveri, J., H. Viitaniemi, E. Leder, and J. Merilä. 2013. Characterizing genic and nongenic molecular markers: Comparison of microsatellites and SNPs. Molecular Ecology Resources 13:377–392.

Feder, M. E., and T. Mitchell-Olds. 2003. Evolutionary and ecological functional genomics. Nature Reviews Genetics 4:649–655.

Fleuren, W. W. M., and W. Alkema. 2015. Application of text mining in the biomedical domain. Methods 74:97–106.

François, O., and K. Caye. 2018. Naturalgwas: An R package for evaluating genomewide association methods with empirical data. Molecular Ecology Resources 18:789–797.

Freeland, J. R., H. Kirk, and S. Petersen. 2011. Molecular Ecology. 2nd edition. John Wiley & Sons, Ltd.

Gamble, T. 2016. Using RAD-seq to recognize sex-specific markers and sex chromosome systems. Molecular ecology 25:2114–2116.

Gauthier, D. T., C. A. Audemard, J. E. L. Carlsson, T. L. Darden, M. R. Denson, K. S. Reece, and J. Carlsson. 2013. Genetic population structure of US Atlantic Coastal Striped Bass (*Morone saxatilis*). Journal of Heredity 104:510–520.

Gautier, M., K. Gharbi, T. Cezard, J. Foucaud, C. Kerdelhué, P. Pudlo, J. M. Cornuet, and A. Estoup. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology 22:3165–3178.

Kumar, G., and M. Kocour. 2017. Applications of next-generation sequencing in fisheries research : A review. Fisheries Research 186:11–22.

Landry, C. R., and N. Aubin-Horth. 2007. Ecological annotation of genes and genomes through ecological genomics. Molecular Ecology 16:4419–4421.

Li, Y. H., and H. P. Wang. 2017. Advances of genotyping-by-sequencing in fisheries and aquaculture. Reviews in Fish Biology and Fisheries 27:535–559.

Morin, P. A., G. Luikart, and R. K. Wayne. 2004. SNPs in ecology, evolution and conservation. Trends in Ecology and Evolution 19:208–216.

Morin, P. A., C. Manaster, S. L. Mesnick, and R. Holland. 2009a. Normalization and binning of historical and multi-source microsatellite data: Overcoming the problems of allele size shift with allelogram. Molecular Ecology Resources 9:1451–1455.

Morin, P. A., K. K. Martien, and B. L. Taylor. 2009b. Assessing statistical power of SNPs for population structure and conservation studies. Molecular Ecology Resources 9:66–73.

Ozerov, M., A. Vasemägi, V. Wennevik, R. Diaz-Fernandez, M. Kent, J. Gilbey, S.

Prusov, E. Niemelä, and J. P. Vähä. 2013. Finding markers that make a difference: DNA pooling and SNP-arrays identify population informative markers for genetic stock identification. PLoS ONE 8.

Pavey, S. A., L. Bernatchez, N. Aubin-Horth, and C. R. Landry. 2012. What is needed for next-generation ecological and evolutionary genomics? Trends in Ecology & Evolution 27:673–678.

Putman, A. I., and I. Carbone. 2014. Challenges in analysis and interpretation of microsatellite data for population genetic studies. Ecology and Evolution 4:4399–4428.

Renaut, S., A. W. Nolte, S. M. Rogers, N. Derome, and L. Bernatchez. 2011. SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (*Coregonus spp.*). Molecular Ecology 20:545–559.

Richards, R. A., and P. J. Rago. 1999. A case history of effective fishery management: Chesapeake Bay Striped Bass. North American Journal of Fisheries Management 19:356–375.

Santure, A. W., and D. Garant. 2018. Wild GWAS—association mapping in natural populations. Molecular Ecology Resources 18:729–738.

Seeb, L. W., A. Antonovich, M. A. Banks, T. D. Beacham, M. R. Bellinger, S. M. Blankenship, M. R. Campbell, N. A. Decovich, J. C. Garza, C. M. Guthrie III, T. A. Lundrigan, P. Moran, S. R. Narum, J. J. Stephenson, K. J. Supernault, D. J. Teel, W. D. Templin, J. K. Wenburg, S. F. Young, and C. T. Smith. 2007. Development of a standardized DNA database for Chinook salmon. Fisheries 32:540–552.

Spasic, I., S. Ananiadou, J. McNaught, and A. Kumar. 2005. Text mining and ontologies in biomedicine: Making sense of raw text. Briefings in Bioinformatics 6:239–251.

Spies, I., K. M. Gruenthal, D. P. Drinan, A. B. Hollowed, D. E. Stevenson, C. M. Tarpey, and L. Hauser. 2020. Genetic evidence of a northward range expansion in the eastern Bering Sea stock of Pacific Cod. Evolutionary Applications 13:362–375.

Sunday, J. M., A. E. Bates, and N. K. Dulvy. 2012. Thermal tolerance and the global redistribution of animals. Nature Climate Change 2:686–690.

Waldman, J. R., and M. C. Fabrizio. 1994. Problems of stock definition in estimating relative contributions of Atlantic Striped Bass to the coastal fishery. Transactions of the American Fisheries Society 123:766–778.

Waldman, J. R., R. A. Richards, W. B. Schill, I. Wirgin, and M. C. Fabrizio. 1997. An Empirical Comparison of Stock Identification Techniques Applied to Striped Bass. Transactions of the American Fisheries Society 126:369–385.

Waldman, J. R., L. Maceda, and I. Wirgin. 2012. Mixed-stock analysis of wintertime aggregations of Striped Bass along the Mid-Atlantic coast. Journal of Applied Ichthyology 28:1–6.

Wirgin, I., L. Maceda, J. Stabile, and C. Mesing. 1997a. An evaluation of introgression of Atlantic coast Striped Bass mitochondrial DNA in a Gulf of Mexico population using formalin-preserved museum collections. Molecular Ecology 6:907–916.

Wirgin, I., J. R. Waldman, L. Maceda, J. Stabile, and V. J. Vecchio. 1997b. Mixed-stock analysis of Atlantic Coast Striped Bass (*Morone saxatilis*) using nuclear DNA and mitochondrial DNA markers. Canadian Journal of Fisheries and Aquatic Sciences 54:2814–2826.

Wirgin, I., L. Maceda, M. Tozer, J. Stabile, and J. Waldman. 2020. Atlantic coastwide population structure of Striped Bass *Morone saxatilis* using microsatellite DNA analysis. Fisheries Research 226:105506.

Zhao, K., M. J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. 2007. An Arabidopsis example of association mapping in structured samples. PLoS Genetics 3:0071–0082.

Zhao, Y., F. Chen, R. Zhai, X. Lin, Z. Wang, L. Su, and D. C. Christiani. 2012. Correction for population stratification in random forest analysis. International Journal of Epidemiology 41:1798–1806.

CHAPTER 1: Evidence of a genetically distinct population of Striped Bass (*Morone saxatilis*) within the Saint John River, New Brunswick, Canada

1. Nathalie M. Leblanc (corresponding author)

Department of Biological Sciences, Canadian Rivers Institute, University of New Brunswick, Saint John, NB, E2L 4L5, Canada. Email: nleblan5@unb.ca

Email: nathalie.leblanc@unb.ca

2. Samuel N. Andrews

Canadian Rivers Institute, Department of Biology, University of New Brunswick, Fredericton, Canada

3. Trevor S. Avery

Departments of Biology and Mathematics & Statistics, Acadia University, Wolfville, NS, B4P2R6, Canada.

4. Gregory N. Puncher

Department of Biological Sciences, Canadian Rivers Institute, University of New Brunswick, Saint John, NB, E2L 4L5, Canada.

Molecular Biology Laboratory, Maurice Lamontagne Institute, Fisheries and Oceans Canada, Mont-Joli, QC, G5H 3Z4, Canada

5. Benjamin I. Gahagan

Massachusetts Division of Marine Fisheries, Annisquam River Marine Fisheries Station, 30 Emereson Avenue Gloucester, MA, United States of America

6. Andrew R. Whiteley

Department of Ecosystem and Conservation Sciences and Wildlife Biology Program,

W. A. Franke College of Forestry and Conservation, University of Montana, Missoula, MT, 59812

7. R. Allen Curry

Canadian Rivers Institute, Department of Biology and Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, Canada

8. Scott A. Pavey

Department of Biological Sciences, Canadian Rivers Institute, University of New Brunswick, Saint John, NB, E2L 4L5, Canada. Email: spavey@unb.ca

Abstract

Sound management of species requires, among other things, careful consideration of their distribution and genetic structure throughout their range. Historically, there were three spawning populations of Striped Bass *Morone saxatilis* occurring within the Bay of Fundy, Canada (Shubenacadie River, Annapolis River, and Saint John River), but today the only known spawning population is found in Shubenacadie River, Nova Scotia. The last spawning event recorded (albeit unsuccessful) in the Saint John River was in 1975 shortly after the completion of the Mactaquac Dam in 1968. Adult Striped Bass from other rivers frequent the Saint John River during much of the year, making the presence of adults uninformative about the status of spawning. In the absence of direct indicators of spawning, such as eggs and larvae, genomic tools can provide insight into the genetic origin of the juvenile Striped Bass in the Saint John River. Tissue samples were taken from Striped Bass (age 1-3, total length 12.2–35.0 cm) captured in the Saint John River, and compared with samples from Shubenacadie River, Hudson River, and Chesapeake Bay. A double digest RAD-seq (ddRAD) technique was used to identify 4,700 single nucleotide polymorphisms (SNPs) and population structure was assessed using population differentiation statistics (F_{ST}) and genetic clustering algorithms. F_{ST} analysis found significant differences among all sample sites, albeit weak differences between Hudson River and Chesapeake Bay samples, and a global F_{ST} of 0.101 ($P < 0.001$). Genetic clustering and DAPC analyses both grouped samples into three clusters; Shubenacadie River, the U.S.A. populations, and the Saint John River juveniles. Based on these findings and the current understanding of Striped Bass juvenile dispersal, there

is strong evidence of a genetically distinct population of Striped Bass within the Saint John River.

Introduction

The native range of Striped Bass *Morone saxatilis* (Walbaum, 1792), extends along the Atlantic coast of North America from the Gulf of St. Lawrence into the Gulf of Mexico, with introduced populations on the Pacific coast and landlocked reservoirs throughout the interior U.S.A. Striped Bass north of Albemarle Sound, North Carolina are anadromous, and adults from Chesapeake Bay and the Hudson River are commonly found in coastal regions from Albemarle Sound to the Bay of Fundy in Canada throughout much of the year (Gauthier et al. 2013). In Canada, there are three locations in which recognized active spawning populations still exist: 1) the Shubenacadie River, Nova Scotia (Bay of Fundy), 2) the Miramichi River in the Southern Gulf of St. Lawrence, and 3) the Rivière-du-Sud basin of the St. Lawrence River (recently reintroduced from the Miramichi River; COSEWIC 2012). Each of these populations are classified into a Designatable Unit (DU) for management purposes: Bay of Fundy, Southern Gulf of St. Lawrence, and St. Lawrence River, respectively, and are thought to be reproductively isolated (COSEWIC 2012). The Bay of Fundy DU once encompassed three spawning populations (Shubenacadie River, Annapolis River, and Saint John River; Figure 1.1), but today only the Shubenacadie River population is known to persist, and thus this DU has been designated Endangered. Conservation of existing populations and recovery of historical spawning groups is important for the persistence of native Striped Bass in the Bay of Fundy (e.g., COSEWIC 2012, Bradford et al. 2015, Andrews et al. 2017).

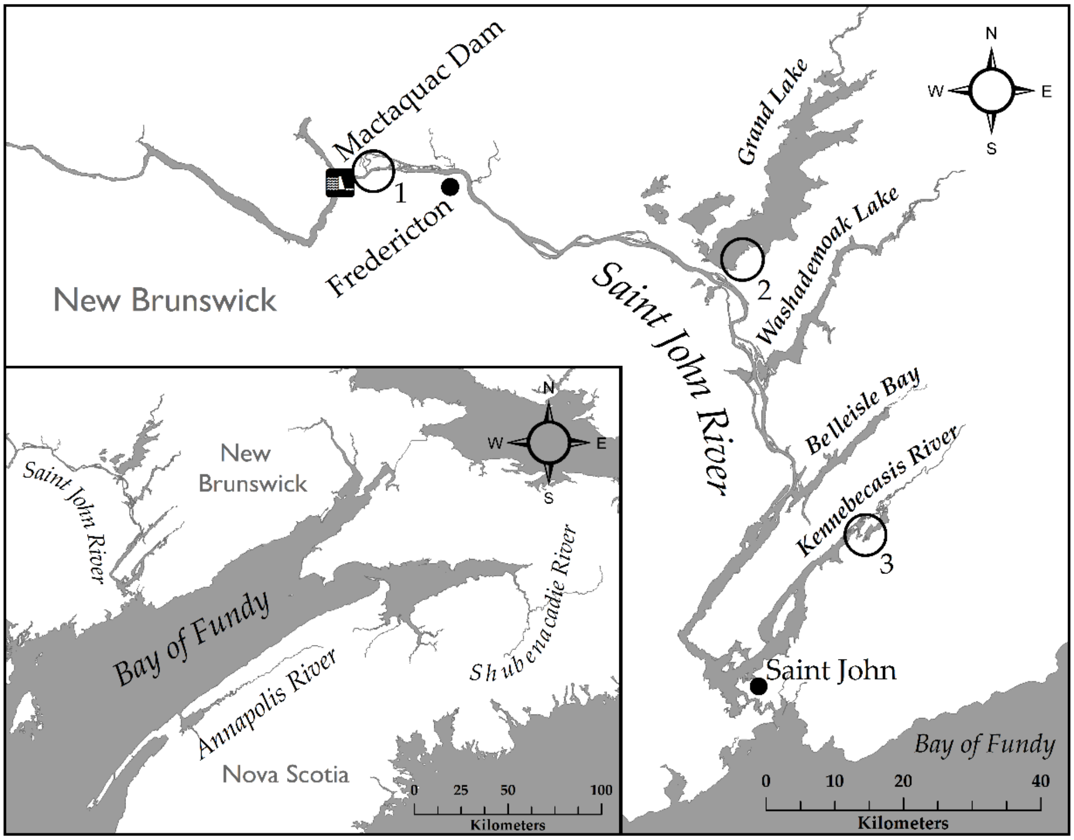


Figure 1.1. In the bottom left, a map of the three rivers in the Bay of Fundy Striped Bass have historically spawned in. To the right, a map of the entire Saint John River with sampling sites marked in numbered open circles: 1) Downstream of Mactaquac Dam, 2) Grand Lake, and 3) Kennebecasis River.

The Saint John River (SJR) is a long (~670 km) tidally influenced river that traverses the provinces of Quebec, New Brunswick and State of Maine before emptying into the Bay of Fundy at the City of Saint John in southeast New Brunswick. The SJR is the second longest river between the Gulf of St. Lawrence and the Gulf of Mexico and has a watershed that spans ~55,000 km^2 and mean annual discharge of ~1100 m^3/s (Cunjak and Newbury 2005). The SJR once hosted a large and widely recognized population of native

Striped Bass that supported both a commercial fishery and renowned recreational fisheries throughout the 19th and most of the 20th century (Rulifson and Dadswell 1995; Andrews et al. 2017). Historical accounts indicate that SJR Striped Bass spawned during spring amongst the islands near the city of Fredericton (Andrews et al. 2017), and that all spawning in the SJR likely ceased following the completion of the Mactaquac Dam in 1968 (Rulifson and Dadswell 1995; Bradford et al. 2012; Andrews et al. 2017). On a single occasion in 1975, eggs were collected in Belleisle Bay on the SJR, suggesting spawning had occurred; however, 95% of the eggs had ruptured chorionic membranes (Dadswell 1975), potentially disrupting spawning and early life survival and contributing to population decline. The population was believed lost (Rulifson and Dadswell 1995; COSEWIC 2004), although the latest evaluation of Striped Bass in the Bay of Fundy lists the population status as ‘uncertain’ following recent evidence documented below (COSEWIC 2012).

In addition to being an historical spawning location, the SJR hosts migrant Striped Bass from the Shubenacadie River and U.S.A. populations, making assessment of the local population difficult (COSEWIC 2012; Bradford et al. 2012). Historically, adults from visiting populations have been common (Wirgin et al. 1995; Bentzen and Paterson 2008); however, the last juvenile Striped Bass (a single age -1 individual) reported in the SJR was captured in 1979 and since then report of small Striped Bass (i.e., <35 cm) have been scarce (COSEWIC 2012; Andrews et al. 2017). Searches for juvenile Striped Bass and direct evidence of spawning in the SJR have been conducted sporadically since 1975 with no success (Douglas et al. 2003; Andrews et al. 2017); however, these surveys were conducted in only a few locations and over short periods.

Two studies have used molecular tools to investigate mixed stocks within the Bay of Fundy. Wirgin et al. (1995) investigated the origins of Striped Bass in the Bay of Fundy using individual adults collected during spring and summer of 1992 and 1993. These samples were profiled using mtDNA restriction fragment length polymorphisms (RFLPs) that had been shown previously to successfully differentiate between Striped Bass from Shubenacadie River and U.S.A. Striped Bass (Wirgin et al. 1993). Wirgin et al (1995) concluded that the proportion of Striped Bass with U.S.A. ancestry within the SJR was 63% in 1992 and 97% in 1993, the remainder originating from the Shubenacadie River. The proportion of Striped Bass with U.S.A. ancestry in the Shubenacadie River in the same years was 1.2% and 6.8%, respectively. Wirgin et al. (1995) suggest in their discussion that a local spawning population does not occur in the SJR. RLFP markers have less power to discriminate among populations than microsatellites or SNPs (Shaw et al. 1999, Liu and Cordes 2004), and the RFLPs used in Wirgin et al. (1995) may be unable to distinguish among individuals from the SJR and the Shubenacadie River or U.S.A. In comparison, Bentzen and Paterson (2008) used 11 microsatellites to analyze adult Striped Bass captured at or near the Mactaquac Dam, on the SJR in 1999–2006. They reported three groups in the SJR with origins from the Shubenacadie River, U.S.A. rivers, and an unknown group they tentatively identified as a “native” SJR population. The proportion of this tentative SJR population varied from 9–85% of the mixed population over seven years of study.

A recent telemetry study that tracked 22 adult Striped Bass over a period of 1–3 years from 2010–2014 found that tagged adults remained in the SJR throughout the year and demonstrate fidelity to specific overwintering sites over multiple years (Andrews et al. 2018). The same study observed Striped Bass leaving the river temporarily during the spawning season as well as migrating to the area downstream from the Mactaquac Dam, an historical spawning location. Andrews et al. (2018) was the first study to provide evidence of adults residing in the SJR throughout the year and migrating to a historical spawning location during the spawning season.

In 2014, commercial fishermen reported large numbers of very small Striped Bass, presumably young-of-the-year, in their nets in the SJR, and juveniles of increasing size have continued to appear every year thereafter. Shubenacadie River juveniles of similar size have been found in low numbers as far as Five Islands in Minas Basin (Rulifson et al. 2008). Miramichi River young-of-the-year were shown to travel 55 km from their native river (Robinson et al. 2004), but Miramichi River Striped Bass of any age have not been shown to travel farther south than the northwestern coast of Nova Scotia (Douglas et al. 2006). Large numbers of age 1 juvenile Striped Bass found in the SJR were likely spawned there, and raise the question of whether these juveniles belong to the same genetically distinct group of adult Striped Bass found by Bentzen and Paterson (2008).

We used genomic methods to re-examine the genetic structure of the SJR Striped Bass. Genome-wide analysis of thousands of markers is best for genetic analyses when only a small number of samples can be obtained, as it has the power to detect genetic divergence and diversity across few individuals (Willing et al. 2012; Nazareno et al. 2017). Genetic structure was examined using juvenile using juvenile Striped Bass age 1–3, sampled from the Kennebecasis River, Grand Lake, and Mactaquac Dam in the SJR system. These sites are possible nursery habitats for Striped Bass (Andrews et al. 2017). Samples were compared with Striped Bass from Shubenacadie River, Hudson River, and Chesapeake Bay.

Methods

Sample collection

Fin clips (tissue samples) were collected from spawning adult Striped Bass captured in the Upper Chesapeake Bay (n = 23, collected in April 2012 with gill nets) and Hudson River (n = 23, collected in April and May 2012 with boat electrofishing). Shubenacadie River (n = 22) scales were collected from adult fish caught by angling in the Stewiacke River from late April to late June 2013–2015. Striped Bass in this tributary of the Shubenacadie River gather for spawning during these months after descending from Grand Lake at the headwaters of the Shubenacadie River (not to be confused with Grand Lake in the SJR watershed). Individuals sampled during this period are considered of Shubenacadie River origin and are used in population surveys to minimize the inclusion of U.S.A. migrants (DFO 2014). In the SJR, Striped Bass were collected from three locations (n = 23): Kennebecasis River (n = 19), Grand Lake (n = 1) and the Mactaquac Dam (n = 3) by means of gill net, fyke net, Gaspereau trap, boat electrofishing and angling (Figure 1.1). Grand Lake and Mactaquac Dam juveniles were caught in 2015. Kennebecasis juveniles were caught in 2014 (n=11), 2015 (n=5), and 2016 (n=1). All juveniles were caught from June–October, following the Canadian spawning period of May–June (Rulifson and Dadswell 1995). All SJR fish were measured and weighed when tissue samples were collected, and age was determined from scales. To ensure accuracy, ageing was performed by multiple researchers on 3–7 scales per fish. The modal age was recorded. Juveniles were 1–3 years of age (total length range 12.2–35.0 cm) and appeared to be a single year class: fish samples caught in 2014 were age 1, in 2015 were age 2, and in 2016 were age 3.

Laboratory

DNA was isolated using the NucleoMag® 96 Tissue (Machery-Nagel, Düren, Germany) kit and an epMotion 5075t (Cat. 5075000302). Libraries were prepared using a modified double-digest restriction-site associated DNA sequencing (ddRAD-seq or ddRAD) protocol developed by Poland et al. (2012). Samples were digested using restriction enzymes *PstI* and *MspI*. In this study, DNA fragment size selection (377–523 bp) was performed before PCR amplification using a Sage Pippin Prep © platform, while Poland et al. (2012) performed size selection after amplification. All individuals were processed on the same lane using paired-end sequencing of 125 bp with an Illumina® HiSeq™ 2500 (San Diego, U.S.A.) at Génome Québec Innovation Centre.

Quality control and data processing

Adaptors were trimmed from the resulting DNA libraries using Cutadapt v. 1.13 (Martin 2011) and quality before and after was assessed by eye with FastQC v.0.11.5 (Andrews 2010). Trimmed sequences were demultiplexed into individual samples using the attached barcodes via Stacks v.1.46 (Catchen et al. 2013) *process radtags* module. Sequences were trimmed to a uniform length of 110 bp, and the paired end option –P was used to simultaneously process both sets of reads. Samples with less than two million reads were not included in subsequent analysis.

Sequences were aligned to the Striped Bass genome (BioProject accession no. PRJNA266827) using BWA v. 0.7.15 (Li and Durbin 2010). Specifically, bwa mem was used with the default parameters except for a minimum seed length (k) of 19 and a maximum seed occurrence of 500. Samtools was used to exclude alignments with a mapq score smaller than 1, or the following flags: 4, 256, 2048. Alignments were output as Binary Alignment Map (BAM) files, a binary version of the Sequence Alignment Map (SAM) text file format designed to efficiently store large amounts of nucleotide information (Li et al. 2009). Stacks 1.46 modules from here on were executed using Éric Normandeau’s Stacks workflow scripts, available on github (https://github.com/enormandeau/stacks_workflow, downloaded August 2016), which makes use of the module rxstacks to remove confounded and poor quality loci, and make corrections to SNP calls by filtering out probable sequencing errors and then reevaluating those individuals as heterozygotes or homozygotes. The 'snp' model type and an alpha of 0.1 was used for this purpose. Loci with a log likelihood of less than -40 were excluded, as well as stacks with a depth of less than 5, loci with more than 20% missing data in a population, and loci that were not present in all populations. SNPs were then filtered using F_{IS} values (F_{IS} -0.3 or lower) that are calculated in the Stacks *populations* module to eliminate highly heterozygous loci as possible paralogs. The remaining SNPs were then filtered again using VCFtools v. 0.1.13 (Danecek et al. 2011) to eliminate all loci with a minor allele frequency less than 0.01 over all individuals. Summary statistics were calculated by *populations*. Sibship tests were run on all populations, one population at a time, using Colony v. 2.0.6.4 (Jones and Wang 2010), to confirm that individuals did not share a parent.

Population structure

Overall F_{ST} was calculated using an AMOVA algorithm implemented in Arlequin v. 3.5.2.2 (Excoffier and Lischer 2010). The same program was used to calculate pairwise F_{ST}, and significance was assessed by way of 10,100 random permutation tests. Structure files output by Stacks were converted to Arlequin input files using PGDSpider v. 2.1.1.0 (Lischer and Excoffier 2012). The R package LEA v. 2.0 (Frichot and François 2015) estimates ancestry coefficients for all individuals given a proposed number of ancestral populations (K) using an algorithm that has been optimized for use with large numbers of genetic markers. Simulations were run using theorized K 1–5 with default settings, without specifying collection site. The probability of a K being valid was calculated using the cross-entropy criterion, and the K with the lowest minimal cross-entropy value across 10 runs was selected as being most probable (Frichot et al. 2014). Striped Bass were assigned to a genetic cluster if their ancestry coefficient was greater than 0.8 (80%) and were otherwise considered admixed. The R package adegenet v. 2.1.1 was used to run Discriminant Analysis of Principal Components (DAPC). DAPC applies Discriminant analysis to a principal component analysis (PCA) to maximize the difference between an assumed number of groups (or clusters), without using *a priori* information about locations or populations (Jombart et al. 2010). DAPC was run using 1–5 assumed numbers of clusters to mirror K 1–5 populations and the number of groups with the lowest Bayesian Information Criterion (BIC) value was selected as most probable number of clusters. A DAPC was run with all axes retained and the a-score calculated to select the optimal number of principal components to retain (Jombart and Collins 2015). DAPC was then run again with a lower number of principal components retained. Principal components explaining the highest amount of variation were plotted against each other, with amount of variation explained represented as DA eigenvalues.

Results

Quality filtering

A total of 84 samples were retained (Table 1.1), and seven samples were excluded due to low sequence coverage (fewer than two million reads each). The excluded samples were from the SJR (n=2; both from Kennebecasis), Shubenacadie River (n=1), and Chesapeake Bay (n=4). The initial SNP catalogue contained 377,931 loci. Filtering removed 373,231 loci in total. Loci were removed based on log likelihood (35,507), low F_{IS} values and non-polymorphic loci (41,892), minor allele frequencies (1,727), and various other filters including minimum stack depth, > 2 alleles and loci not present in all populations (294,105). Of these remaining loci, 72–80% were polymorphic in Chesapeake Bay, Hudson River, and the SJR, whereas 55% were polymorphic in the Shubenacadie River. Expected and observed heterozygosities were similar (average difference of 0.0034; Table 1.1). Observed heterozygosities ranged from 0.262–0.295, which is lower than previous studies on Striped Bass using microsatellites. This is a consequence of marker type because microsatellites typically have many alleles (sometimes 30 or more) and also have very high heterozygosity rates (Balloux and Lugon-Moulin 2002). SNPs, which have two alleles each, correspondingly have lower heterozygosity rates. Because observed heterozygosity varies based on marker type, it is best interpreted relative to other studies and species analyzed using the same kind of marker. Observed heterozygosity of SJR Striped Bass was comparable to that of U.S.A. Striped Bass sampled in this study. Percent polymorphic loci was lower than U.S.A. populations and higher than the Shubenacadie River population. Sibship analysis showed no evidence for full- or half-siblings in any of the populations, confirming that the juvenile Striped Bass collected in the SJR were not part of the same family unit.

Table 1.1. Number of individuals (# Ind.), number of polymorphic loci (Poly. loci) and percent polymorphic loci (in brackets), observed heterozygosity (Obs. het.), and expected heterozygosity (Exp. het.) calculated for 84 Striped Bass samples amplified at 4,700 SNP loci. CHPK = Chesapeake Bay, HUD = Hudson River, SHUB = Shubenacadie, SJR = Saint John River.

Population	# Ind.	Poly. loci	Obs. het.	Exp. het.
CHPK	19	4647 (83%)	0.279	0.278

HUD	23	4865 (87%)	0.262	0.269
SHUB	21	3448 (62%)	0.295	0.297
SJR	21	4342 (78%)	0.270	0.273

Table 1.2. Pairwise comparisons of 84 Striped Bass samples from four sampling sites (CHPK = Chesapeake Bay, HUD = Hudson River, SHUB = Shubenacadie, SJR = Saint John River) using 4,700 SNP loci. FST values are located above the diagonal and P-values are below the diagonal.

	CHPK	HUD	SHUB	SJR
CHPK	*	0.019	0.159	0.080
HUD	0.000	*	0.149	0.074
SHUB	0.000	0.000	*	0.115
SJR	0.000	0.000	0.000	*

Population structure

The overall F_{ST} among all sites was 0.101 (p-value < 0.001). All sites were significantly different from each other using pairwise F_{ST} values ($p \leq 0.001$; Table 1.2), confirming that there is reproductive isolation at each sampling location. Bonferroni is a very conservative method of pairwise correction (Rice 1988), thus, F_{ST} values that remained significantly different between sites with standard Bonferroni correction exceed a rather conservative threshold (corrected p-value of 0.008). Upper Chesapeake Bay and Hudson River Striped Bass were the least genetically differentiated pair (F_{ST} = 0.019), while the Shubenacadie River was most genetically divergent from both Chesapeake Bay and Hudson River (F_{ST} = 0.160 and 0.150, respectively). Juvenile Striped Bass from the SJR were quite distinct from the other two genetic clusters but more closely related to the U.S.A. populations (F_{ST} = 0.076 and 0.081) than to the Shubenacadie River (F_{ST} = 0.115).

The most probable number of ancestral genetic populations based on LEA's entropy measurement was three (K = 3; Appendix A) and assignment of individuals to each cluster strongly corresponds to geographic area (Figure 1.2). Figure 1.2A depicts mean assignment of individuals in each site to a cluster, while Figure 1.2B depicts assignment of individual Striped Bass to each cluster. Almost all U.S.A. individuals (19/19 from Chesapeake Bay and 20/23 individuals from the Hudson River) were assigned to the U.S.A. cluster (CHPK + HUD) with ≥ 80% ancestry coefficient. A second cluster (SHUB) consisted of all Shubenacadie River individuals, assigned with > 92% ancestry coefficient. Only samples taken from the SJR (15/21 individuals) were assigned to a third cluster (putatively SJR) and were assigned with > 80% ancestry coefficient. The remaining six individuals were admixed. Three admixed individuals had about equal assignment to both the U.S.A. cluster and the putative SJR cluster, one about equal assignment to both the SHUB cluster and the putative SJR cluster, and two had 27–28% assignment to the putative SJR cluster and 69% assignment to the U.S.A. cluster (Figure 1.2B).

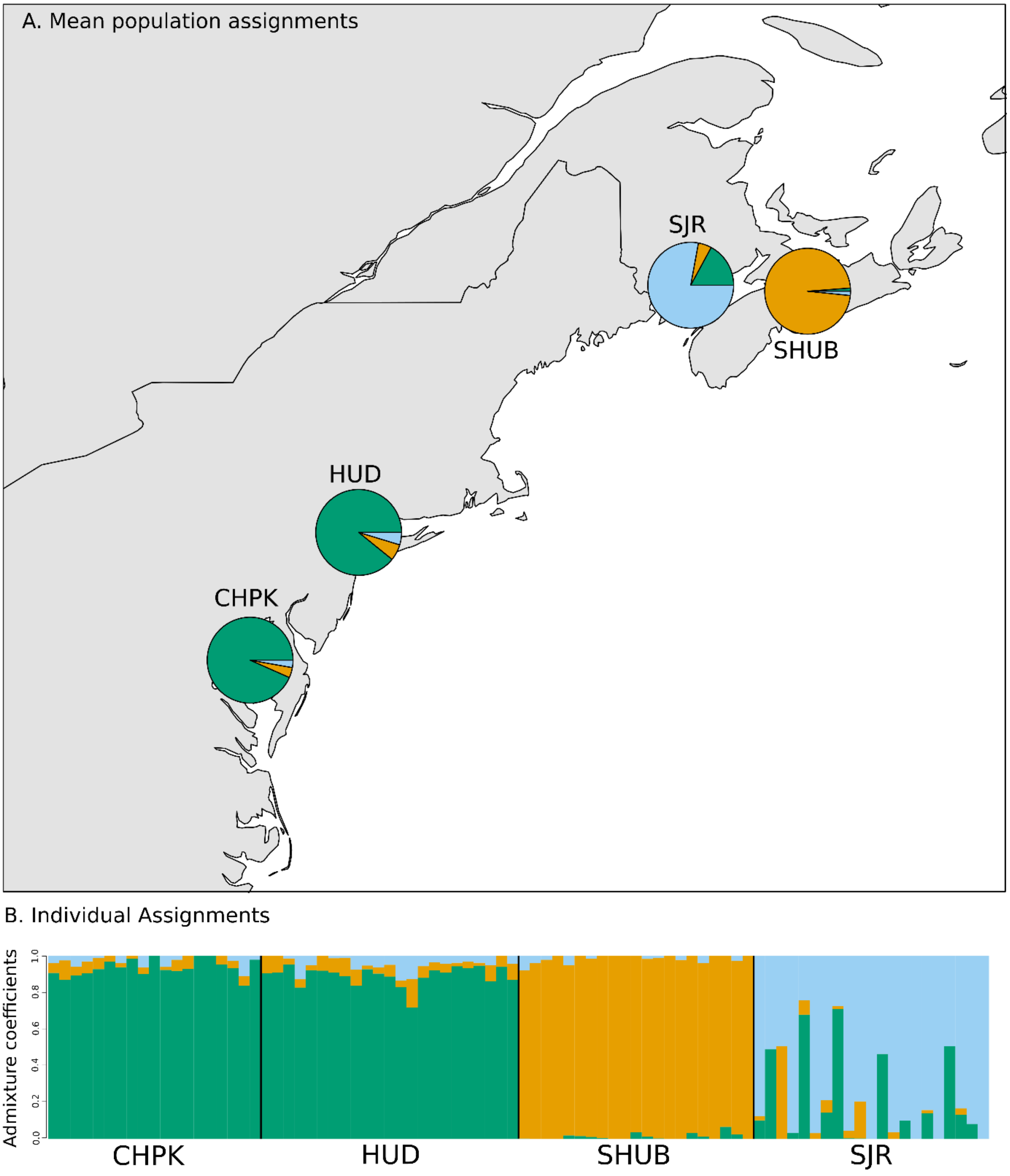


Figure 1.2. A) Mean ancestry coefficient, measured by LEA, of Striped Bass, *Morone saxatilis*, collected at four sites (CHPK = Chesapeake Bay, HUD = Hudson River, SHUB = Shubenacadie River, SJR = Saint John River) to three genetic clusters, calculated using 4,700 SNP loci. B) Individual ancestry coefficients of

Striped Bass in each site to the three genetic clusters. Individual Striped Bass are represented by vertical bars, with percent probability of assignment to each cluster represented by colours. SHUB = Shubenacadie River, SJR = Saint John River, HUD = Hudson River, CHPK = Chesapeake River.

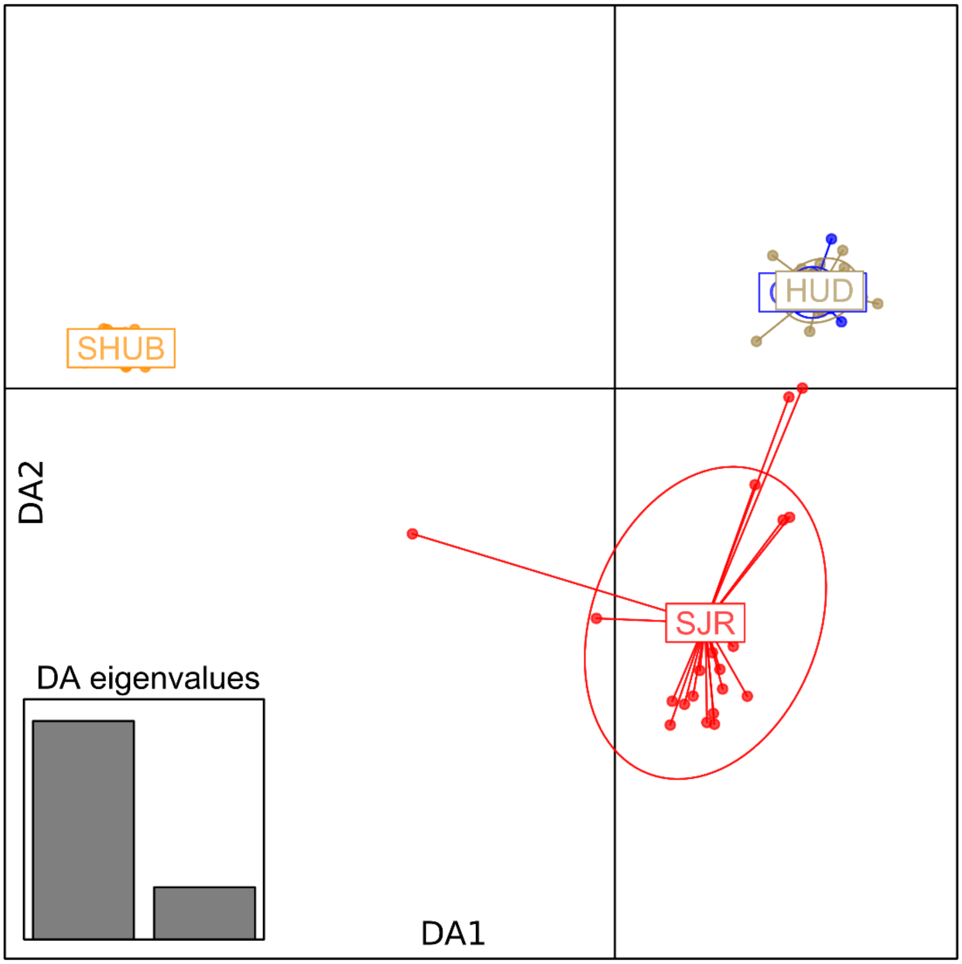


Figure 1.3. DAPC plot of Striped Bass, *Morone saxatilis*, populations collected in four sites (Chesapeake Bay (CHPK), Hudson River (HUD), Shubenacadie River (SHUB), and Saint John River (SJR)), constructed using 4,700 SNPs. Individual Striped Bass are represented by symbols depicted in the legend, and a line connects the dot to the site it was sampled in. Distance between dots corresponds to genetic distance along two discriminant functions.

In DAPC, the most likely number of groups was three (Figure 1.3). The recommended number of principal components to retain to prevent overfitting varied from 1–10, and all analyses were run with 10 components retained. When three groups were assumed, Shubenacadie River samples formed a distinct cluster from other populations and Chesapeake Bay and Hudson River samples overlapped. Saint John River was intermediate between the two along the first principal component and separated along the second, with three individuals located closer to the Chesapeake Bay or Shubenacadie River groups.

Discussion

The juvenile Striped Bass from the SJR displayed a mix of three genetic groups: parents originating from the Shubenacadie River, U.S.A. rivers, and an unassignable but clearly single parentage source. All SJR juvenile Striped Bass in our study (n = 21, age 1–3) showed some assignment to a putative SJR cluster and overall were highly divergent from all other populations according to both F_{ST} and DAPC measures. The genetic cluster of these juveniles could represent a persistent population within the SJR. Our results match the study by Bentzen and Paterson (2008) who found three similar genetic groups and that 23% of adult SJR Striped Bass sampled near the Mactaquac Dam belonged to a distinct genetic cluster not seen in the Shubenacadie River or U.S.A. populations. Though spawning in the SJR remains to be confirmed, the probability that these unassigned juvenile Striped Bass are migrants from a previously unknown population is low. Further efforts should ideally be made to confirm and characterize this local population of Striped Bass. The presence of genetically distinct juvenile Striped Bass in the SJR, despite existing survey methods finding no evidence of eggs or larvae, highlights the need for a better understanding of Striped Bass ecology in the SJR to effectively restore the species to one of its native rivers.

Several alternative explanations are possible for the origin of the putative SJR juveniles, including an unsampled source population not represented in our baseline collection. Spawning formerly occurred in the Kennebec River, Maine (Little 1995) but the population was believed to be extirpated in the 1930s (Squiers et al. 1984). The river was stocked using Hudson River and hatchery-raised Striped Bass from 1982–1991 (Flagg and Squiers 1992, cited in Rulifson and Laney 1999) and spawning is suspected to have begun anew in recent years due to the reappearance of juveniles in the area (Greene et al. 2009). Bentzen and Paterson (2008) showed juvenile Striped Bass caught in the Kennebec River were genetically indistinguishable from those in the Hudson River, suggesting that it is not the source of SJR juveniles. Further, there are no reports of new, undefined populations along the coast between the SJR and the Hudson River. While our collection of Striped Bass from the Chesapeake Bay was limited to only one spawning area, our results conform to prior work, which found that Striped Bass within Chesapeake Bay are more like each other than to Hudson River (Gauthier et al. 2013) and even greater genetic discrepancy between the Chesapeake and Canadian populations are expected. In Canadian waters, the next closest source of juveniles is the Shubenacadie River, > 100 km away. Adults from the Shubenacadie River have been found in the SJR

(Bradford et al. 2015; Andrews et al. 2017), but aside from a single admixed juvenile, none of the SJR juveniles analyzed in this study are genetically like Shubenacadie River Striped Bass. Bentzen and Paterson (2008) showed genetic differences between the unknown genetic signature in the SJR and the Miramichi River population, the next closest Canadian spawning population to the SJR, and neither Bentzen and Paterson (2008), nor Wirgin et al. (1995) found instances of Miramichi River adult migrants within the Saint John River (the Miramichi River RFLP genotype was well documented at the time of Wirgin et al. 1995).

Another possible source of this genetic population is recolonization by a very small number of migrant Striped Bass several generations ago, with no further gene flow. The small number of breeders and resulting inbreeding can result in low genetic diversity within the colonized population and rapid genetic divergence from the source population due to a founder effect (*sensu* Templeton 1980). Empirical studies on one-time founder effects in recently colonized populations have found little (F_{ST} = 0.02–0.05; Hawley et al. 2005; Eales et al. 2008) to no (Clegg et al. 2002; Melany et al. 2018) genetic divergence between the introduced and source populations after five to fifty generations. Within the SJR, the continued presence of adult migrants makes it likely that if occasional migrant breeding is happening it was not a one-time occurrence, as seen in most cases of founder effect. Ongoing gene flow between the SJR and Shubenacadie River / U.S.A. is supported by the presence of admixture in some SJR juveniles. The juvenile Striped Bass in our samples also have similar levels of heterozygosity and polymorphism to U.S.A. populations, indicating similar levels of genetic diversity. The presence of gene flow combined with the observed genetic diversity make the possibility of a founder effect less likely (Melany et al. 2018). Regardless of whether SJR juveniles represent a remnant of the pre-Mactaquac Dam population or a founder effect-type recolonization, the evidence of a third genetic cluster presented here suggests that the SJR population is self-sustaining, that gene flow with adjacent populations is present, and that successful spawning may be sporadic. The presence of a local spawning population is supported by the existence of resident Striped Bass within the river, as found through telemetry data (Andrews et al. 2018).

Temporal variability of spawning success is evidenced by the size distribution of juvenile Striped Bass in our study, which were primarily from one year (likely 2013). Bentzen and Paterson (2008) also reported that individuals assigned to the putative SJR population were of similar body size and probably from one cohort. Year class dominance is seen in populations throughout the Striped Bass range (Ulanowicz and Polgar 1980; Bradford et al. 2015) and particularly within Canadian populations, possibly due to variable spawning conditions (Peer and Miller 2014, Douglas et al. 2009). In the SJR there has been no evidence of spawning since 1975 (Dadswell 1975), but repeated, though infrequent, reports of juvenile Striped Bass in commercial fisheries (Andrews et al 2017) may be evidence of the temporal nature of successful spawning in the SJR.

Striped Bass populations in Canadian waters have grown significantly in recent years (DFO 2018) and adults have been found > 1,000 km from their rivers of origin. Adults from the Miramichi River were caught in Labrador (Andrews et al. 2019; Avery, unpublished data), and juvenile Striped Bass from the Shubenacadie River have been caught in the Petitcodiac River (Redfield 2018). The potential for interbreeding within and among rivers combined with possible range expansion introduces a complicated set of challenges for Striped Bass managers. Current assessment of Striped Bass endangered status is partly based on an absence of evidence that U.S.A. migrants spawn in Canadian waters (COSEWIC 2012). Further studies are essential to fill knowledge gaps that make monitoring and managing this putative SJR population difficult. Infrequent spawning success impacts the efficacy of attempts to observe eggs or larvae, especially when survey effort is also sporadic. Future surveys should incorporate past known spawning areas, new acoustic telemetry results, and be completed annually over a long period. Additional genetic studies should be done using a greater number of juveniles, ideally from multiple cohorts, and including less likely migrant origins such as the Miramichi River and the small population that has recently been observed in the Kennebec River.

The year class observed in this paper should continue to be monitored, and consistent monitoring of both presence and abundance of other year classes should be done to identify patterns in spawning success and identify possible reasons why 2013 was successful. By identifying commonalities in environmental conditions between the successful production of multiple dominant year classes, we can predict which years are likely to be successful in the future. Key nursery habitats for these juveniles should be identified and protected, and timing and movement of migrant adults should be measured to produce information on when and where they are present within the SJR. Acoustic tracking will be an invaluable tool in answering many of these questions by allowing tracking of the movement of individual juveniles and adult migrants in the SJR throughout the year, and these data should be combined with genetic analysis to determine the origin of the Striped Bass being tracked. The re-establishment of a spawning population in one of the two historical spawning rivers in the Bay of Fundy (Annapolis River and SJR) was identified as a management goal by the Department of Fisheries and Oceans (DFO 2006, cited in Bradford et al. 2015). The presence of these genetically distinct juveniles in the SJR may indicate that Striped Bass are currently spawning in the SJR, but a greater understanding of the life history of these Striped Bass is needed before devising a management plan.

Acknowledgements

Anglers and commercial fishermen in the Saint John River and Shubenacadie River provided valuable local knowledge, and caught many of the sampled Shubenacadie River Striped Bass. Sequencing services were provided by Génome Québec Innovation Centre. We would also like to thank the Striped Bass Research Team for coordinating sampling from the Shubenacadie River, Beth Versak (MD DNR) for coordinating sampling from the Chesapeake Bay, and Andy Kahnle (NYSDEC Bureau of Marine Resources, Hudson River Fisheries Unit) coordinating sampling from the Hudson River. Research was funded by NSERC, the Canadian Rivers Institute, Canada Research Chair, New Brunswick Innovation Foundation, the Canada Foundation for Innovation, and the Massachusetts Division of Marine Fisheries.

Literature Cited

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Andrews, S. N., T. Linnansaari, R. A. Curry, and M. J. Dadswell. 2017. The misunderstood Striped Bass of the Saint John River, New Brunswick: past, present, and future. North American Journal of Fisheries Management 37(1):235–254.

Andrews, S. N., Wallace, B., Gautreau, M., Linnansaari, T. and Curry, R. A., 2018. Seasonal movements of Striped Bass *Morone saxatilis* in a large tidal and hydropower regulated river. Environmental Biology of Fishes, pp.1-10.

Andrews, S. N., M. J. Dadswell, C. F. Buhariwalla, T. Linnansaari, and R. A. Curry. 2019. Looking for Striped Bass in Atlantic Canada: The Reconciliation of Local, Scientific, and Historical Knowledge. Northeastern Naturalist 26:1–30.

Balloux, F., and Lugon-Moulin, N. 2002. The estimation of population differentiation with microsatellite markers. Molecular ecology 11(2):155-165.

Bentzen, P., and I. G. Paterson. 2008. Report: genetic analysis of Striped Bass collected by Kingsclear First Nation in the Saint John River, New Brunswick. Report to the Department of Fisheries and Oceans, Dartmouth, Nova Scotia. p. 1-22.

Bradford, R. G., P. LeBlanc, and P. Bentzen. 2012. Update status report on Bay of Fundy Striped Bass (*Morone saxatilis*). DFO Page Canadian Science Advisory Secretariat Research Document 2012/021.

Bradford, R. G., E. A. Halfyard, T. Hayman, and P. LeBlanc. 2015. Overview of 2013 Bay of Fundy Striped Bass biology and general status. Canadian Science Advisory Secretariat Research Document.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22(11):3124–3140.

Clegg, S. M., S. M. Degnan, J. Kikkawa, C. Moritz, A. Estoup, and I. P. F. Owens. 2002. Genetic consequences of sequential founder events by an island-colonizing bird. Proceedings of the National Academy of Sciences 99(12):8127–8132.

COSEWIC (Committee on the status of endangered wildlife in Canada). 2004. COSEWIC assessment and status report on the Striped Bass (*Morone saxatilis*) in Canada. COSEWIC, Canadian Wildlife Service Environment Canada, Ottawa.

COSEWIC (Committee on the status of endangered wildlife in Canada). 2012. COSEWIC assessment and status report on the Striped Bass (*Morone saxatilis*) in Canada. COSEWIC, Canadian Wildlife Service Environment Canada, Ottawa.

Cunjak R. A., Newbury R. W. 2005. Atlantic coast rivers of Canada, Chapter 21. Pages 939–980 *in* A. C. Benke, C. E. Cushing, editors. Rivers of North America. Elsevier Inc, San Diego, California.

Dadswell, M. J. 1975. Mercury, DDT, and PCB content of certain fishes from the Saint John River estuary, New Brunswick. Transactions of the Atlantic Chapter, Canadian Society of Environmental Biologists annual meeting. Huntsman Marine Laboratory, St. Andrews, New Brunswick.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

DFO (Department of Fisheries and Oceans). 2006. Recovery Potential Assessment for the St. Lawrence River, southern Gulf of St. Lawrence and Bay of Fundy Striped Bass

(*Morone saxatilis*) Populations. DFO Canadian Science Advisory Secretariat Science Advisory Response 2006/053.

DFO. 2014. Recovery potential assessment for the Bay of Fundy Striped Bass (*Morone saxatilis*) designatable unit. DFO Canadian Science Advisory Secretariat Science Advisory Report 2014/053.

DFO. 2018. Spawner abundance and biological characteristics of Striped Bass (*Morone saxatilis*) in the southern Gulf of St. Lawrence in 2017. DFO Canadian Science Advisory Secretariat Science Response 2018/016

Douglas, S. G., R. G. Bradford, and G. Chaput. 2003. Assessment of Striped Bass (*Morone saxatilis*) in the maritime provinces in the context of species at risk. DFO Canadian Science Advisory Secretariat Research Document 2003/008.

Douglas, S. G., Chaput, S., and Caissie, D. 2006. Assessment of status and recovery potential for Striped Bass (*Morone saxatilis*) in the southern Gulf of St. Lawrence. Canadian Science Advisory Secretariat.

Douglas, S. G., G. Chaput, J. Hayward, and J. Sheasgreen. 2009. Prespawning, spawning, and postspawning behavior of Striped Bass in the Miramichi River. Transactions of the American Fisheries Society 138:121-134.

Eales, J., R. S. Thorpe, and A. Malhotra. 2008. Weak founder effect signal in a recent introduction of Caribbean Anolis. Molecular Ecology 17(6):1416–1426.

Excoffier, L., and H. E. L. Lischer. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10(3):564–567.

Flagg, L. N., and T. S. Squiers. 1992. Striped Bass restoration in Maine, pp. 137-140. In

NOAA and USFWS. 1992. Abstracts of the emergency Striped Bass study annual workshop. Atlantic States Marine Fisheries Commission, Washington, DC.

Frichot, E., F. Mathieu, T. Trouillon, G. Bouchard, and O. François. 2014. Fast and efficient estimation of individual ancestry coefficients. Genetics, 196(4):973-983.

Frichot, E., and O. François. 2015. LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution 6(8):925–929.

Gauthier, D. T., C. A. Audemard, J. E. L. Carlsson, T. L. Darden, M. R. Denson, K. S. Reece, and J. Carlsson. 2013. Genetic population structure of US Atlantic coastal Striped Bass (*Morone saxatilis*). Journal of Heredity 104(4):510–520.

Greene, K. E., Zimmerman, J. L., Laney, R. W., & Thomas-Blate, J. C. (2009). Atlantic coast diadromous fish habitat: A review of utilization, threats, recommendations for conservation, and research needs. Atlantic States Marine Fisheries Commission Habitat Management Series, 464.

Hawley, D. M., D. Hanley, A. A. Dhondt, and I. J. Lovette. 2006. Molecular evidence for a founder effect in invasive House Finch (*Carpodacus mexicanus*) populations experiencing an emergent disease epidemic. Molecular Ecology 15(1):263–275.

Jombart, T., S. Devillard and F. Balloux. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC genetics 11(1):94

Jombart, T., and C. Collins. 2015. A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0. 0. London: Imperial College London, MRC Centre for Outbreak Analysis and Modelling.

Jones, O., and J. Wang. 2010. COLONY: a program for parentage and sibship inference from multilocus genotype data. Molecular Ecology Resources 10:551–555.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26(5):589–95.

Lischer, H. E. L., and L. Excoffier. 2012. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics 28(2):298–299.

Little, M. J. 1995. A report on the historic spawning grounds of the Striped Bass, "*Morone saxatilis*." Maine Naturalist 3(2):107–113.

Liu, Z. J., and J. F. Cordes. 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture 238(1–4):1–37.

Malaney, J. L., C. W. Lackey, J. P. Beckmann, and M. D. Matocq. 2018. Natural rewilding of the Great Basin: Genetic consequences of recolonization by Black Bears (*Ursus americanus*). Diversity and Distributions 24(2):168–178.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17(1):10.

Nazareno, A. G., Bemmels, J.B., Dick, C.W., and Lohmann, L.G. (2017). Minimum sample sizes for population genomics: An empirical study from an Amazonian plant species. Molecular Ecology Resources, 17(6):1136-1147.

Peer, A. C., and T. J. Miller, 2014. Climate change, migration phenology, and fisheries management interact with unanticipated consequences. North American Journal of

Fisheries Management 34:94-110.

Poland, J. A., P. J. Brown, M. E. Sorrells, and J. L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7(2).

Redfield, E. 2018. Petitcodiac Fish Trap Results, 2017 Season. Fort Folly Habitat Recovery, Dorchester New Brunswick.

Rice, W. R. 1989. Analyzing Tables of Statistical Tests. Evolution 43(1):223–225.

Robinson, M., S. Courtenay, T. Benfey, L. Maceda, and I. Wirgin. 2004. Origin and Movements of Young-of-the-Year Striped Bass in the Southern Gulf of St. Lawrence, New Brunswick. Transactions of the American Fisheries Society 133(2):412–426.

Rulifson, R. A., and M. J. Dadswell. 1995. Life history and population characteristics of Striped Bass in Atlantic Canada. Transactions of the American Fisheries Society 124:477–507.

Rulifson, R. A., and W. Laney. 1999. Striped Bass stocking programs in the United States: ecological and resource management issues. Page Canadian Stock Assessment Secretariat Research Document 99/007.

Rulifson, R. A., S. A. McKenna, and M. J. Dadswell. 2008. Intertidal habitat use, population characteristics, movement, and exploitation of Striped Bass in the inner Bay of Fundy, Canada. Transactions of the American Fisheries Society 137(1):23–32.

Shaw, P. W., C. Turan, J. M. Wright, M. O'connell, and G. R. Carvalho. 1999. Microsatellite DNA analysis of population structure in Atlantic Herring (*Clupea* *harengus*), with direct comparison to allozyme and mtDNA RFLP analyses. Heredity 83(4):490–499.

Squiers, T. S., Jr. 1984. Quarterly report on the Kennebec River anadromous stock evaluation, Project AFC-23-3. Department of Marine Resources, Augusta, Maine.

Templeton, A. R. 1980. The theory of speciation via the founder principle. Genetics 94(9):1011–1038.

Ulanowicz, R. E., and T. T. Polgar. 1980. Influences of anadromous spawning behavior and optimal environmental conditions upon Striped Bass (*Morone saxatilis*) year-class success. Canadian Journal of Fisheries and Aquatic Sciences 37:143–154.

Willing, E. M., Dreyer, C., and Van Oosterhout, C. 2012. Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. PloS one, 7(8):e42649.

Wirgin, I., T.-L. Ong, L. Maceda, J. R. Waldman, D. Moore, and S. Courtenay. 1993. Mitochondrial DNA variation in Striped Bass (*Morone saxatilis*) from Canadian rivers. Canadian Journal of Fisheries and Aquatic Sciences 50:80–87.

Wirgin, I., B. Jessop, S. Courtenay, M. Pedersen, S. Maceda, and J. R. Waldman. 1995. Mixed stock analysis of Striped Bass in rivers of the Bay of Fundy as revealed by mitochondrial DNA. Canadian Journal of Fisheries and Aquatic Sciences 52(5):961–970.

CHAPTER 2: Genomic Population Structure of Striped Bass (*Morone saxatilis*) from the Gulf of St. Lawrence to Cape Fear River

1. Nathalie M. Leblanc (corresponding author)

Department of Biological Sciences, Canadian Rivers Institute, University of New Brunswick, Saint John, NB, E2L 4L5, Canada. Email: nleblan5@unb.ca

Email: nathalie.leblanc@unb.ca

2. Benjamin I. Gahagan

Massachusetts Division of Marine Fisheries, Annisquam River Marine Fisheries Station, 30 Emereson Avenue Gloucester, MA, United States of America

Email: ben.gahagan@state.ma.us

3. Samuel N. Andrews

Department of Biology, Canadian Rivers Institute, University of New Brunswick, Fredericton, Canada

Email: samuel.andrews@unb.ca

4. Trevor S. Avery

Departments of Biology and Mathematics & Statistics, Acadia University, Wolfville, NS, B4P2R6, Canada.

Email: trevor.avery@acadiau.ca

5. Gregory N. Puncher

Department of Biological Sciences, Canadian Rivers Institute, University of New Brunswick, Saint John, NB, E2L 4L5, Canada.

Molecular Biology Laboratory, Maurice Lamontagne Institute, Fisheries and Oceans Canada, Mont-Joli, QC, G5H 3Z4, Canada

Email: Gregory.Puncher@unb.ca

6. Benjamin J. Reading

Department of Applied Ecology, North Carolina State University, Raleigh, NC, 27695, United States of America

Pamlico Aquaculture Field Laboratory, North Carolina State University, Aurora, NC 27806, United States of America

Email: bjreadin@ncsu.edu

7. Colin F. Buhariwalla

Nova Scotia Fisheries and Aquaculture, Inland Fisheries Division, 91 Beeches Road, Pictou, NS, BOK 1H0

Email: Colin.Buhariwalla@novascotia.ca

8. R. Allen Curry

Department of Biology and Faculty of Forestry and Environmental Management, Canadian Rivers Institute, University of New Brunswick, Fredericton, Canada

Email: racurry@unb.ca

9. Andrew R. Whiteley

Department of Ecosystem and Conservation Sciences and Wildlife Biology Program,

W. A. Franke College of Forestry and Conservation, University of Montana, Missoula, MT, 59812

Email: andrew.whiteley@mso.umt.edu

10. Scott A. Pavey

Department of Biological Sciences, Canadian Rivers Institute, University of New Brunswick, Saint John, NB, E2L 4L5, Canada. Email: spavey@unb.ca

Email: scottapavey@gmail.com

Abstract

Striped Bass, *Morone saxatilis* (Walbaum, 1792), is an anadromous fish species that supports fisheries throughout North America and is native to the North American Atlantic Coast. Due to long coastal migrations that span multiple jurisdictions, a detailed understanding of population genomics is required to untangle demographic patterns, understand local adaptation, and characterize population movements. This study used 1256 single nucleotide polymorphism (SNP) loci to investigate genetic structure of 477 Striped Bass sampled from 15 locations spanning the North American Atlantic coast from the Gulf of St. Lawrence, Canada to the Cape Fear River, United States (US). We found striking differences in neutral divergence among Canadian sites, which were isolated from each other and US populations, compared with US populations that were much less isolated. Our SNP dataset was able to assign 99% of Striped Bass back to six reporting groups, a 39% improvement over previous genetic markers. Using this method, we found (1) evidence of admixture within Saint John River, indicating that migrants from the US and from Shubenacadie River occasionally spawn in the Saint John River; (2) Striped Bass collected in the Mira River, Cape Breton, Canada were found to be of both Miramichi River and US origin ; (3) juveniles in the newly restored Kennebec River population had small and nonsignificant differences from the Hudson River; and (4) tributaries within the Chesapeake Bay showed a mixture of homogeny and small differences among each other. This study introduces new hypotheses about the dynamic zoogeography of Striped Bass at its northern range and has important implications for the local and international management of this species.

Introduction

The Striped Bass, *Morone saxatilis* (Walbaum, 1792), is a facultative anadromous and economically important fish with a native range extending along the Atlantic coast of North America from the St. Lawrence River, Quebec to the St John's River, Florida, as well as a native population in the Apalachicola–Chattahoochee–Flint river system in the Gulf of Mexico (Setzler et al. 1980, Wirgin et al. 2005; Figure 2.1). Individuals from the Hudson River to the Roanoke River can move long distances, some moving 400–1000 km along the Atlantic Coast (Mather et al. 2010, Kneebone et al. 2014, Callihan et al. 2015), and along with some Canadian populations are known to enter non-natal rivers (Grothues et al. 2009, Kneebone et al. 2014, LeBlanc et al. 2018). Migratory populations within the US are currently managed as two separate stocks: the Roanoke River, and all US populations north of the Roanoke River (Atlantic States Marine Fisheries Commission (ASMFC), 2019). Populations south of the Roanoke River and Albemarle Sound are generally considered nonmigratory (Bjorgo et al. 2000), Striped Bass in the Gulf of St. Lawrence are thought to be isolated from the rest of the range (Rulifson and Dadswell 1995), and it is unknown whether Bay of Fundy populations travel further than the Gulf of Maine (Department of Fisheries and Oceans (DFO), 2014). Striped Bass throughout its range experienced severe population declines from the 1960s to 1980s, leading to extensive temporary and permanent closures of commercial and recreational fisheries (Carmichael et al. 1998, Richards and Rago 1999, Andrews et al. 2019a). Multi-state emergency management measures implemented by the ASMFC in the US resulted in the recovery of most US populations during the 1990s and 2000s (Richards and Rago 1999, ASMFC 2019), although abundance estimates have since declined in the 2010s

(ASMFC, 2019). In Canada, the closure of commercial fisheries and restrictions on recreational fishing led to the recovery of some populations (Miramichi River and Shubenacadie River) and not others (Saint John River, Annapolis River, and St. Lawrence River; Andrews et al. 2019a).

Effective management of a highly migratory species requires knowledge of the connectivity between populations and the seasonal mixing rates of multi-origin stocks in relation to spatial dynamics within the species range. Striped Bass populations have complex life histories and often exhibit multiple migratory components, or contingents (Clark 1968, Secor 1999, Secor et al. 2001, Gahagan et al. 2015, Andrews et al. 2017), which appear tied to ontogenic development (Gahagan et al. 2015, Conroy et al. 2015) and population size (Waldman et al. 1990, Callihan et al. 2014). Partial migration, where some individuals in a population are resident while others migrate, and contingent behaviors further complicate management of Striped Bass as harvest in coastal waters may be on mixed stocks from multiple populations and also from specific behavioral subsets of those populations. Coastal migrations can facilitate genetic exchange between populations because mixing away from natal rivers may lead to some individuals straying and spawning in non-natal rivers; an infrequent yet measurable occurrence (e.g. Gauthier et al. 2013; LeBlanc et al. 2018). In anadromous fishes, straying allows moving, expanding or contracting its range in response to environmental changes (Pess et al. 2014). Striped Bass inhabiting areas once covered by the Laurentide ice sheet, i.e., the

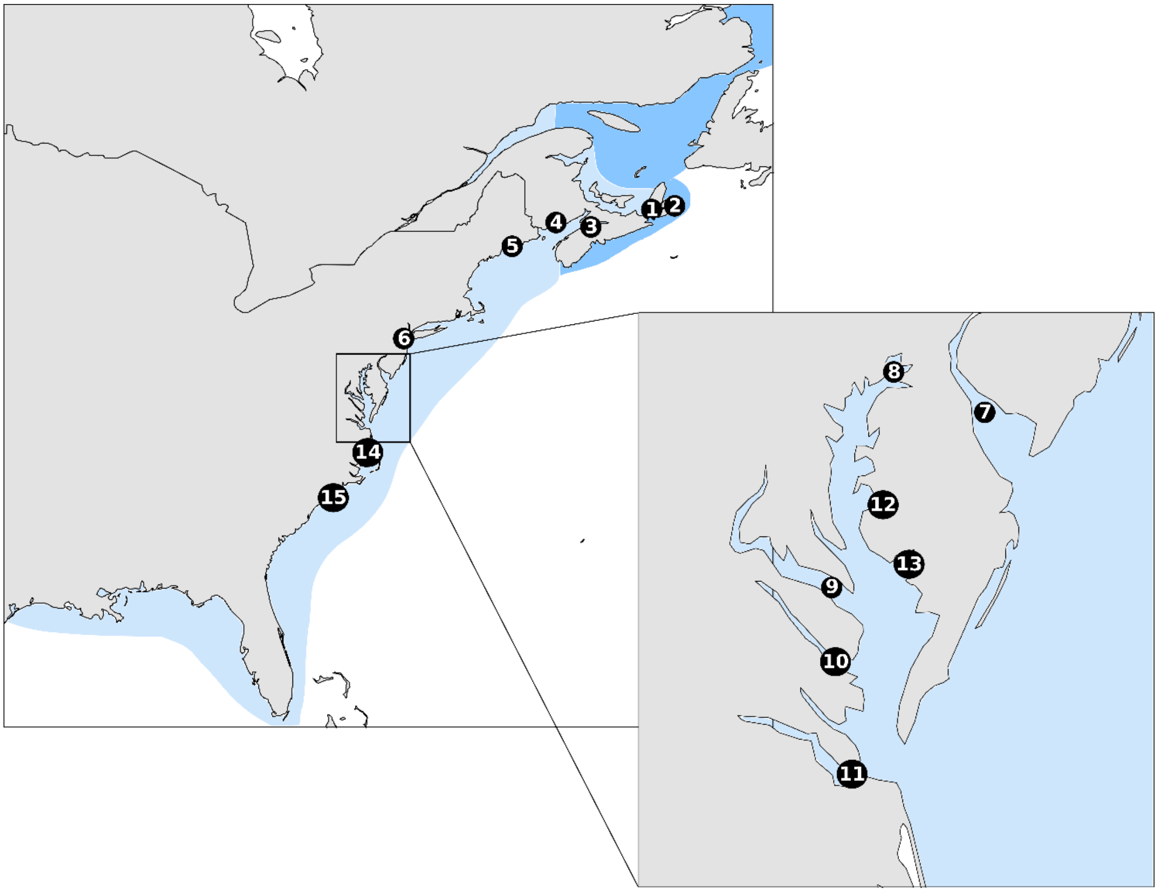


Figure 2.1. On the left, a map showing the current range of Striped Bass along the North American Atlantic Coast. Potential additions to the range where Striped Bass have been reported in or may inhabit are marked in darker green. Sampling sites marked in numbered circles as follows and listed according to the location of the river mouth. 1) Bras d'Or Lake, Nova Scotia, 2) Mira River, Nova Scotia, 3) Shubenacadie River, Nova Scotia, 4) Saint John River, New Brunswick, 5) Kennebec River, Maine, 6) Hudson River, New York/New Jersey 14) Roanoke River, North Carolina, 15) Cape Fear River, North Carolina. On the right, a close up of Delaware River and Chesapeake Bay, with sampling locations marked as follows. 7) Delaware River, New Jersey/Delaware 8) upper Chesapeake Bay, Maryland 9) Potomac River, Maryland 10) Rappahannock River, Virginia, 11) James River, Virginia, 12) Choptank River, Maryland 13) Nanticoke River, Maryland.

entirety of the Bay of Fundy and Gulf of St. Lawrence, must have descended from southern migrants colonizing these rivers within the last 10,000 years as the glaciers retreated (Pielou 1991, Curry 2007).

Over the past five decades, Striped Bass stock discrimination has been attempted using many techniques producing inconsistent results when matching individuals to more than two reference populations (Waldman and Fabrizio 1994, Waldman et al. 2012). The Chesapeake Bay is usually considered the primary source of migratory Striped Bass found along the North American Atlantic Coast, with the Hudson River occasionally providing large numbers and the Delaware and Roanoke Rivers previously considered to have a negligible contribution (Wirgin et al. 1997b, Richards and Rago 1999). Mixed-stock analyses have found that stock composition can vary dramatically. Hudson River Striped Bass can contribute 14–89% of coastal aggregations in different seasons and locations and from year to year (Fabrizio 1987, Wirgin et al. 1993a). Existing mixed-stock methods are often unable to reliably differentiate Roanoke River and Chesapeake Bay individuals and consequently the Roanoke River population is often merged with the Chesapeake Bay in reference groups, making it difficult to track relative contribution of Roanoke River Striped Bass to the current coastal groups (Waldman and Fabrizio 1994, Waldman et al. 2012). The Delaware River is often not considered in coastal stocks, because the Delaware River population is small and was not expected to contribute to coastal aggregations in previous decades (Waldman and Fabrizio 1994, Waldman et al. 2012); however, acoustic telemetry showed Delaware River Striped Bass make up 14-20% of Striped Bass caught off the coast of Massachusetts (Kneebone et al., 2014). A mixed-stock analysis that can reliably distinguish among stocks that exhibit varying degrees of mixing in the coastal environment could substantially improve Striped Bass management.

In addition to the ongoing attempts to characterize Striped Bass migration, the last decade has seen shifts in the existing range of several populations. Large-sized Striped Bass in the Roanoke River population, previously considered largely resident because few tagged fish have been caught outside the river, have been recently shown to migrate approximately 500–600 km to New Jersey (Callihan et al. 2015). Striped Bass from the Miramichi River, which is considered the only spawning population in the Gulf of St. Lawrence (Robinson et al. 2004), have been caught off the Labrador coast following a decade of strong population growth (DFO 2018, Andrews et al. 2019a). These apparent range expansions have been attributed to increased ocean temperature (DFO 2018), increased population size (Callihan et al. 2014, Andrews et al. 2019a) and an increase in the number of older, larger adults that are more likely to migrate longer distances (Callihan et al. 2014). These emerging migrations highlight the need to apply more sophisticated population discrimination tools to best inform management.

Several attempts have been made to use genetic markers in mixed-stock analysis of Atlantic Coast Striped Bass (Wirgin et al. 1993a, 1997b, Brown et al. 2005, Gauthier et al. 2013), within the Bay of Fundy (Wirgin et al. 1995), and the Gulf of St. Lawrence (Robinson et al. 2004). Two studies have comprehensively investigated the genetic structure among the major migratory populations of the North American Atlantic Coast

(Gauthier et al. 2013, Wirgin et al. 2020). Previous studies have found consistent genetic differences among known Canadian populations (Wirgin et al. 1993b, Bentzen and Paterson 2008), and lower but significant differences between regions such as the Hudson River and Chesapeake Bay (Wirgin et al. 1997b, Gauthier et al. 2013); however, rivers in close proximity to each other, particularly the Chesapeake Bay and Delaware River, have had inconsistent results (see Brown et al. 2005). Most recently, Gauthier et al. (2013) and Wirgin et al. (2020) found very low but significant differences among rivers within the Chesapeake Bay using 14 and 8 microsatellites, respectively, but both were unable to assign a high number of individuals to a river of origin.

Genotyping-by-sequencing (GBS) can be used to construct large panels of single nucleotide polymorphisms (SNPs) throughout the genome of an individual organism (Poland et al. 2012, Narum et al. 2013). The SNP panels created by GBS can discriminate among closely related populations of anadromous fishes such as Alewife (*Alosa pseudoharengus*) and Blueback Herring (*A. aestivalis*; Baetscher et al. 2017) and Atlantic Salmon (*Salmo salar*; Bourret et al. 2013). Large numbers of SNPs also facilitate identification of genes or regions showing signs of selection, by examining which of the hundreds or thousands of SNPs show significantly greater differentiation among populations (Allendorf et al. 2010). While these outlier analyses are biased toward detection of single loci with strong signals of selection over more subtle polygenic adaptation (Rockman 2012), they can serve as a starting point for identifying adaptive differences between populations. Moreover, inclusion of outlier loci in tests of population differentiation can disproportionately bias results (Allendorf and Seeb 2000, Luikart et al.

2003). Once identified, these loci can then be removed from analyses of genetic structure, migration, and effective population size, and examined separately to gain insights into adaptive selection that may be occurring in a population and highlight potential candidate genes for future studies.

In this work, we employ next-generation sequencing to examine the genetics of Striped Bass from 14 locations across the native range, from the Gulf of St. Lawrence to the southernmost edge of the migratory range in the Roanoke River. (Figure 2.1). We sample two locations (Hudson River and Delaware River) in two different years to assess temporal stability of populations. We include samples from six tributaries within the Chesapeake Bay to examine small-scale spatial differences. Also included are samples from the Cape Fear River, which has a supportive breeding program to maintain a Striped Bass population in-river, the recently restored Kennebec River, and from the Mira River on the northeastern coast of Nova Scotia, which is speculated to host a spawning aggregation of Striped Bass (Buhariwalla 2018). We assess neutral genetic structure and characteristics of SNPs that show signs of selection, and we test the ability of our SNP dataset to assign Striped Bass back to their natal population.

Methods

Sample Collection

Fin clips and scales were taken from Striped Bass from multiple collections (Table 2.1). Age of sampled individuals differed by location. YOY juveniles were individuals less than 1 year old (<15 cm long). Saint John juveniles were 1–4 years old and largely spawned in the year 2013. Ages for Saint John River juveniles were obtained from scales. Adults were sexually mature individuals aged 4 years and older. All adults collected were in spawning condition at time of sampling, except for Bras d'Or Lake, Mira River, and Shubenacadie River. Shubenacadie origin Striped Bass migrate to the Stewiacke-Shubenacadie systems from overwintering sites during the sampling period (DFO 2014, Keyser et al. 2016). Adult bass caught during this period are assumed to be of Shubenacadie River origin for the purpose of population surveys (DFO 2014). Putative Miramichi River origin Striped Bass were included using fin clips taken from Striped Bass caught in the Bras d'Or Lake, Cape Breton that have previously been examined using microsatellites and found to match the Miramichi River population (Bentzen et al. 2014). These samples will hereafter be referred to as Bras d'Or–Miramichi individuals.

Laboratory

DNA was isolated using either NucleoMag® 96 Tissue (Machery-Nagel, Düren, Germany) kit on an epMotion 5075t (Cat. 5075000302), or the E.Z.N.A. Tissue DNA Kit (Omega Bio-Tek, Doraville, CA). Libraries containing 96 individuals each were prepared using a double-digest restriction-site associated DNA sequencing (ddRAD-seq or ddRAD) protocol developed by Poland et al. (2012) and modified as described in LeBlanc et al. (2018). For samples re-used from LeBlanc (2018), existing sequence data was used in this study. For additional samples, individuals were randomized so that each lane contained individuals from multiple locations and sequenced using Illumina®

HiSeq™ 2500 or Illumina® HiSeq™ 4000 (San Diego, U.S.A.) at Génome Québec Innovation Centre.

Table 2.1. Number, collection date, type of tissue, and age of fish for each of 15 locations Striped Bass samples were collected in.

Location	n	Date collected	Type	Age
BD-MICHI	19	June-Nov 2012-2014	Fin Clips	Adults
MIRA	22	April-June 2013-2017	Fin Clips	Adults
SHUB	33	2014-2017	Scales	Adults
SJR	32	July-Sept 2014-2017	Fin clips	Juveniles
KEN	16	April-May 2012	Fin clips	Juveniles (YOY)
HUD 2012	34	May, 2014	Fin clips	Adults, spawning condition
HUD 2014	21	April-May, 2012	Fin clips	Adults, spawning condition
DEL 2012	28	April 2014	Fin clips	Adults, spawning condition
DEL 2014	29	April 2012	Fin clips	Adults, spawning condition
CHPK	27	July, Sept 2011	Fin clips	Adults, spawning condition
POT	33	April-May 2014	Fin clips	Juveniles (YOY)
RAPP	32	April 20124	Fin clips	Adults, spawning condition
JAMES	33	August, Sept 2011	Fin clips	
CHOP	33	June, Dec 2011	Fin clips	Juveniles (YOY)
NANTI	33	April 2014	Fin clips	Juveniles (YOY)
ROA	30	April-May 2015	Fin clips	Adults, spawning condition
CF	22	April-May 2015	Fin clips	Adults, spawning condition

Abbreviations: BD-MICHI = Bras d'Or-Miramichi, MIRA = Mira River, SHUB = Shubenacadie River, SJR = Saint John River, KEN = Kennebec River, HUD = Hudson River, DEL = Delaware River, CHPK = Upper Chesapeake Bay, POT = Potomac River, RAPP = Rappahannock River, JAMES = James River, CHOP = Choptank River, NANTI = Nanticoke River, ROA = Roanoke River, CF = Cape Fear

Quality Control and Analysis

SNPs were demultiplexed and filtered using modified versions of Eric Normandeau's Stacks workflow scripts, available on github (https://github.com/enormandeau/stacks_workflow, downloaded August 2016). Cutadapt v. 1.13 (Martin 2011) was used to trim adapters from the raw sequences using a maximum error rate (e) of 0.2 and a minimum read length (m) of 50. FastQC v. 0.11.5 (Babraham Bioinformatics) was used to assess sequence quality before and after. Sequences were then trimmed to a uniform length of 85 bp and demultiplexed using the *process radtags* module of Stacks v. 1.46 (Catchen et al. 2013) using the paired end option –P. BWA version 0.7.15 (Li and Durbin 2010) was used to align sequences to the Striped Bass genome (BioProject accession number PRJNA266827) using a minimum seed length (k) of 19, a maximum seed occurrence of 500, and otherwise default parameters. Samtools was used to exclude alignments with a mapq score smaller than 1, or the following flags: 4, 256, 2048. The stacks module *pstacks* identified reference aligned loci with a minimum depth (m) of 4 using the "snp" model type and an alpha of 0.1. Loci were assembled into a catalogue using *cstacks*, *sstacks*, and *rxstacks* with default settings, and unclear or unlikely haplotypes, as well as SNPs with a log likelihood < 45, were pruned from the dataset. Using the *populations* module, SNPs were further filtered to remove all loci with a stack depth <5, with greater than 20% missing data in any given location, and any loci not amplified in all locations. We examined the output of populations and removed loci with an Fis < -0.3 to eliminate possible paralogs, and used VCFTools 0.1.13 (Danecek et al. 2011) to remove any loci with a minor allele frequency < 0.01, and plink v. 1.90 (Chang et al. 2015) was used to remove loci with an r2 above 0.2.

Structure files created by Stacks were converted to the appropriate input files for downstream analyses using PGDSpider v. 2.1.1.0 (Lischer and Excoffier 2012). Sibship analyses were carried out in Colony2 v. 2.0.6.5 (Jones and Wang 2010) on each population separately to ensure individuals were not closely related. Full sibling pairs identified with a probability of >0.5 were removed from subsequent analyses. Percent polymorphism of loci in each population was reported by the Stacks *populations* module, and expected and observed heterozygosity were calculated using the R package *adegenet* v. 2.1.1 (Jombart 2008).

An initial pairwise F_{ST} analysis was conducted in Arlequin v. 3.5.2.2 (Excoffier and Lischer 2010), with significance assessed using 10,000 random permutation tests. Individuals caught in the Hudson River and Delaware River in 2012 and 2014 were grouped by location and year in order to assess whether the genetic profile of each location differed from year to year. After confirming no significant differences between years, the two sampling years were pooled together for outlier analyses.

Constructing a neutral SNP panel and assessing adaptive selection

Outlier loci were removed prior to subsequent population genetic analyses, and a subset of outliers were examined separately. Existing outlier analyses are known to detect high numbers of false positives alongside true outlier loci (De Villemereuil et al. 2014, Lotterhos and Whitlock 2014), and a common method of controlling for this is to examine which loci are flagged as having non-neutral divergence patterns by more than one analysis software (De Villemereuil et al. 2014). In the absence of outgroup genotypes or known neutral loci, we assessed whether any of our SNPs were under balancing or divergent selection using two methods. BAYESCAN v. 2.1 (Foll and Gaggiotti 2008) was run with 100,000 iterations, using a burn-in of 50,000, a thinning interval of 10, and a sample size of 5K. Prior odds were set to 1,000 to minimize false positives while retaining power to detect outliers (Lotterhos and Whitlock 2014). We also used the recently developed R package OutFLANK (Whitlock and Lotterhos 2015) with Hmin > 0.1 to identify an additional set of outliers. Unlike previous outlier tests like Bayescan, outFLANK uses distribution of allele frequencies across all loci to account for differences in genetic structure among populations (Whitlock and Lotterhos 2015). Loci identified as outliers at a q-value <= 0.05 by either method were removed to create a dataset of putatively neutral loci for genetic structure analyses. Loci identified as outliers by both methods were mapped to one of 35,010 scaffolds contained in the published Striped Bass genome using the JBrowse genome browser (Skinner et al. 2009) to identify associated genes showing signatures of selection, and allele frequencies were calculated in Arlequin to investigate divergence patterns across populations.

Connectivity of Striped Bass locations through population genetic structure

Population structure was assessed using both traditional F_{ST} and clustering algorithms. Overall hierarchal F_{ST} and pairwise F_{ST} was calculated in Arlequin, and significance was assessed using 10,000 random permutation tests. Hierarchal population groupings for overall F_{ST} were made based on patterns of differentiation seen in clustering analyses and previous studies. Pairwise F_{ST} values were also calculated on all locations, and pairwise significance was assessed using a chi-square test implemented in the R package *strataG* v. 2.1 (Archer et al. 2017), and corrected to account for multiple tests using the False

Discovery Rate method detailed in Benjamini & Hochberg (1995). Chi-square tests have high power and low false positive rates when used on large numbers of bi-allelic loci, as found in SNP datasets (Ryman et al. 2006).

Isolation-by-distance (IBD) was assessed using mantel tests implemented in Arlequin v. 3.5.2.2 (Excoffier and Lischer 2010). Isolation-by-distance was assessed on all locations, on only Canadian locations, only US locations, and on locations within Chesapeake Bay and Delaware River, using approximate distance between rivers. When calculating distances between rivers, we assumed that Striped Bass make use of the Cape Cod Canal, and that Striped Bass move between the Chesapeake Bay and the Delaware River via the Chesapeake and Delaware Canal based upon tagging study results (Kneebone et al. 2014, Gahagan et al. 2015).

The R package LEA v. 2.0 (Frichot and François 2015) estimates ancestry coefficients for all individuals using sparse Non-Negative Matrix Factorization (sNMF), an algorithm that has been optimized for use with large numbers of genetic markers. Scenarios were run using 1–20 theorized number of distinct genetic populations (K), with 10 repetitions per K value, on all individuals as well as on only US individuals. To ensure results were not biased by differences in sampling size, sNMF was run a second time with a maximum of 30 individuals from any given genetic cluster found in the initial run. The probability of a K being valid was calculated using the cross-entropy criterion. The K values with the lowest minimal cross-entropy value were considered most probable as the true number of ancestral populations (Frichot et al. 2014). Where the lowest entropy was unclear, clustering results for the lowest K values were manually inspected for informative grouping and consistency across repetitions. Population structure was also assessed using Discriminant Analysis of Principal Components (DAPC; Jombart et al. 2010), implemented in the R package *adegenet*, using 1-20 assumed clusters (K). The number of putative clusters with the lowest Bayesian information criterion value was chosen to evaluate population groupings. Another DAPC analysis was conducted with samples taken from Canadian rivers excluded, using the same methods described above.

Assessing the power of SNPs and reference pool for population assignment

We tested whether our SNP panel could accurately assign individuals to populations of origin using a leave-one-out protocol implemented in GeneClass2 v. 2.0 (Piry et al. 2004), using the Rannala & Mountain 1997 Bayesian method (Rannala and Mountain 1997). Assignment success was compared to results from another genetic assignment algorithm implemented in the R package *rubias*, again using a leave-one-out protocol. Using this protocol, each individual is assigned to a region using a reference panel composed of all individuals except the one being tested.

We tested assignment success of all sample locations separately, as well as assignment to pooled groups according to previous population groupings used in Gauthier et al. (2013). In both cases, we considered an individual assigned to a population if the confidence score for assignment to that population was 80% or above.

Results

Filtering

The initial SNP catalogue contained 756,713 loci. After filtering for Ln Likelihood less than -40, the catalogue contained 670,167 loci. After filtering out loci with stack depths of less than five, more than 20% missing data, more than two alleles, and loci present in fewer than 17 populations, the SNP catalogue contained 7884 loci. After filtering by F_{IS} values, removing loci with minor allele frequencies less than 0.01 and loci in linkage disequilibrium with at least one other locus, and removing non-polymorphic loci, we had 1291 loci. Average read depth across loci for each individual was 55 (range = 9–171), and average read depth for loci across all individuals was 55 (range = 17–131). Sibship analysis found two possible full sibship pairs in the Kennebec River and three in the Chesapeake Bay; one individual from each pair was removed. In addition, the Cape Fear dataset contained three possible full sibship pairs, one trio, and one group of five individuals. One individual from each group was retained and the rest were excluded from downstream analyses.

Expected and observed heterozygosity levels ranged from 0.25 to 0.38 and observed heterozygosity did not deviate more than 0.02 from expected heterozygosity in any location. Canadian rivers had slightly lower observed heterozygosity (H_O = 0.26–0.33) compared to rivers south of the Bay of Fundy (H_O = 0.35–0.38; Table 2.2). Similarly, individuals in the Bras d’Or–Miramichi and Shubenacadie River populations had the lowest proportion of polymorphic loci among all locations; almost one quarter of all loci genotyped were fixed in Bras d’Or–Miramichi individuals (Table 2.2). All other sampled locations had >95% polymorphic loci.

Table 2.2. Table shows summary statistics of Striped Bass samples from 15 locations amplified at 1,291 SNP loci. Values obtained when all samples are included are in brackets.

		# poly.			
Location	**n**	**loci**	**% poly.**	**Ho**	**He**
BD-MICHI	19	987	0.76	0.26	0.25
MIRA	22	1234	0.96	0.27	0.29
SHUB	33	1126	0.87	0.28	0.28
SJR	32	1271	0.98	0.33	0.32
KEN	16	1281	0.99	0.36	0.35
HUD 2012	34	1289	1.00	0.36	0.36
HUD 2014	21	1282	0.99	0.37	0.36
DEL 2012	28	1287	1.00	0.36	0.36
DEL 2014	29	1289	1.00	0.37	0.36
CHPK	27	1286	1.00	0.38	0.36
POT	33	1289	1.00	0.36	0.36
RAPP	32	1288	1.00	0.38	0.36
JAMES	33	1290	1.00	0.37	0.36
CHOP	33	1283	0.99	0.36	0.36
NANTI	33	1285	1.00	0.36	0.35
ROA	30	1288	1.00	0.37	0.37
CF	22	1277	0.99	0.36	0.35

Abbreviations: N = Number of individuals, # poly. Loci = # of loci that are polymorphic within a population, % poly. = proportion of loci that are polymorphic in a population, Ho = observed heterozygosity, He = expected heterozygosity. BD-MICHI = Bras d’Or-Miramichi, MIRA = Mira River, SHUB = Shubenacadie River, SJR = Saint John River, KEN = Kennebec River, HUD = Hudson River, DEL = Delaware River, CHPK = Upper Chesapeake Bay, POT = Potomac River, RAPP = Rappahannock River, JAMES = James River, CHOP = Choptank River, NANTI = Nanticoke River, ROA = Roanoke River, CF = Cape Fear.

Outliers

Outlier analyses identified 35 total outlier loci: Bayescan identified 13 loci as possible outliers, compared to 25 loci found by outflank, and 3 loci were identified by both analyses. All 35 potential outliers were excluded from downstream genetic structure analyses, while the three loci identified by both approaches were examined further as putative adaptive loci. These three loci were located on three different scaffolds and were given names according to their scaffold number and base pair position on the scaffold (scaffold_bp). Locus 4437_41108 is located 41,108 base pairs into a large (77,288 bp) scaffold, Msax_4437, inside an intron of insulin-like growth factor 2b *(igf2b)*. The remaining two outliers, 25891_222 and 27535_2519, were located on short (2825 and 5316 bp, respectively) scaffolds with no known genes.

Examination of allele frequencies of the three putative outliers revealed that all three loci possessed one allele that was at or near fixation in individuals within Gulf of St. Lawrence and Shubenacadie River and at very low frequencies in US locations (Figure 2.2), with maximum allele frequency differences of 0.85–0.98. In 25891_222 and 27535_2519, allele frequencies in Saint John River fish were slightly lower than other Canadian locations, while the major Canadian allele of 4437_41108 was present at about a frequency of 0.50 in the Saint John River, Kennebec River, and Hudson River (Figure 2.2).

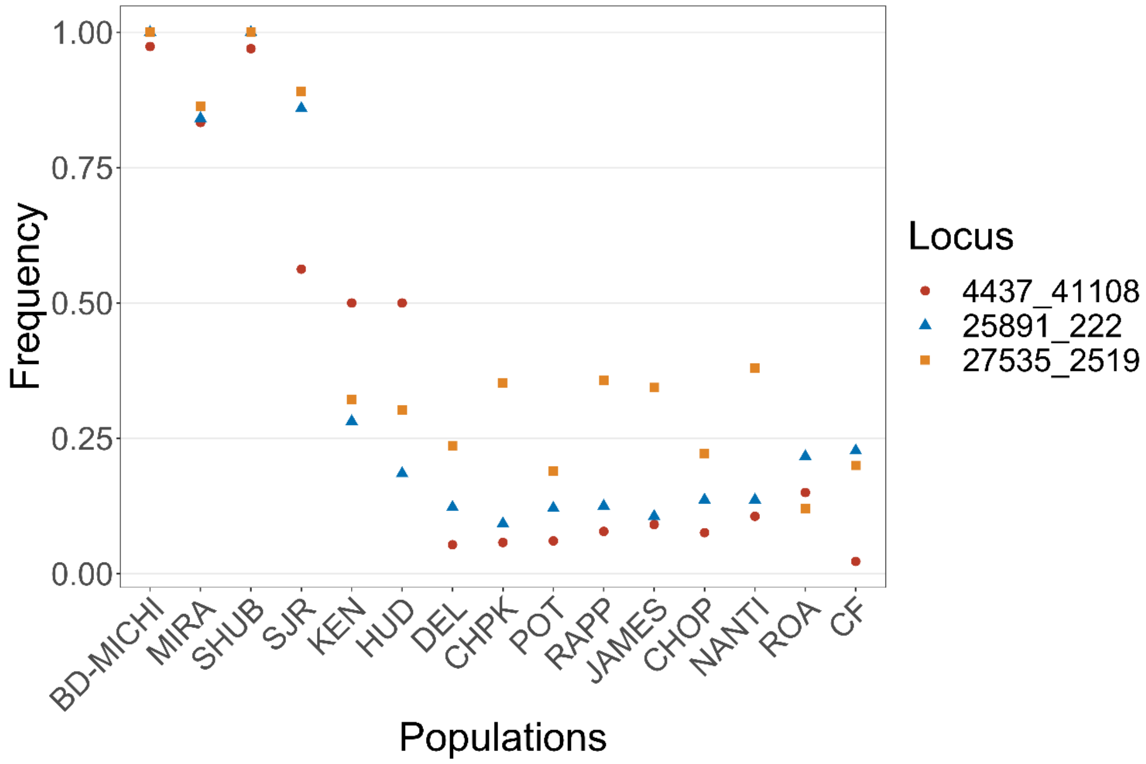


Figure 2.2. Allele frequencies of three loci identified as outliers in Bayescan and outFLANK in Striped Bass populations along the North American Atlantic Coast. The frequency of the major allele of each locus in Bras d'Or-Miramichi Striped Bass is plotted across 15 locations, displayed from north to south. BD-MICHI = Bras d'Or-Miramichi, MIRA = Mira River, SHUB = Shubenacadie River, SJR = Saint John River, KEN = Kennebec River, HU = Hudson River, DEL = Delaware River, CHPK = Upper Chesapeake Bay, POT = Potomac River, RAPP = Rappahannock River, JAMES = James River, CHOP = Choptank River, NANTI = Nanticoke River, ROA = Roanoke River, CF = Cape Fear.

Population Structure

Population structure analyses were conducted using 1256 SNPs deemed to be neutrally evolving after outlier analyses. Overall F_{ST} was 0.086 and highly significant (p<0.001), while pairwise values varied from 0 to 0.20. Correction for multiple testing did not change significance of any pairwise F_{ST} values. Canadian populations tended to be highly genetically distinct while populations in the US migratory range were less genetically differentiated. The highest pairwise F_{ST} values occurred between Canadian rivers and all other locations (save for F_{ST} between the Bras d'Or–Miramichi individuals and Mira River), with F_{ST} values ranging from 0.09 to 0.20. Pairwise F_{ST} between Mira River and Bras d'Or–Miramichi River was 0.007 and non-significant (p-value = 0.08; Table 2.3). Among the three US regions identified with genetic clustering, F_{ST} values ranged from 0.012 to 0.035, while within-region F_{ST} values were lower (0 to 0.011). Within the Delaware River–Chesapeake Bay region, 20 of 28 comparisons were significant (Table 2.3). The majority (18) of significant comparisons were between James River individuals and all other locations in this region, as well as between populations in the Nanticoke and Choptank Rivers, the only two rivers sampled along the eastern side of the Chesapeake Bay, and all other locations within the region. The Delaware River, upper Chesapeake Bay, the Potomac River, and the Rappahannock River all had very low and mostly non-significant F_{ST} values with each other. The Kennebec River had very low (0.004 to 0.008) but significant ($p < 0.01$) pairwise F_{ST} compared with Hudson River individuals. Similarly, the Roanoke River and Cape Fear River had small ($F_{ST} = 0.004$) but significant differences ($p < 0.001$; Table 2.3).

	BD-					HUD	HUD	DEL	DEL								
	MICHI	MIRA	SHUB	SJR	KEN	2012	2014	2012	2014	CHPK	POT	RAPP	JAMES	CHOP	NANTI	ROA	CF
BD-MICHI	*																
MIRA	0.007	*															
SHUB	0.201**	0.164**	*														
SJR	0.179**	0.134**	0.127**	*													
KEN	0.186**	0.132**	0.149**	0.091**	*												
HUD 2012	0.179**	0.130**	0.153**	0.094**	0.008*	*											
HUD 2014	0.183**	0.131**	0.156**	0.092**	0.004**	0.000	*										
DEL 2012	0.187**	0.134**	0.160**	0.095**	0.014**	0.015**	0.012**	*									
DEL 2014	0.182**	0.129**	0.154**	0.093**	0.012**	0.017**	0.014**	0.001	*								
CHPK	0.188**	0.133**	0.161**	0.097**	0.014**	0.017**	0.013**	0.000	0.000	*							
POT	0.179**	0.127**	0.154**	0.093**	0.013**	0.014**	0.011**	0.002*	0.000	0.002*	*						
RAPP	0.182**	0.130**	0.155**	0.094**	0.015**	0.018**	0.014**	0.000	0.001	0.000	0.002*	*					
JAMES	0.183**	0.131**	0.156**	0.098**	0.013**	0.015**	0.011**	0.003**	0.003**	0.004**	0.001*	0.004**	*				
CHOP	0.187**	0.134**	0.161**	0.097**	0.020**	0.024**	0.018**	0.002*	0.003**	0.002*	0.008**	0.001*	0.010**	*			
NANTI	0.191**	0.137**	0.161**	0.100**	0.022**	0.024**	0.021**	0.003**	0.004**	0.002*	0.008**	0.002*	0.011**	0.000	*		

ROA	0.187**	0.136**	0.155**	0.099**	0.028**	0.027**	0.027**	0.022**	0.022**	0.025**	0.021**	0.024**	0.020**	0.029**	0.029**	*	
CF	0.193**	0.140**	0.161**	0.101**	0.031**	0.029**	0.027**	0.027**	0.025**	0.028**	0.025**	0.028**	0.025**	0.034**	0.035**	0.004**	*

Table 2.3. Pairwise F_{ST} comparisons of Striped Bass samples from 15 locations using 1,256 putatively neutral SNP loci, with significance calculated using a chi-square test.

** indicates p-value < 0.001; * indicates p-value < 0.05. BD-MICHI = Bras d'Or-Miramichi, MIRA = Mira River, SHUB = Shubenacadie River, SJR = Saint John River, KEN = Kennebec River, HUD = Hudson River, DEL = Delaware River, CHPK = Upper Chesapeake Bay, POT = Potomac River, RAPP = Rappahannock River, JAMES = James River, CHOP = Choptank River, NANTI = Nanticoke River, ROA = Roanoke River, CF = Cape Fear.

Isolation-by-distance (IBD) and differentiation patterns differed among Canadian and US locations. Significant isolation-by-distance was found when all locations were considered (r = 0.84, p < 0.001). Within Chesapeake Bay, there was no significant isolation by distance pattern (r < 0.001, p = 0.50), but when US locations in North Carolina, Hudson River, and Kennebec River were included IBD became significant (r = 0.61, p = 0.03). When only Canadian populations were considered, there was no significant IBD (r = 0.49, p = 0.21). When all samples were run using DAPC, the most likely number of clusters was four, as estimated using the Bayesian Information Criterion (Appendix B). Canadian Striped Bass formed three groups, and all US Striped Bass were assigned to a fourth group (Figure 2.3; Table 2.4). This general pattern was seen when DAPC was run assuming five and six genetic groups (Appendix C). Using LEA, the number of genetic clusters (K) with the lowest entropy across ten runs was six (Appendix B). We visualized clustering patterns for K values 4 through 7 to identify hierarchal clustering patterns as K increases (Figure 2.4). In all simulations, Canadian Striped Bass clustered into the same three groups as in DAPC. North Carolina rivers separated into their own cluster at K=5, while Kennebec River and Hudson River separated at K=6, and at K=7 the two rivers on the eastern coast of Chesapeake Bay (Nanticoke River and Choptank River) primarily belong to the seventh cluster (Figure 2.4). The same clustering pattern was seen when US samples were analyzed separately from Canadian samples (Appendix D, E). When LEA was run with balanced sampling numbers, the lowest entropy was K = 4 as seen in DAPC analyses. Canadian locations clustered into three regions, while all US Striped Bass were clustered together. Mean assignment per location remained high when K was increased to 6, with the same clustering pattern seen in the full dataset (Appendix D).

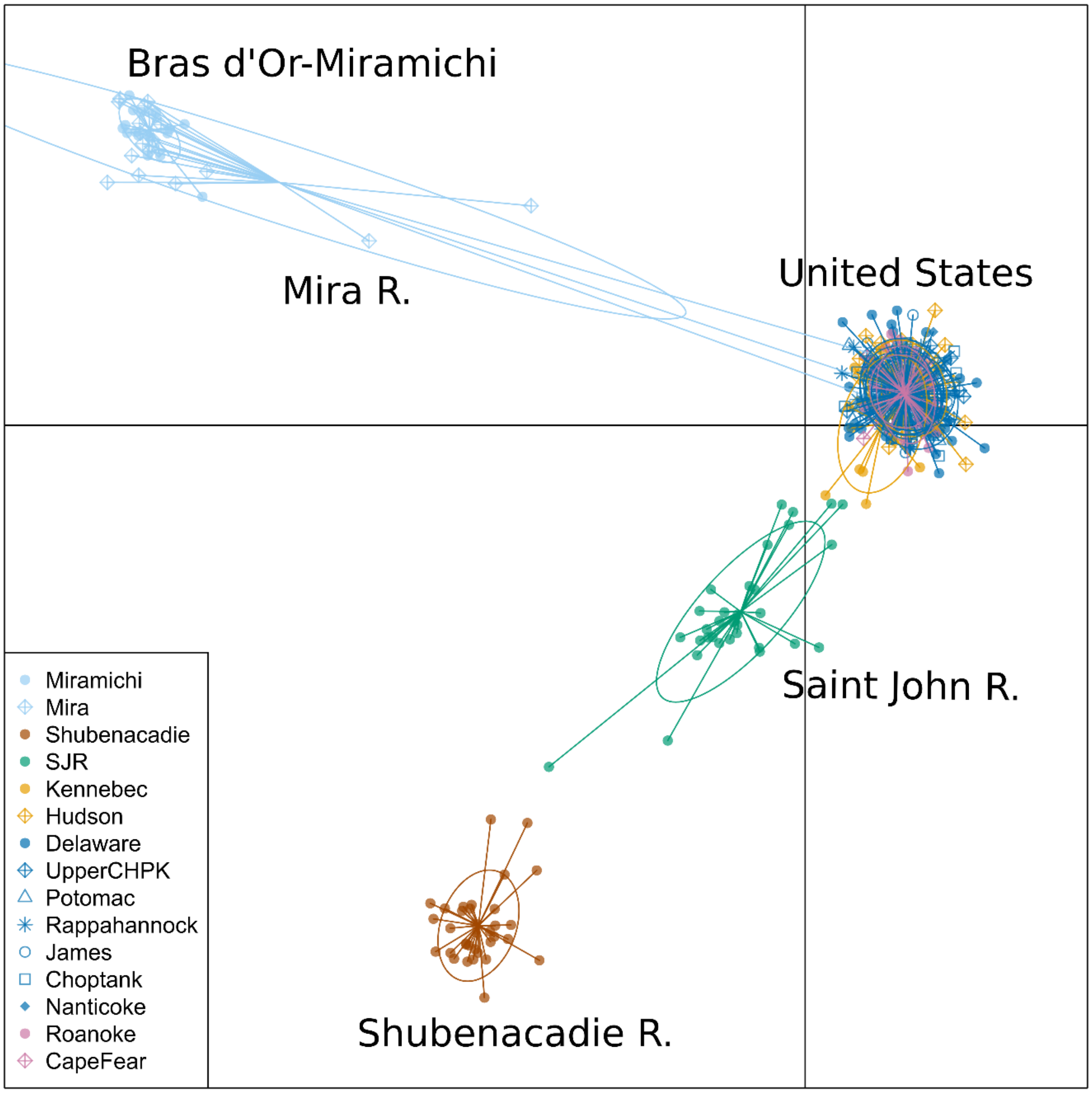


Figure 2.3. DAPC plot of Striped Bass, *Morone saxatilis*, populations collected in 15 locations, constructed using 1,256 SNPs. Individual Striped Bass are represented by symbols depicted in the legend, and a line connects the dot to the site it was sampled in. Distance between dots corresponds to genetic distance along two discriminant functions. Major groupings are labelled according to which populations are contained within.

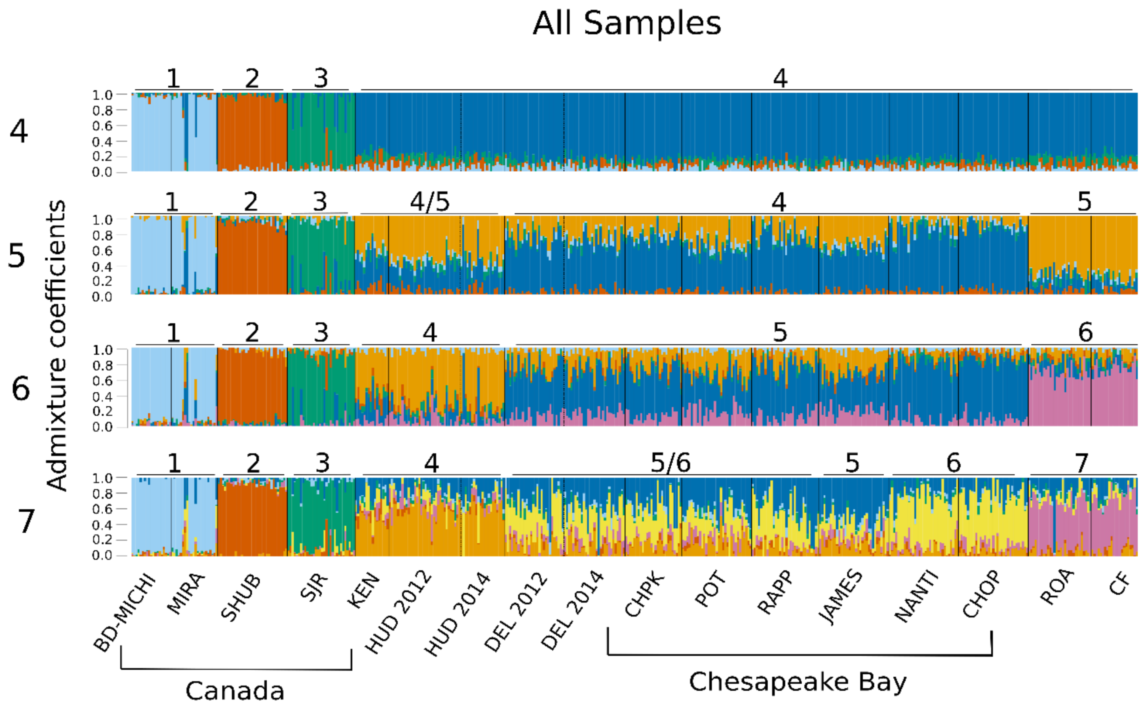


Figure 2.4. Individual admixture coefficients of Striped Bass in each site to 4, 5, 6, and 7 genetic clusters. Individual Striped Bass are represented by vertical bars, with percent genotype similarity to each cluster represented by colours. Clusters are numbered and populations are labelled with the cluster they most resemble. Population shorthands are as follows: BD-MICHI = Bras d’Or-Miramichi, MIRA = Mira River, SHUB = Shubenacadie River, SJR = Saint John River, KEN = Kennebec River, HUD = Hudson River, DEL = Delaware River, CHPK = Upper Chesapeake Bay, POT = Potomac River, RAPP = Rappahannock River, JAMES = James River, CHOP = Choptank River, NANTI = Nanticoke River, ROA = Roanoke River, CF = Cape Fear.

Table 2.4. Mean ancestry coefficients of Striped Bass from 15 locations to 6 genetic clusters identified by LEA, using 1,256 putatively neutral SNP loci.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
BD-MICHI	0.94	0.01	0.01	0.01	0.01	0.01
MiraRiver	0.79	0.02	0.02	0.06	0.1	0.02
Shubenacadie	0.03	0.9	0.03	0.02	0.01	0.01
SJR	0.02	0.05	0.79	0.03	0.08	0.02
Kennebec	0.04	0.09	0.05	0.52	0.23	0.07
Hudson12	0.04	0.05	0.04	0.67	0.12	0.08
Hudson14	0.04	0.04	0.04	0.66	0.16	0.05
Delaware12	0.04	0.03	0.04	0.19	0.59	0.12
Delaware14	0.04	0.05	0.03	0.19	0.55	0.14
Chesapeake	0.04	0.03	0.03	0.19	0.61	0.1
Potomac	0.04	0.03	0.05	0.25	0.47	0.16
Rappahannock	0.03	0.03	0.03	0.16	0.64	0.1
James	0.04	0.03	0.02	0.26	0.46	0.19
Choptank	0.05	0.04	0.05	0.09	0.73	0.05
Nanticoke	0.03	0.05	0.03	0.05	0.75	0.08
Roanoke	0.03	0.04	0.04	0.1	0.12	0.68
CapeFear	0.03	0.03	0.03	0.09	0.08	0.73

Due to the genetic distinctness of Canadian Striped Bass, it was possible to identify both putative migrants and admixed individuals within these populations. Admixture was seen in a single individual in the Mira River and eight individuals in the Saint John River. Admixed individuals within the Saint John River had approximately equal assignment to the Saint John River cluster and either the Chesapeake Bay or Shubenacadie clusters. The single admixed individual in the Mira River had equal assignment to the Bras d’Or–Miramichi cluster and to both the Hudson River and Chesapeake Bay. Additionally, three individuals caught in the Mira River appear to be migrants from Hudson River or Chesapeake Bay. When migrants and admixed individuals were removed from Mira

River both observed and expected heterozygosity for this population became 0.25, and the proportion of polymorphic loci dropped to 75% from 96%. Genetic variation within the Saint John River population also lowered slightly with the removal of admixed individuals: observed and expected heterozygosity became 0.31 and percent polymorphism became 94% from 98%. Significance of pairwise F_{ST} values remained the same with and without migrants and admixed individuals. F_{ST} between Mira River and Bras d'Or–Miramichi Striped Bass became 0.001.

Assignment

Analyses were performed at two spatial resolutions to determine the geographic scale to which reliably natal assignments could be made. When individuals were compared to all 15 collection locations in GeneClass 2, 53% were assigned back to their collection location (Appendix F). Of the remaining individuals, 71% were assigned to a different river including 91% of Mira River Striped Bass assigned to Bras d'Or-Miramichi, 59% of Cape Fear River Striped Bass assigned to the Roanoke River, and 48% of Striped Bass from the Delaware River and the Chesapeake Bay assigned to a different river in that area. Assignment patterns seen in North Carolina and Delaware River–Chesapeake Bay correspond to the geographic regions used in previous studies, which grouped North Carolina rivers together, and the Delaware River with all Chesapeake Bay rivers (Gauthier et al. 2013). Assignment success and accuracy was similar when analyzed with *rubias*, with 51% of individuals assigned back to their collection location and 76% of the remaining individuals assigned with high likelihood to a different river.

In the second analysis, Striped Bass were pooled into six geographic regions or proposed reporting groups: the Gulf of St. Lawrence, Shubenacadie River, Saint John River, Kennebec–Hudson River, Delaware–Chesapeake Bay, and Roanoke–Cape Fear. The three regions containing US Striped Bass correlate to groupings made by Gauthier et al. (2013), and all six regions correspond to one of the six genetic clusters identified in LEA. Under this scenario, both GeneClass2 and *rubias* assigned 99% of individuals to a reporting group: 97% to the group from which it was collected, and 2% to a different group (Table 2.5).

Table 2.5. Self-assignment of Striped Bass samples from 6 regions (proposed reporting groups) in GeneClass2 using 1,256 putatively neutral SNP loci. Individuals were considered to belong to a reporting group if they were assigned with a confidence score of 80% or more. Rows correspond to the location individuals were collected in, while columns correspond to assigned reporting group.

Location	GoSL	SHUB	SJR	KEN-HUD	DEL-CHPK	N. Carolina	Unassigned
GoSL	37				4		
SHUB		33					
SJR			28		2		2
KEN-HUD				65	4	1	1
DEL-CHPK					247		1
ROA-CF					1	51	

GoSL = Gulf of St. Lawrence, SHUB = Shubenacadie River, SJR = Saint John River, KEN-HUD = Kennebec River and Hudson River, DEL-CHPK = Delaware River and all rivers within Chesapeake Bay, ROA-CF = Roanoke River and Cape Fear River.

Discussion

We present the most complete examination of Striped Bass genetic structure across their native range using SNP loci. Previous genetic studies of Striped Bass have used genetic markers such as RFLPs (e.g. Wirgin et al. 1993a, 1997a), VNTRs (Laughlin and Turner 1996), mitochondrial sequences (e.g. Wirgin et al. 1997a), microsatellites (Brown et al. 2005, Bentzen and Paterson 2008, Gauthier et al. 2013), and SNP loci (LeBlanc et al. 2018) to assess the genetic structure of Striped Bass across portions of its range. Only two studies have included a thorough coverage of all major migratory populations (Hudson River to Roanoke River), and of those only Wirgin et al. (2020) has included Canadian populations. In addition, this is the first published study to include samples from Mira River, which hosts a largely unknown group of Striped Bass that may support a spawning aggregation (Buhariwalla 2018, Andrews et al. 2019a), and to document the presence of US origin Striped Bass on the northeastern coast of Nova Scotia. Our study found significant genetic structure partitioned into six regions, and much greater differentiation of Canadian regions from each other and all US regions. Our SNP panel was able to assign Striped Bass to one of these six regions with a 99% success rate. We also identified three SNP loci that show signs of selection across the sampled Striped Bass range.

Genetic diversity

Genetic diversity was highest among sampled US locations and lowest in the relatively isolated populations in the Gulf of St. Lawrence and the Shubenacadie River. All sampled US populations had similar mean observed heterozygosity, including the small and hatchery-supported Cape Fear River, suggesting that genetic variation comparable to the rest of the US range is being maintained in this population. The Kennebec River population similarly does not show signs of reduced genetic diversity, despite the Kennebec River's recent restoration with stocked Hudson River Striped Bass (G. Whippelhauser, pers. comm.). Within Canada, the highest observed heterozygosity was seen in the Saint John River, only slightly lower than values seen in the US, suggesting this population has retained substantial genetic diversity despite its apparent decline in numbers since the 1970s (Andrews et al. 2017, LeBlanc et al. 2018). Diversity values calculated for all populations were comparable to patterns seen in a recent range-wise microsatellite study (Wirgin et al. 2020). Lower genetic diversity in northern, previously glaciated locations has been seen in other anadromous fish species such as American Shad (*Alosa sapidissima*; Hasselman et al. 2013), and is expected in populations on the edge of a range expansion (Hewitt 1996, Bernatchez and Wilson 1998). The lowest observed heterozygosity values seen in this study (0.26–0.28) were similar to observed heterozygosity seen in other anadromous fishes examined using SNP markers, such as Blueback Herring (Ho=0.28–0.30), Alewife (Ho=0.22–0.27; Baetscher et al. 2017), and Chinook Salmon (*Oncorhynchus tshawytscha;* Ho=0.26–0.32; Clemento et al. 2011).

Outlier loci represent regions of major effect

Most ecologically relevant traits are thought to be polygenic, involving small allele frequency differences of many genes (Pavey et al. 2015, Yeaman 2015). All three outliers identified in this study showed high allele frequency changes among populations (0.85–0.97 maximum allele frequency differences), which suggests the presence of single genes or regions of major effect. The two un-annotated outlier loci identified in this study (25891_222 and 27535_2519) showed a strong tendency toward fixation for one allele in Canadian populations, and very low frequency of that allele in southern populations, while locus 4437_41108 showed a tendency toward fixation of allele *A* in Shubenacadie River and Gulf of St. Lawrence Striped Bass and very high allele frequencies of allele *B* in populations south of the Hudson River. Within Striped Bass in the Saint John River, Kennebec River, and Hudson River, in contrast, both alleles were maintained at relatively equal frequency. Further characterization of the Striped Bass genome and anchoring of existing scaffolds into linkage groups will be invaluable for placing these outlier loci into a wider genomic context, and sequencing of *igf2b* in northern vs southern individuals will shed light on whether locus 4437_41108 is associated with non-synonymous mutations within this gene.

Canadian populations are highly distinct

Rivers near the Gulf of St. Lawrence (Bras d'Or–Miramichi and Mira River), the Shubenacadie River, and the Saint John River were consistently, highly differentiated from each other and from US populations (F_{ST} = 0.13–0.20). Phylogeographic theory predicts that populations founded after the last glacial retreat will show less intraspecific divergence than their southern counterparts (Bernatchez and Wilson 1998). Unexpectedly high divergence in Canadian populations has been seen in other anadromous fishes along the North American Atlantic coast, and has been attributed both to the circuitous coastline created by the Nova Scotia peninsula as well as a complex hydrography within the Bay of Fundy that drives differentiation of native fish populations (McConnell et al. 1997, King et al. 2001, Hasselman et al. 2013). Variation in habitat is known to drive differentiation of anadromous fish species such as Atlantic Salmon (Bradbury et al. 2014) and Dolly Varden Char (*Salvelinus malma*; Bond et al. 2014). The Shubenacadie River, in particular, is the only tidal bore river wherein Striped Bass are known to successfully spawn (Rulifson and Dadswell 1995), and the extreme environmental conditions that eggs and larvae must tolerate in this river may contribute to its increased population differentiation (Rulifson and Tull 1999). Unexpectedly high genetic divergence in Canadian populations could also be the result of small initial colonization sizes driving changes in allele frequencies that persist to the present day (Excoffier and Ray 2008).

Genetic similarity between the Mira River and Bras d'Or-Miramichi Striped Bass indicate that these two groups have the same origin. It is likely that Striped Bass currently residing in the Mira River migrated from the Miramichi River at some point after the formation of suitable estuarine habitat and nursery areas some 500–800 years ago (Andrews et al. 2019a). While Striped Bass in the Mira River appear behaviorally distinct, demonstrating multiannual residency and spring upstream migration shown in an acoustic telemetry study in 2012–2015 (Buhariwalla 2018, Andrews et al. 2019a), our data suggest that this potential spawning population is not genetically distinct from Striped Bass found within the Gulf of St. Lawrence.

Also identified from Mira River samples were individuals of putative US origin (3/22 samples). Recent evidence that Striped Bass move between the US Atlantic coast and the northeastern coast of Nova Scotia is scarce. In 1983 a single fish tagged in the Kouchibouguac River in the Gulf of St. Lawrence in 1983 was later recaptured in the Wye River, Maryland (Hogans and Melvin 1984), indicating that this fish likely passed through the northeastern coast of Nova Scotia. In contrast, none of the hundreds of Striped Bass with internal acoustic tags in the Roanoke River, Hudson River, New England coast, Bay of Fundy, and Miramichi River (Douglas et al. 2003, Pautzke et al. 2010, Broome 2014, Callihan et al. 2015, Gahagan et al. 2015, Andrews et al. 2018) have ever been detected passing the Halifax Line of acoustic receivers on the eastern coast of Nova Scotia (Andrews et al. 2019a). Thousands of Striped Bass in the Gulf of St. Lawrence (Hogans and Melvin 1984, Douglas et al. 2003, DFO 2010), the Bay of Fundy (Rulifson and Dadswell 1995, Broome 2014), and along the US coast (Waldman et al. 1990, Richards and Rago 1999, Pautzke et al. 2010) have been externally tagged from the 1960s to the present day (Andrews et al. 2019a), only one of which has ever been caught on the far eastern shores of Nova Scotia (Douglas et al. 2003, Andrews et al. 2018). This apparent isolation may be caused by a physical isolation of the Gulf of St. Lawrence before the Canso Strait opened post-glacier retreat (Shaw and Courtney 2002) and after the Canso Causeway was built in 1955 (Vilks et al. 1975), or influenced by a sharp temperature change between the two water bodies (Rulifson and Dadswell 1995). A "genetic breakpoint" has been described in several other species along eastern Nova Scotia at ~45°N (close to the City of Halifax; Stanley et al. 2018). Increasing ocean temperatures are predicted to drive Striped Bass populations north, but this remains a poorly studied region.

The presence of a genetically distinct population of Striped Bass in the Saint John River following its suggested extirpation in the 1970s has been debated for over a decade (Andrews et al. 2017). Two previous studies have found evidence of unique genotypes distinct from US and Shubenacadie River Striped Bass, and present in adults (Bentzen and Paterson 2008) and juveniles (LeBlanc et al. 2018). A third study examined a mixture of 17 juveniles and 25 adults collected from the Saint John River in 2014 and found that all fish showed admixture between Shubenacadie River and US genotypes with no unique cluster (Wirgin et al., 2020). The 17 juveniles examined by Wirgin et al. (2020) are included in this present study, along with 15 additional juveniles collected in 2015–2017. In contrast to Wirgin et al. 2020’s results, most juveniles we examined showed a distinct genetic signature and admixture was only seen between the Saint John River cluster and either Shubenacadie or US genotypes. We detected no Shubenacadie-US hybrids. Adult and juvenile Saint John River Striped Bass examined in both studies are also part of an ongoing (6+ year) acoustic telemetry study. Initial telemetry results clearly demonstrate differing migratory patterns between adults genotyped as Shubenacadie origin (which left the river to spawn), adults genotyped as belonging to the Saint John River cluster (which migrated upstream) and those of US origin Striped Bass (Andrews et al. 2019b).

Striped Bass from US populations have been found in Minas Basin (Bay of Fundy; Rulifson et al. 2008) and are thought to enter the Shubenacadie River as well, this study found no evidence that migrants successfully spawn in the Shubenacadie River. In contrast, juveniles from the Saint John River were admixed with Shubenacadie River and Chesapeake Bay populations. It is unknown how often this gene flow occurs now or prior to the population's apparent disappearance in the 1970s. All admixed juveniles we examined had approximately equal assignment to the Saint John River and the Shubenacadie River/Chesapeake Bay clusters, suggesting intraspecific F1 hybrids. The first-generation hybrids and the distinctness of the Saint John River–US genotypes suggests that these admixed juveniles may be a recent development. There is little information about the proportion of US migrants in the Saint John River before the population crash and no information about possible admixed individuals (Andrews et al. 2017). Larger Striped Bass are more likely to migrate and to travel far (Callihan et al. 2014, DFO 2018, Andrews et al. 2019a), and as Striped Bass populations recover there is an increase in the number of older, larger individuals making migrations (Callihan et al. 2014). We hypothesize that the admixed juveniles result from small numbers of local spawners making admixed offspring more prevalent, increased migration from recovering populations, and a climate-induced northward range shift.

United States Regional Structure

In contrast to Canadian Striped Bass, F_{ST} values among US locations were much lower (F_{ST} = 0.000–0.035), and support three regions with low but significant genetic divergence: 1) Hudson River and Kennebec River, 2) Delaware River and Chesapeake Bay, and 3) North Carolina Rivers. Overall, our results suggest individual spawning populations within the Delaware River and Chesapeake Bay make up a large metapopulation connected by extensive gene flow, with lesser amounts of gene flow between this region and populations to the north and south.

Connectivity among populations of Striped Bass along the Atlantic Coast has been investigated in several previous studies (Brown et al. 2005, Bentzen and Paterson 2008, Able et al. 2012, Gauthier et al. 2013, Callihan et al. 2015), and is influenced by gene flow, stocking, and possibly recolonization following local extirpation. Striped Bass between Maine and North Carolina are highly migratory (>1000 km; Callihan et al. 2015), compared to the more restricted migratory range of populations in Canada and the apparent complete residency of populations south of Roanoke River, North Carolina. Several tagging studies have previously documented movement of Striped Bass among Chesapeake Bay, Kennebec River, Hudson River, and Roanoke River (Waldman et al. 1990, Dorazio et al. 1994, Kneebone et al. 2014, Callihan et al. 2015, Gahagan et al. 2015). The presence of an isolation-by-distance pattern of differentiation among US locations but not among Canadian locations further supports gene flow among populations in this range.

Our study also investigates the current genetic profile of the recently restored Kennebec River population of Striped Bass. The Kennebec River is one of several rivers in Maine that likely once hosted a native population of Striped Bass (Little 1995). This population declined and was likely extirpated by the late 1930s, and was subsequently stocked with over 260,000 Striped Bass juveniles from 1982 to 1991 from the Hudson River in an attempt to restore the population (G. Whippelhauser, pers. comm.). Considering stocking, it is unsurprising that the juvenile Striped Bass caught in the Kennebec River in this study were most similar to the Hudson River. F_{ST} values between the two rivers were low (F_{ST} = 0.008) but significant, and similar to values seen among some rivers within the Chesapeake Bay, indicating a similar level of relatedness despite the large geographic distance between them (approximately 620 km from mouth to mouth). This similarity was also seen in a recent microsatellite study that examined Kennebec River juveniles (Wirgin et al. 2020), which found no statistically significant difference between the Kennebec River and Hudson River.

Within the Chesapeake Bay and Delaware River, F_{ST} values were very low. The highest values were seen when comparing James River individuals to other rivers in the Bay, as well as Nanticoke and Choptank Rivers, both located on the east coast of the Bay, to rivers on the west coast. A small amount of differentiation between the east and west coast of Chesapeake Bay was also seen in the most recent microsatellite study done on Striped Bass (Wirgin et al. 2020) and may be due to the channel of deeper water that runs through the center of the Bay. The lowest F_{ST} values within the Chesapeake-Delaware region were seen between individuals from the head of Chesapeake Bay and Delaware River, supporting the hypothesis that these two groups of Striped Bass are not genetically distinct from one another. Previous genetic studies have found conflicting results on whether the growing Striped Bass population in the Delaware was distinct from the Chesapeake Bay. An analysis of mitochondrial length frequency differences of the recovered Delaware River Striped Bass found significant differences in minor length frequency alleles from Chesapeake Bay Striped Bass (Waldman and Wirgin 1994). Minor length frequency differences were also seen among tributaries within the Chesapeake Bay (Wirgin et al. 1993a), and microsatellite studies which found significant F_{ST} values between the Delaware River and the Chesapeake Bay also found F_{ST} values of the same magnitude among tributaries within the bay (Gauthier et al. 2013). Decades of observations of adult Striped Bass using the Chesapeake and Delaware Canal to transit between the Chesapeake and Delaware estuaries during spawning season (Nichols and Miller 1967, Koo and Wilson 1972, Kneebone et al. 2014) support the likelihood that the Delaware River population receives a high amount of gene flow from Chesapeake Bay Striped Bass, on a similar magnitude as seen among rivers within the Bay. Whether the current Delaware River Striped Bass are descended from a remnant population that was genetically similar to the Chesapeake Bay or whether they are descended from Chesapeake Bay Striped Bass that recolonized the river, it seems clear that Delaware River Striped Bass are part of a complex network of gene flow among the tributaries of the Chesapeake Bay.

Assignment

Self-assignment tests were performed on the SNP panel generated in this study to assess its utility as a reference dataset for future mixed-stock analyses. Previous attempts to use genetic markers for mixed-stock analysis have met with limited success. Most recently, a study conducted self-assignment tests using GeneClass2 on 14 microsatellites and was able to assign 60% of Striped Bass from the Hudson River, Chesapeake Bay (including the Delaware River), North Carolina, and South Carolina to a region of origin (Gauthier et al. 2013). Our SNP panel showed the highest assignment success when overlapping populations were grouped into the same reporting groups used by Gauthier et al. (2013). We were able to assign 99% of Striped Bass to a region of origin with >80% confidence. When individuals were assigned to river of origin (rather than region of origin) assignment success was much lower and individuals were mis-assigned to other rivers within the same region, reflecting the low genetic differentiation among these rivers. The assignment success rate seen within regions is likely an indication that rivers within a region are not demographically independent.

Statistical biases when using large panels of SNP loci have been identified in assignment tests that use simulated individuals to predict assignment accuracy of a set of loci (Anderson et al. 2008) and the uniformly high confidence values seen in this present study appear to corroborate this. In light of emerging techniques allowing high-throughput genotyping of large numbers (>1000) of loci (Ali et al. 2016), researchers looking to assess stock composition of increasingly closely related populations should interpret confidence scores with these issues in mind when choosing a geographic resolution in which to assign fish. In addition, admixed individuals seen in the Saint John River were assigned to one of their parent populations with high confidence, suggesting that assignment in both GeneClass2 and *rubias* is insensitive to the presence of admixed individuals. When performing mixed-stock analysis on locations with large numbers of hybrid individuals, assignment may be better conducted using a genetic clustering algorithm such as those found in LEA or STRUCTURE. Overall, our SNP panel constitutes a significant improvement over other genetic markers in assigning Striped

Bass to regional areas along the Atlantic coast and will be invaluable to the development of a highly accurate and reliable genetic tool for mixed-stock analysis of the species across the central and northern portion of their range.

Conclusion

Striped Bass have been thought to exhibit a high degree of natal homing (Pess et al. 2014); but recent genetic and telemetry studies indicate the species expresses more variability in homing to their natal river (e.g. Callihan et al. 2015; Gahagan et al. 2015). Studies document skipped spawning and straying among populations (Kneebone et al. 2014, Gahagan et al. 2015). Low or nonexistent genetic structure among tributaries in the Chesapeake Bay and the connected Delaware River (see also Brown et al. 2005; Gauthier et al. 2013) suggests that straying or colonization among rivers in this region is common. Canadian populations at the northern range limit exhibited greater genetic isolation, but with evidence of hybridization with US individuals in the Saint John River and detection of US individuals in the Mira River. Genetic structure in the north may relate to the relatively recent opening of the rivers, i.e., post-glaciation about 10,00 years ago and/or more recent climatic changes and population increases pushing US migrants farther north. Regardless, the variable exchanges among rivers provides a zoogeographic dynamic with important implications for the local and international management of Striped Bass.

This study represents the first contribution that used genotyping-by-sequencing to facilitate highly accurate mixed-stock analyses of Striped Bass along the Atlantic Coast, including stock compositions in the Bay of Fundy and ongoing characterization of Striped Bass along the Nova Scotian coast. This mixed-stock analysis method will be especially valuable if climate change influences shifts in the range of Striped Bass and results in increased mixing of different spawning populations across international borders, allowing for early detection and appropriate responses in management and policy.

Acknowledgements

We would like to acknowledge Paul Bentzen, who provided tissue samples from Miramichi River Striped Bass in Bras d’Or Lake. We would like to thank Dr. Michael Dadswell for spearheading an initial project to investigate Striped Bass in Cape Breton. We would also like to thank Scott Young and his late father Keith Young for their generous donation of time and knowledge in assisting us to capture and sample Striped Bass in the Saint John River, New Brunswick.

We would like to thank the Striped Bass Research Team for coordinating sampling from the Shubenacadie River, as well as the local anglers who provided many of the samples from this river. We thank Beth Versak (Maryland Department of Natural Resources) for coordinating sampling from the Chesapeake Bay, as well as Andy Kahnle (NYSDEC Bureau of Marine Resources, Hudson River Fisheries Unit) and Matt Fisher (Delaware

DNREC Division of Fish and Wildlife) for coordinating sampling from the Hudson River. For coordinating the collection of Kennebec River juveniles, we would like to thank Gail Wippelhauser and Jason Bartlett, from the Maine Department of Marine Resources. We would like to thank Jeremy McCargo, as well as the North Carolina Division of Marine Fisheries and Jeff Evans and Kevin Dockendorf (North Carolina Wildlife Resources Commission) for collection of Roanoke River and Cape Fear River samples.

Striped Bass is a priority species of the United States Department of Agriculture (USDA) National Research Support Project 8 (NRSP-8, *National Animal Genome Project*; USDA 0018889) and B.J.R. is the Co-Coordinator of the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry*.

Génome Québec Innovation Centre provided sequencing services, and funding was provided by NSERC, the Canadian Rivers Institute, Canada Research Chair, New Brunswick Innovation Foundation, the Canada Foundation for Innovation, the Massachusetts Division of Marine Fisheries, the North Carolina Division of Marine Fisheries, and the Foundation for Food and Agriculture Research (New Innovator Award).

Data Availability Statement

Raw sequencing data are available on the US National Center for Biotechnology Information, BioProjects [PRJNA627492]. Data input files are available in the Dryad repository, https://doi.org/10.5061/dryad.9kd51c5dd.

Literature Cited

Able, K. W., T. M. Grothues, J. T. Turnure, D. M. Byrne, and P. Clerkin. 2012. Distribution, movements, and habitat use of small Striped Bass (*Morone saxatilis*) across multiple spatial scales. Fishery Bulletin 110:176–192.

Ali, O. A., S. M. O’Rourke, S. J. Amish, M. H. Meek, G. Luikart, C. Jeffres, and M. R. Miller. 2016. RAD capture (Rapture): Flexible and efficient sequence-based genotyping. Genetics 202:389–400.

Allendorf, F. W., and L. W. Seeb. 2000. Concordance of genetic divergence among Sockeye Salmon populations at allozyme, nuclear DNA, and mitochondrial DNA markers. Evolution 54:640–651.

Allendorf, F. W., P. A. Hohenlohe, and G. Luikart. 2010. Genomics and the future of conservation genetics. Nature reviews. Genetics 11:697–709.

Anderson, E. C., R. Waples, and S. T. Kalinowski. 2008. An improved method for predicting the accuracy of genetic stock identification. Canadian Journal of Fisheries and Aquatic Sciences 65:1475–1486.

Andrews, S. N., T. Linnansaari, R. A. Curry, and M. J. Dadswell. 2017. The misunderstood Striped Bass of the Saint John River, New Brunswick: Past, present, and future. North American Journal of Fisheries Management 37:235–254.

Andrews, S. N., B. Wallace, M. Gautreau, T. Linnansaari, and R. A. Curry. 2018. Seasonal movements of Striped Bass *Morone saxatilis* in a large tidal and hydropower regulated river. Environmental Biology of Fishes 101:1549–1558.

Andrews, S. N., M. J. Dadswell, C. F. Buhariwalla, T. Linnansaari, and R. A. Curry. 2019a. The misunderstood Striped Bass of the Saint John River, New Brunswick: past, present, and future. Northeastern Naturalist 26:1–30.

Andrews, S. N., T. Linnansaari, N. Leblanc, S. Pavey, and R. A. Curry. 2019b. Interannual variation in spawning success of Striped Bass (*Morone saxatilis*) in the Saint John River, New Brunswick. River Research and Applications 36:13–24.

Archer, F. I., P. E. Adams, and B. B. Schneiders. 2017. stratag : An R package for manipulating, summarizing and analysing population genetic data. Molecular Ecology Resources 17:5–11.

ASMFC. 2019. Summary of the 2019 Benchmark Stock Assessment for Atlantic Striped Bass. ASMFC, Washington, D.C.

Baetscher, D. S., D. J. Hasselman, K. Reid, E. P. Palkovacs, and J. C. Garza. 2017. Discovery and characterization of single nucleotide polymorphisms in two anadromous alosine fishes of conservation concern. Ecology and Evolution 7:6638–6648.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57:289–300.

Bentzen, P., and I. G. Paterson. 2008. Report: genetic analysis of Striped Bass collected by Kingsclear First Nation in the Saint John River, New Brunswick. Report to the Department of Fisheries and Oceans, Dartmouth, Nova Scotia. p. 1-22.

Bentzen, P., M. Mcbride, and I. G. Paterson. 2014. Report: genetic analysis of Striped Bass collected in Bras d'Or Lake. Report to the Eskasoni Fish and Wildlife Commission, Eskasoni, Nova Scotia. p. 1-15.

Bernatchez, L., and C. C. Wilson. 1998. Comparative phylogeography of Nearctic and Palearctic fishes. Molecular Ecology 7:431–452.

Bjorgo, K. A., J. J. Isely, and C. S. Thomason. 2000. Seasonal movement and habitat use by Striped Bass in the Combahee River, South Carolina. Transactions of the American Fisheries Society 129:1281–1287.

Bond, M. H., P. A. Crane, W. A. Larson, and T. P. Quinn. 2014. Is isolation by adaptation driving genetic divergence among proximate Dolly Varden char populations? Ecology and Evolution 4:2515–2532.

Bourret, V., M. P. Kent, C. R. Primmer, A. Vasemägi, S. Karlsson, K. Hindar, P. McGinnity, E. Verspoor, L. Bernatchez, and S. Lien. 2013. SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). Molecular Ecology 22:532–551.

Bradbury, I. R., L. C. Hamilton, M. J. Robertson, C. E. Bourgeois, A. Mansour, and J. B. Dempson. 2014. Landscape structure and climatic variation determine Atlantic Salmon genetic connectivity in the Northwest Atlantic. Canadian Journal of Fisheries and Aquatic Sciences 71:246–258.

Broome, J. E. 2014. Population characteristics of Striped Bass (*Morone saxatilis*

Walbaum, 1792) in Minas Basin and patterns of acoustically detected movements within Minas Passage (Master's thesis). Acadia University, Wolfville, NS, Canada.

Brown, K. M., G. A. Baltazar, and M. B. Hamilton. 2005. Reconciling nuclear microsatellite and mitochondrial marker estimates of population structure: breeding population structure of Chesapeake Bay Striped Bass (*Morone saxatilis*). Heredity 94:606–15.

Buhariwalla, C. 2018. Documenting aspects of the ecology of Striped Bass *Morone saxatilis* (Walbaum, 1792) in northeastern Nova Scotia (Master's thesis). Acadia University, Wolfville, NS, Canada.

Callihan, J. L., C. H. Godwin, and J. A. Buckel. 2014. Effect of demography on spatial distribution: Movement patterns of the Albemarle Sound-Roanoke River stock of Striped Bass (*Morone saxatilis*) in relation to their recovery. Fishery Bulletin 112:131–143.

Callihan, J. L., J. E. Harris, and J. E. Hightower. 2015. Coastal migration and homing of Roanoke River Striped Bass. Marine and Coastal Fisheries 7:301–315.

Carmichael, J. T., S. L. Haeseker, and J. E. Hightower. 1998. Spawning migration of telemetered Striped Bass in the Roanoke River, North Carolina. Transactions of the American Fisheries Society 127:286–297.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124–3140.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:7.

Clark, J. 1968. Seasonal movements of Striped Bass contingents of Long Island Sound and the New York Bight. Transactions of the American Fisheries Society 97:320–343.

Clemento, A. J., A. Abadía-Cardoso, H. A. Starks, and J. C. Garza. 2011. Discovery and characterization of single nucleotide polymorphisms in Chinook Salmon, *Oncorhynchus tshawytscha*. Molecular Ecology Resources 11:50–66.

Conroy, C., P. Piccoli, and D. Secor. 2015. Carryover effects of early growth and river flow on partial migration in Striped Bass *Morone saxatilis*. Marine Ecology Progress Series 541:179–194.

Curry, R. A. 2007. Late glacial impacts on dispersal and colonization of Atlantic Canada and Maine by freshwater fishes. Quaternary Research 67:225–233.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

DFO. 2010. Assessment of the habitat quality and habitat use by the Striped Bass population of the St. Lawrence Estuary, Quebec. DFO Canadian Science Advisory Secretariat. Science Advisory Report. 2010/069.

DFO. 2014. Recovery potential assessment for the Bay of Fundy Striped Bass (*Morone saxatilis*) designatable unit. DFO Canadian Science Advisory Secretariat Science Advisory Report 2014/053.

DFO. 2018. Spawner abundance and biological characteristics of Striped Bass (*Morone saxatilis*) in the southern Gulf of St. Lawrence in 2017. DFO Canadian Science Advisory Secretariat Science Response 2018/016

Dorazio, R. M., K. A. Hattala, C. B. McCollough, and J. E. Skjeveland. 1994. Tag recovery estimates of migration of Striped Bass from spawning areas of the Chesapeake Bay. Transactions of the American Fisheries Society 123:950–963.

Douglas, S. G., R. G. Bradford, and G. Chaput. 2003. Assessment of Striped Bass (*Morone saxatilis*) in the Maritime provinces in the context of species at risk. Fisheries and Oceans Canada, Center for Science Advice (CSA), Gulf Region, Dartmouth, NS, Canada.

Excoffier, L., and N. Ray. 2008. Surfing during population expansions promotes genetic revolutions and structuration. Trends in Ecology and Evolution 23:347–351.

Excoffier, L., and H. E. L. Lischer. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular ecology resources 10:564–567.

Fabrizio, M. C. 1987. Growth-invariant discrimination and classification of Striped Bass stocks by morphometric and electrophoretic methods. Transactions of the American Fisheries Society 116:728–736.

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

Frichot, E., F. Mathieu, T. Trouillon, G. Bouchard, and O. François. 2014. Fast and efficient estimation of individual ancestry coefficients. Genetics 196:973–83.

Frichot, E., and O. François. 2015. LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution 6:925–929.

Gahagan, B., D. Fox, and D. Secor. 2015. Partial migration of Striped Bass: Revisiting the contingent hypothesis. Marine Ecology Progress Series 525:185–197.

Gauthier, D. T., C. A. Audemard, J. E. L. Carlsson, T. L. Darden, M. R. Denson, K. S. Reece, and J. Carlsson. 2013. Genetic population structure of US Atlantic Coastal Striped Bass (*Morone saxatilis*). Journal of Heredity 104:510–520.

Grothues, T. M., K. W. Able, J. Carter, and T. W. Arienti. 2009. Migration patterns of Striped Bass through nonnatal estuaries of the U.S. Atlantic Coast. American Fisheries Society Symposium:135–150.

Hasselman, D. J., D. Ricard, and P. Bentzen. 2013. Genetic diversity and differentiation in a wide ranging anadromous fish, American shad (*Alosa sapidissima*), is correlated with latitude. Molecular Ecology 22:1558–1573.

Hewitt, G. M. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. Biological Journal of the Linnean Society 58:247–276.

Hogans, W., and G. Melvin. 1984. Kouchibouguac National Park Striped Bass (*Morone saxatilis*) fishery survey, New Brunswick. Aquatic Industries Limited. St Andrew's.

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

Jombart, T., S. Devillard, and F. Balloux. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 11:94.

Jones, O. R., and J. Wang. 2010. COLONY: a program for parentage and sibship inference from multilocus genotype data. Molecular Ecology Resources 10:551–555.

Keyser, F. M., J. E. Broome, R. G. Bradford, B. Sanderson, and A. M. Redden. 2016.

Winter presence and temperature-related diel vertical migration of Striped Bass (*Morone saxatilis*) in an extreme high-flow passage in the inner Bay of Fundy. Canadian Journal of Fisheries and Aquatic Sciences 73:1777–1786.

King, T. L., S. T. Kalinowski, W. B. Schill, A. P. Spidle, and B. A. Lubinski. 2001. Population structure of Atlantic Salmon (*Salmo salar l.*): A range-wide perspective from microsatellite DNA variation. Molecular Ecology 10:807–821.

Kneebone, J., W. S. Hoffman, M. J. Dean, D. A. Fox, and M. P. Armstrong. 2014. Movement patterns and stock composition of adult Striped Bass tagged in Massachusetts coastal waters. Transactions of the American Fisheries Society 143:1115–1129.

Koo, T. S. Y., and J. S. Wilson. 1972. Sonic tracking Striped Bass in the Chesapeake and Delaware Canal. Transactions of the American Fisheries Society 101:453–462.

Laughlin, T. F., and B. J. Turner. 1996. Hypervariable DNA markers reveal high genetic variability within Striped Bass populations of the lower Chesapeake Bay. Transactions of the American Fisheries Society 125:49–55.

LeBlanc, N. M., S. N. Andrews, T. S. Avery, G. N. Puncher, B. I. Gahagan, A. R. Whiteley, R. A. Curry, and S. A. Pavey. 2018. Evidence of a genetically distinct population of Striped Bass within the Saint John River, New Brunswick, Canada. North American Journal of Fisheries Management 38:1339–1349.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 26:589–95.

Lischer, H. E. L., and L. Excoffier. 2012. PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics 28:298–

299.

Little, M. J. 1995. A report on the historic spawning grounds of the Striped Bass, "*Morone saxatilis*." Maine Naturalist 3:107–113.

Lotterhos, K. E., and M. C. Whitlock. 2014. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. Molecular Ecology 23:2178–2192.

Luikart, G., P. R. England, D. Tallmon, S. Jordan, and P. Taberlet. 2003. The power and promise of population genomics: from genotyping to genome typing. Nature Reviews Genetics 4:981–994.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10.

Mather, M. E., J. T. Finn, S. M. Pautzke, D. Fox, T. Savoy, H. M. Brundage, L. A. Deegan, and R. M. Muth. 2010. Diversity in destinations, routes and timing of small adult and sub-adult Striped Bass *Morone saxatilis* on their southward autumn migration. Journal of Fish Biology 77:2326–2337.

McConnell, S. K. J., D. E. Ruzzante, P. T. O'Reilly, L. Hamilton, and J. M. Wright. 1997. Microsatellite loci reveal highly significant genetic differentiation among Atlantic salmon (*Salmo salar L.*) stocks from the east coast of Canada. Molecular Ecology 6:1075–1089.

Narum, S. R., C. A. Buerkle, J. W. Davey, M. R. Miller, and P. A. Hohenlohe. 2013. Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology 22:2841–2847.

Nichols, P. R., and R. V. Miller. 1967. Seasonal movements of Striped Bass, Roccus saxatilis (Walbaum), tagged and released in the Potomac River, Maryland, 1959-61. Chesapeake Science 8:102.

Pautzke, S. M., M. E. Mather, J. T. Finn, L. A. Deegan, and R. M. Muth. 2010. Seasonal use of a New England estuary by foraging contingents of migratory Striped Bass. Transactions of the American Fisheries Society 139:257–269.

Pavey, S. A., J. Gaudin, E. Normandeau, M. Dionne, M. Castonguay, C. Audet, and L. Bernatchez. 2015. RAD sequencing highlights polygenic discrimination of habitat ecotypes in the panmictic American Eel. Current Biology 25:1666–1671.

Pess, G. R., T. P. Quinn, S. R. Gephard, and R. Saunders. 2014. Re-colonization of Atlantic and Pacific rivers by anadromous fishes: Linkages between life history and the benefits of barrier removal. Reviews in Fish Biology and Fisheries 24:881–900.

Pielou, E. C. 1991. After the ice age: The return of life to glaciated North America.

Piry, S., A. Alapetite, J.-M. Cornuet, D. Paetkau, L. Baudouin, and A. Estoup. 2004. GENECLASS2: A software for genetic assignment and first-generation migrant detection. Journal of Heredity 95:536–539.

Poland, J. A., P. J. Brown, M. E. Sorrells, and J. L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253.

Rannala, B., and J. L. Mountain. 1997. Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences 94:9197–9201.

Richards, R. A., and P. J. Rago. 1999. A case history of effective fishery management: Chesapeake Bay Striped Bass. North American Journal of Fisheries Management 19:356–375.

Robinson, M., S. Courtenay, T. Benfey, L. Maceda, and I. Wirgin. 2004. Origin and Movements of Young-of-the-Year Striped Bass in the Southern Gulf of St. Lawrence, New Brunswick. Transactions of the American Fisheries Society 133(2):412–426.

Rockman, M. V. 2012. The QTN program and the alleles that matter for evolution: All that's gold does not glitter. Evolution 66:1–17.

Rulifson, R. A., and M. J. Dadswell. 1995. Life history and population characteristics of Striped Bass in Atlantic Canada. Transactions of the American Fisheries Society 124:477–507.

Rulifson, R. A., and K. A. Tull. 1999. Striped Bass spawning in a tidal bore river: The Shubenacadie Estuary, Atlantic Canada. Transactions of the American Fisheries Society 128:613–624.

Rulifson, R. A., S. A. McKenna, and M. J. Dadswell. 2008. Intertidal habitat use, population characteristics, movement, and exploitation of Striped Bass in the Inner Bay of Fundy, Canada. Transactions of the American Fisheries Society 137:23–32.

Ryman, N., S. Palm, C. André, G. R. Carvalho, T. G. Dahlgren, P. E. JORDE, L. Laikre, L. C. Larsson, A. Palmé, and D. E. Ruzzante. 2006. Power for detecting genetic divergence: differences between statistical methods and marker loci. Molecular Ecology 15:2031–2045.

Secor, D. . 1999. Specifying divergent migrations in the concept of stock: the contingent hypothesis. Fisheries Research 43:13–34.

Secor, D. H., J. R. Rooker, E. Zlokovitz, and V. S. Zdanowicz. 2001. Identification of riverine, estuarine, and coastal contingents of Hudson River Striped Bass based upon otolith elemental fingerprints. Marine Ecology Progress Series 211:245–253.

Setzler, E. M., W. R. Boynton, K. V Wood, H. H. Zion, L. Lubbers, N. K. Mountford, P. Frere, L. Tucker, and J. A. Mihursky. 1980. Synopsis of biological data on Striped Bass, *Morone saxatilis,* (Walbaum).

Shaw, J., and R. C. Courtney. 2002. Postglacial coastlines of Atlantic Canada. Page Geological Survey of Canada.

Skinner, M. E., A. V Uzilov, L. D. Stein, C. J. Mungall, and I. H. Holmes. 2009. JBrowse: a next-generation genome browser. Genome research 19:1630–8.

Stanley, R. R. E., C. DiBacco, B. Lowen, R. G. Beiko, N. W. Jeffery, M. Van Wyngaarden, P. Bentzen, D. Brickman, L. Benestan, L. Bernatchez, C. Johnson, P. V. R. Snelgrove, Z. Wang, B. F. Wringe, and I. R. Bradbury. 2018. A climate-associated multispecies cryptic cline in the northwest Atlantic. Science Advances 4:eaaq0929.

Vilks, G., C. T. Schafer, and D. A. Walker. 1975. The influence of a causeway on oceanography and foraminifera in the Strait of Canso, Nova Scotia. Canadian Journal of Earth Sciences 12:2086–2102.

De Villemereuil, P., É. Frichot, É. Bazin, O. François, and O. E. Gaggiotti. 2014. Genome scan methods against more complex models: When and how much should we trust them? Molecular Ecology 23:2006–2019.

Waldman, J. R., D. J. Dunning, Q. E. Ross, and M. T. Mattson. 1990. Range dynamics of Hudson River Striped Bass along the Atlantic Coast. Transactions of the American Fisheries Society 119:910–919.

Waldman, J. R., and M. C. Fabrizio. 1994. Problems of stock definition in estimating relative contributions of Atlantic Striped Bass to the coastal fishery. Transactions of the American Fisheries Society 123:766–778.

Waldman, J. R., and I. I. Wirgin. 1994. Origin of the present Delaware River Striped Bass population as shown by analysis of mitochondrial DNA. Transactions of the American Fisheries Society 123:15–21.

Waldman, J. R., L. Maceda, and I. Wirgin. 2012. Mixed-stock analysis of wintertime aggregations of Striped Bass along the Mid-Atlantic coast. Journal of Applied Ichthyology 28:1–6.

Whitlock, M. C., and K. E. Lotterhos. 2015. Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F(ST). The American naturalist 186 Suppl:S24-36.

Wirgin, I., L. Maceda, J. R. Waldman, and R. N. Crittenden. 1993a. Use of mitochondrial DNA polymorphisms to estimate the relative contributions of the Hudson River and Chesapeake Bay Striped Bass stocks to the mixed fishery on the Atlantic Coast. Transactions of the American Fisheries Society 122:669–684.

Wirgin, I., T.-L. Ong, L. Maceda, J. R. Waldman, D. Moore, and S. Courtenay. 1993b. Mitochondrial DNA variation in Striped Bass (*Morone saxatilis*) from Canadian rivers. Canadian Journal of Fisheries and Aquatic Sciences 50:80–87.

Wirgin, I., B. Jessop, S. Courtenay, M. Pedersen, S. Maceda, and J. R. Waldman. 1995. Mixed stock analysis of Striped Bass in rivers of the Bay of Fundy as revealed by mitochondrial DNA. Canadian Journal of Fisheries and Aquatic Sciences 52:961–970.

Wirgin, I., L. Maceda, J. Stabile, and C. Mesing. 1997a. An evaluation of introgression of

Atlantic coast Striped Bass mitochondrial DNA in a Gulf of Mexico population using formalin-preserved museum collections. Molecular Ecology 6:907–916.

Wirgin, I., J. R. Waldman, L. Maceda, J. Stabile, and V. J. Vecchio. 1997b. Mixed-stock analysis of Atlantic Coast Striped Bass (*Morone saxatilis*) using nuclear DNA and mitochondrial DNA markers. Canadian Journal of Fisheries and Aquatic Sciences 54:2814–2826.

Wirgin, I., D. Currie, N. Roy, L. Maceda, and J. R. Waldman. 2005. Introgression of nuclear DNA (nDNA) alleles of stocked Atlantic coast Striped Bass with the last remaining native gulf of Mexico population. North American Journal of Fisheries Management 25:464–474.

Wirgin, I., L. Maceda, M. Tozer, J. Stabile, and J. Waldman. 2020. Atlantic coastwide population structure of Striped Bass *Morone saxatilis* using microsatellite DNA analysis. Fisheries Research 226:105506.

Yeaman, S. 2015. Local adaptation by alleles of small effect. The American naturalist 186 Suppl:S74-89.

CHAPTER 3: Comparing mixed models and Random Forest association tests using naturalGWAS and a Striped Bass SNP dataset

1. Nathalie M. Leblanc (corresponding author)

Department of Biological Sciences, Canadian Rivers Institute, University of New Brunswick, Saint John, NB, E2L 4L5, Canada. Email: nleblan5@unb.ca

Email: nathalie.leblanc@unb.ca

2. Scott A. Pavey

Department of Biological Sciences, Canadian Rivers Institute, University of New Brunswick, Saint John, NB, E2L 4L5, Canada. Email: spavey@unb.ca

Abstract

In this study, we used naturalGWAS to test the performance of Zhao's Random Forest method in comparison to an uncorrected Random Forest test and two mixed models. We created 100 sets of phenotypes, corresponding to 4 effect sizes and 2, 5, 10, 20, or 30 causal loci, simulated from 7319 empirical SNPs generated from Striped Bass sampled from three distinct locations with high amounts of genetic structure and intraspecific hybrids. All association methods were evaluated for their ability to detect genotype-phenotype associations based on power, false discovery rates, and raw number of false positives. We found that all association methods had low power to detect all causal loci both when phenotypes were highly polygenic and when phenotypes were highly correlated with population structure. Zhao's Random Forest produced the fewest false positives across all tests and performed well when more liberal multiple testing corrections were used, allowing the method to achieve higher power than the next best performing method, CATE. We then used both CATE and Zhao's Random Forest to test for associations between Striped Bass genotype and condition factor, and found no significant loci.

Introduction

The advent of next generation sequencing techniques has given researchers unprecedented opportunity to examine areas of adaptation within even non-model species' genomes, as well as drawing associations between genomic markers and phenotypes (Ekblom and Galindo 2011). The ease of generating large amounts of genetic data has resulted in a shift toward both thorough genetic characterization of ecological model species, and investigations into non-model species that go beyond population-level genetics (Ekblom and Galindo 2011). Genome-wide association studies (GWAS) allow for the detection of causal loci without the use of controlled crosses (Korte and Ashley 2013), and GWAS have successfully used reduced representation SNP libraries to detect the genetic basis of a wide variety of phenotypes, such as migration timing (Brieuc et al. 2015), disease resistance (Barría et al. 2018), or growth (Moghadam et al. 2007).

An ongoing challenge when drawing associations between genotype and phenotypic traits is in detecting causal loci when phenotypes are not influenced by a small number of loci of large effect. There is a growing recognition that most phenotypes are likely influenced by large numbers of loci that each contribute to the phenotype in small ways and in combination with other loci (Rockman 2012; Visscher et al. 2012). Typical tests for genomic association use mixed models, which control for background genetic variation at the same time they assess each locus's association with a trait (Vilhjálmsson and Nordborg 2013, Yang et al. 2014). Since they were first introduced as a method of association testing, numerous modifications have been proposed to improve the mixed model's computational efficiency and ability to correct for neutral population structure while still retaining power to detect the genetic basis of phenotypic traits correlated with that population structure (Frichot et al. 2013, Zhou and Stephens 2013). However, statistical tests such as mixed models examine loci individually for associations with a phenotype of interest, and therefore have low power for detecting loci that contribute very small effects in combination with other loci (Brieuc et al. 2018).

In response to these challenges, machine learning algorithms such as Random Forest have been adopted for their ability to analyze large numbers of loci simultaneously, incorporating locus-locus interactions to identify the best predictors of a trait (Brieuc et al. 2018). Random Forest analyses have been effective at discriminating among distinct populations (Sylvester et al. 2017), predicting phenotypes of individuals based on a given genotype (Bureau et al. 2005, Díaz-Uriarte and Alvarez de Andrés 2006), and identifying potentially adaptive genes (Brieuc et al. 2015, Healy et al. 2018). Studies examining the efficacy of this method are slowly accumulating, with a variety of proposed improvements. A 2003 study found that Random Forest can identify candidate genes associated with a phenotype some, but not all of the time (Bureau et al. 2003). In 2012, Zhao et al. introduced a method of correcting for population structure by transforming phenotype and genotype data with regression-based residuals (Zhao et al. 2012), and this remains one of the most accessible correction methods for Random Forest (Brieuc et al. 2018). This method of correction was later evaluated as part of a general assessment of the utility of Random Forest for ecological association studies, which found that the method reduces population structure-associated false positives but may not reduce overall false positives (Brieuc et al. 2018). In both cases, Zhao’s method of correction for Random Forest tests were evaluated using small numbers of causal loci (1 and 5). As

Random Forest analyses are thought to be promising methods for dealing with polygenic traits (Brieuc et al. 2018), an examination of the performance of corrected and uncorrected Random Forest on identifying causal loci of polygenic traits is essential.

Genotype association methods have been evaluated using simulated datasets with defined demographic histories, but empirical studies often violate the assumptions present in demographic models and the performance of tests in less than ideal datasets is often not measured (François and Caye 2018). A recent R package, naturalGWAS, enables researchers to simulate quantitative phenotypic data from empirical genetic datasets in order to test which GWAS methods have the best power to correctly discriminate loci associated with those traits (François and Caye 2018). Performing simulations on individual datasets ensures that researchers do not base their choice on results reported by studies that have different demographic frameworks (Santure and Garant 2018).

We used naturalGWAS to test the power of Zhao's Random Forest method in comparison to an uncorrected Random Forest test and two mixed models to detect association patterns in a dataset characteristic of many small-scale studies, prior to choosing an association method for empirical condition factor measurements in the same dataset. Individuals were sampled from four locations that comprise three genetically distinct groups (F_{ST} 0.1–0.16; LeBlanc et al. 2020) and admixed individuals present in one population, and individuals were genotyped at a relatively low marker density (7319 SNPs). Using these empirical genotypes, we tested the ability of these techniques to identify the causal loci of 100 sets of simulated phenotypes that varied in effect size and number of causal variants.

Methods

Sample Collection

Tissue was collected from Striped Bass individuals from four locations: Saint John River, Shubenacadie River, Roanoke River, and Cape Fear River. Individuals from Roanoke River and Cape Fear River are extremely genetically similar (LeBlanc et al. 2020) and were considered one population in this study. Fin clips from Saint John River Striped Bass (n=82) aged 0 to 4 were collected July–Sept 2014–2017 via gill net, fyke net, Gaspereau trap, boat electrofishing, and angling. These included individuals used in chapters 1 and 2, as well as additional individuals not included in previous chapters to maintain consistency in sample sizes among locations. Ages for Saint John River juveniles were obtained from scales. Scales were collected from adult Striped Bass in Shubenacadie River (n=33) caught by angling on May 2014–2017 when native Shubenacadie bass are known to migrate from overwintering sites and migrants have not yet arrived in the river (DFO 2014). Fin clips were collected from adults in spawning condition in Roanoke and Cape Fear River (n=56) in April and May 2015 by boat electrofishing, during the North Carolina Division of Marine Fisheries and North Carolina Wildlife Resources Commission annual stock assessment.

Laboratory

DNA was isolated using the NucleoMag® 96 Tissue (Machery-Nagel, Düren, Germany) kit and an epMotion 5075t (Cat. 5075000302). Libraries of 96 individuals each were prepared using a double-digest restriction-site associated DNA sequencing (ddRAD-seq or ddRAD) protocol modified from Poland et al. (2012) as described in previous chapters, using restriction enzymes *PstI* and *MspI* and a Sage Pippin Prep © platform for size selection before amplification. Libraries were sent to Génome Québec Innovation Centre and processed using paired-end sequencing on Illumina® HiSeq™ 2500 (San Diego, U.S.A.).

Data Filtering

Raw sequences were demultiplexed using the *process_radtags* module of stacks v. 2.2 (Catchen et al. 2013). Demultiplexed reads were aligned to the most recently published Striped Bass genome (NCBI BioProject PRJNA532441) using the barrows-weeler aligner (BWA mem; Li and Durbin 2010). Alignments were kept if they had an alignment score of 10 or more, were properly paired, and were not supplementary, secondary, or unmapped reads (flags: -f 3, -F 1796). Aligned samples were discarded if sorted alignment files were less than 50 Mb (approximately one tenth of the size of the largest alignment file (600 Mb). Using stacks 2, a catalogue of 1,184,785 SNPs was built and genotypes were imputed using the Maruki-Lynch Bayesian method, which is optimized to accurately genotype low coverage loci (Maruki and Lynch 2015, 2017). A maximum soft-clipping of 10% was allowed for paired sequences (2.7% sequences failed).

For each dataset, an initial missing data threshold of 50% was applied, resulting in 64,513 loci. Loci were then excluded if they had an Fis value < -0.3 or an observed heterozygosity > 0.6 in all populations, using statistics output by the populations module, leaving 22,461 loci. Using VCFTools (Danecek et al. 2011), we excluded loci with a minor allele frequency less than 1% and loci with more than two alleles, loci that had more than 30% missing data, and genotypes with a read depth less than 7, which produced 7,505 loci. We used the VCFTools command --missing-indv to examine missing data per individual after this filtering and excluded any samples with greater than 30% missing data. After removing 7 North Carolina samples, the remaining individuals had 0.4% to 16% missing data, with a mean missing data of 4%. We also used the VCFTools command --relatedness to exclude all but one of any full sibling groups, removing 1 individual. Finally, linkage disequilibrium was calculated on all samples using PLINK (Purcell and Chang n.d., Chang et al. 2015), and one locus from every pair with an r2 above 0.8 was randomly removed, resulting in a final SNP dataset of 7,319 loci.

Simulations

A number of methods have been created and used to measure associations between continuous phenotypes and genotypes, that attempt to correct for confounding factors such as population structure, isolation by distance, environmental effects on genotypes. We use *naturalGWAS* to measure the power of association tests to identify true causal variants in complex empirical datasets. We assessed the power and false discovery rates (FDR) of five GWAS methods, detailed below. This dataset contains individuals from three groups of Striped Bass, all of which have substantial neutral genetic differences with each other. Environmental effects on anadromous fish are complex due to the wide geographic range many individuals inhabit throughout their lives. While climate variables based on sampling location cannot replicate the true environmental exposure of individuals, they can help to simulate that fish from one spawning location have different histories compared to fish from another, particularly when locations are far apart and/or migration ranges do not overlap.

We created phenotypes drawn from 2, 5, 10, 20 or 30 causal variants (SNPs that contribute to the phenotype), at effect sizes of 10, 100, 1000, and 10000. Variants had additive effects on phenotype, and effect sizes are constant values applied equally to all causal loci in the model, producing distributions wherein loci with smaller allele frequencies have stronger observed effects. This matches expected distributions of most polygenic traits in humans, which are composed of a combination of a few rare variants of large effect and many common variants of lower effect (Timpson et al. 2018). Each combination of causal loci number and effect size was repeated five times for a total of 300 phenotypes generated. For all phenotypes, a gene-environment interaction of 0.5 (moderate environmental influence) was assumed and association signal was modified by 40 confounding factors. For each analysis, the number of true causal SNPs detected by a GWAS method was recorded, along with the number of false positives. A SNP was considered a false positive if it was located outside of a 35 kb window centred around a causal SNP. This was a liberal LD window based on distance to baseline linkage disequilibrium in Striped Bass (unpublished data).

We ran five different association tests on these simulated phenotypes, implemented in the following R packages. For all analyses, missing data was imputed using the *impute* function in the *LEA* v2.8.0 R package. This uses ancestry estimates of each sample (calculated using *snmf*) to impute missing genotypes for each sample. We used the R function *pca* to visualize a PCA scree-plot of our genetic data, and chose the number of factors as 3 for use in downstream analyses where required. The R package *lfmm v1.0* was used to conduct latent factor mixed model analysis (LFMM) that corrects for shared ancestry of samples using latent factors (Frichot et al. 2013), with 3 factors. Confounder adjusted testing and estimation (CATE) is a recent implementation of latent factor analysis that corrects for batch effects as well as unmeasured covariates associated with the measured phenotype, often found in high-throughout experiments (Wang et al. 2017). We implemented this analysis using the R package *cate v1.1* (Wang et al. 2017) and set the number of factors as 3. We compared both LFMM and CATE results to a latent factor analysis, Oracle, that corrects for population stratification by including the confounders used when creating simulated phenotypes as fixed effects in a monogenic phenotype model. Oracle is included in the *naturalGWAS* v0.1.0 R package. This test serves as a basepoint reference for the performance of confounder correction; all other tests that attempt to estimate confounders are expected to do so less well. Finally, we investigated the performance of two variants of the machine learning Random Forest analysis. The uncorrected Random Forest algorithm was implemented using the R package randomforest v4.6-14 (Liaw and Wiener 2015), using an ntree of 2000 and an mtry of 72. For all analyses, phenotypes were scaled to have a variance of 1. We also used a method of correcting for population structure proposed by Zhao et al. (2012) that is perhaps the most widely used method of correcting for population structure currently (Brieuc et al. 2018). Using this method, phenotype and genotype values were regressed against population or group membership, and the residuals were used as “corrected” phenotype and genotype values that Random Forest models were then built on.

For both Random Forest methods, we calculated empirical p-values for the Gini importance measure to facilitate direct comparisons with other GWAS methods. When permuting features in a Random Forest, a new model must be trained on the permuted data to reduce bias in the resulting significance values (Hooker and Mentch 2019). Simulated phenotypes were randomly permuted across samples, destroying the association between phenotype and causal loci, and a new Random Forest model was trained on the resulting dataset. This was repeated 1000 times per phenotype to create a null distribution of importance values, which were used to calculate p-values using the empPvals function implemented in the R package *qvalue* (Storey and Tibshirani 2003).

Several methods of correcting for multiple testing are common in genetic studies, and all involve a balance between reducing false positives and retaining power to detect true positives. Sequential Bonferroni Correction (Rice 1989) was introduced as a less conservative variant on the strict Bonferroni correction; however it has also been criticized for being overly conservative, and too likely to reject true positives. A more recent method is known as the false discovery rate, often implemented as a qvalue, which is increasingly popular for its less conservative thresholds (Benjamini and Hochberg 1995). We compared these two methods of correction in our simulation by calculating significance using both Storey's false discovery rate method implemented in the R package *qvalue* using a FDR threshold of 0.05 and 0.2, and also manually adjusting p-values using the sequential Bonferroni method.

For all tests, power was calculated by identifying the proportion of causal variants, or SNPs within 35k base pairs of a causal variant, successfully identified as significant after correction for multiple tests. False positives were defined as any significant loci that was located farther than 35k base pairs from a causal locus, and false discovery rates were calculated as the proportion of false positives divided by the total number of significant loci.

Simulated phenotypes varied in how correlated they were with population structure within the dataset (Appendix G). We investigated whether this correlation had an impact on test results using a Kruskal-Wallace test to compare phenotype value with membership in one of the three population groupings (Saint John River, Shubenacadie River, and North Carolina rivers). Significance values from this test were used as a proxy for the degree of correlation between phenotype and population structure, and these were compared to number of false positives using linear regression.

Striped Bass Condition Factor

We evaluated association between the same empirical genotype data and condition factor values calculated using mass and total length measurements of all individuals in the dataset. Based on the performance of association tests in simulations, we chose both Zhao's Random Forest and CATE for use on our empirical data. Based on each test's false positive rates at different thresholds of correction for multiple tests, we evaluated CATE using an FDR of 0.05 and Zhao's Random Forest using an FDR of 0.2.

Results

Power

We assessed power of all association tests to detect causal loci in general, as well as in scenarios with highly polygenic traits. LFMM had the highest mean power at all correction levels (41-59%), followed by CATE (40-51%), Oracle (38-48%), Zhao's Random Forest (35-48%) and uncorrected Random Forest (32-47%) and power decreased for all tests as number of causal loci increased (Figure 3.1). To assess the effect of highly polygenic phenotypes, we compared simulations with 2 and 5 causal loci, representing phenotypes with few causal loci, to simulations with 20 and 30 causal loci, which represent highly polygenic phenotypes. Mean power across all tests with 5 or fewer loci underlying the phenotype was 76% (CATE: 84%, LFMM: 82%, ORACLE: 83%, RF: 74%, Zhao's RF: 74%). With 20 or more underlying the phenotype mean power was 10% (CATE: 10%, LFMM: 18%, ORACLE: 7%, RF: 8%, Zhao's RF: 8%). All tests had the highest power when a false discovery rate of 0.2 was used to correct for multiple tests

(47-59%; Figure 3.1) and lowest when sequential Bonferroni correction was used (32-41%). All effect sizes produced a wide range of powers and increasing effect size did not have a noticeable effect on power (mean power 40-46%).

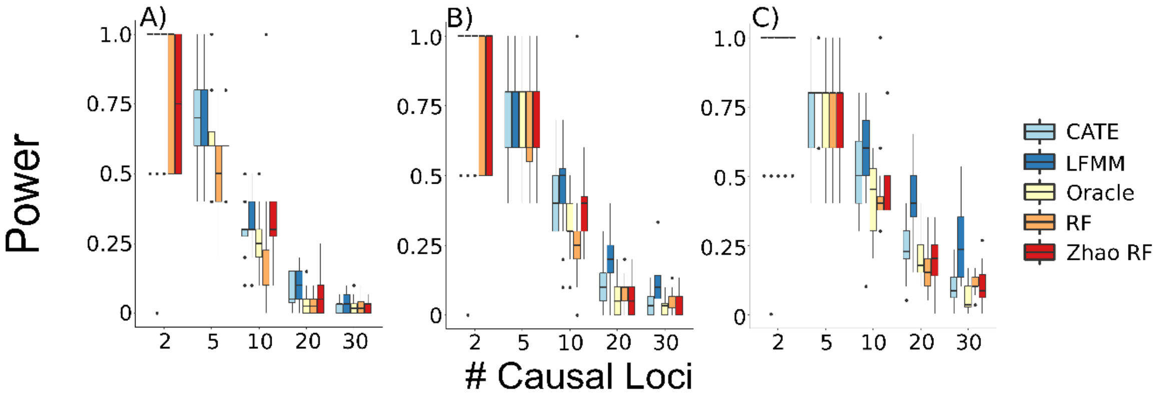


Figure 3.1. Boxplot of power achieved by 5 association tests run on phenotypes simulated from empirical genotypes where 2, 5, 10, 20, and 30 loci were used to determine phenotype value. Power is calculated from significance values corrected for multiple testing using A) Sequential Bonferroni correction, B) False Discovery Rate of 0.05, C) False Discovery Rate of 0.2.

False Positives and False Discovery Rate

False Discovery Rates increased with more liberal correction methods, but Zhao's Random Forest maintained relatively low false discovery rates even when more liberal correction methods were used (Figure 3.2). Using sequential Bonferroni correction, Zhao's Random Forest had an average false discovery rate of 8%, and at a false discovery rate threshold of 0.2 the same test had an average false discovery rate of 38%. By comparison, false discovery rates for LFMM when sequential Bonferroni correction was used were 41%, and for uncorrected Random Forest they were 51%. Similarly, CATE had the same false discovery rate at an FDR threshold of 0.05 as LFMM did when corrected with sequential Bonferroni.

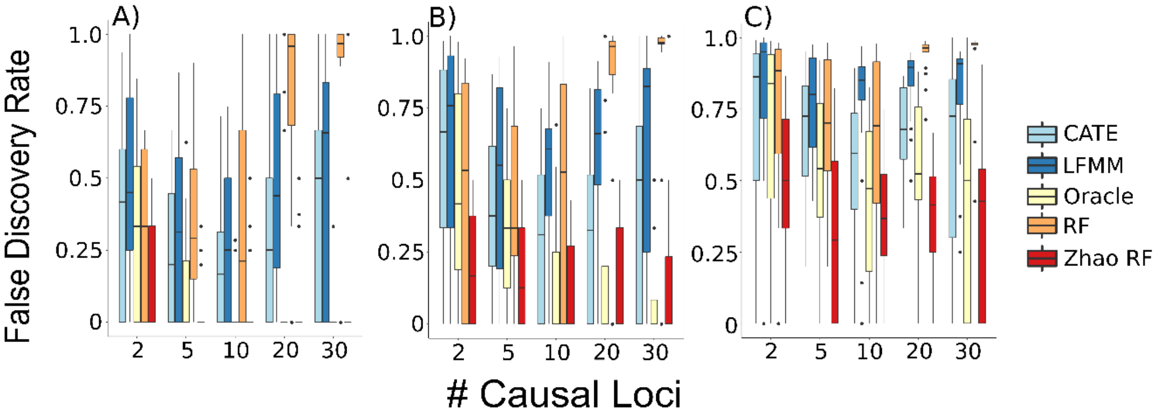


Figure 3.2. Boxplot of false discovery rates achieved by 5 association tests run on phenotypes simulated from empirical genotypes where 2, 5, 10, 20, and 30 loci were used to determine phenotype value. Power is calculated from significance values corrected for multiple testing using A) Sequential Bonferroni correction, B) False Discovery Rate of 0.05, C) False Discovery Rate of 0.2.

Because false discovery rates were relatively high across all tests, we also examined trends in number of false positives across different numbers of causal loci. Among mixed model tests, the lowest number of false positives was seen in the Oracle test (mean: 5.5), implying that the main source of false positives in mixed models was confounding factors within the genome, such as population stratification. Uncorrected Random Forest had false positive numbers comparable to other tests when there were 10 or fewer causal loci, and saw an abrupt increase in false positive numbers when there were 20 or more causal loci (Figure 3.3). LFMM had the next highest numbers of false positives across all levels of polygeny.

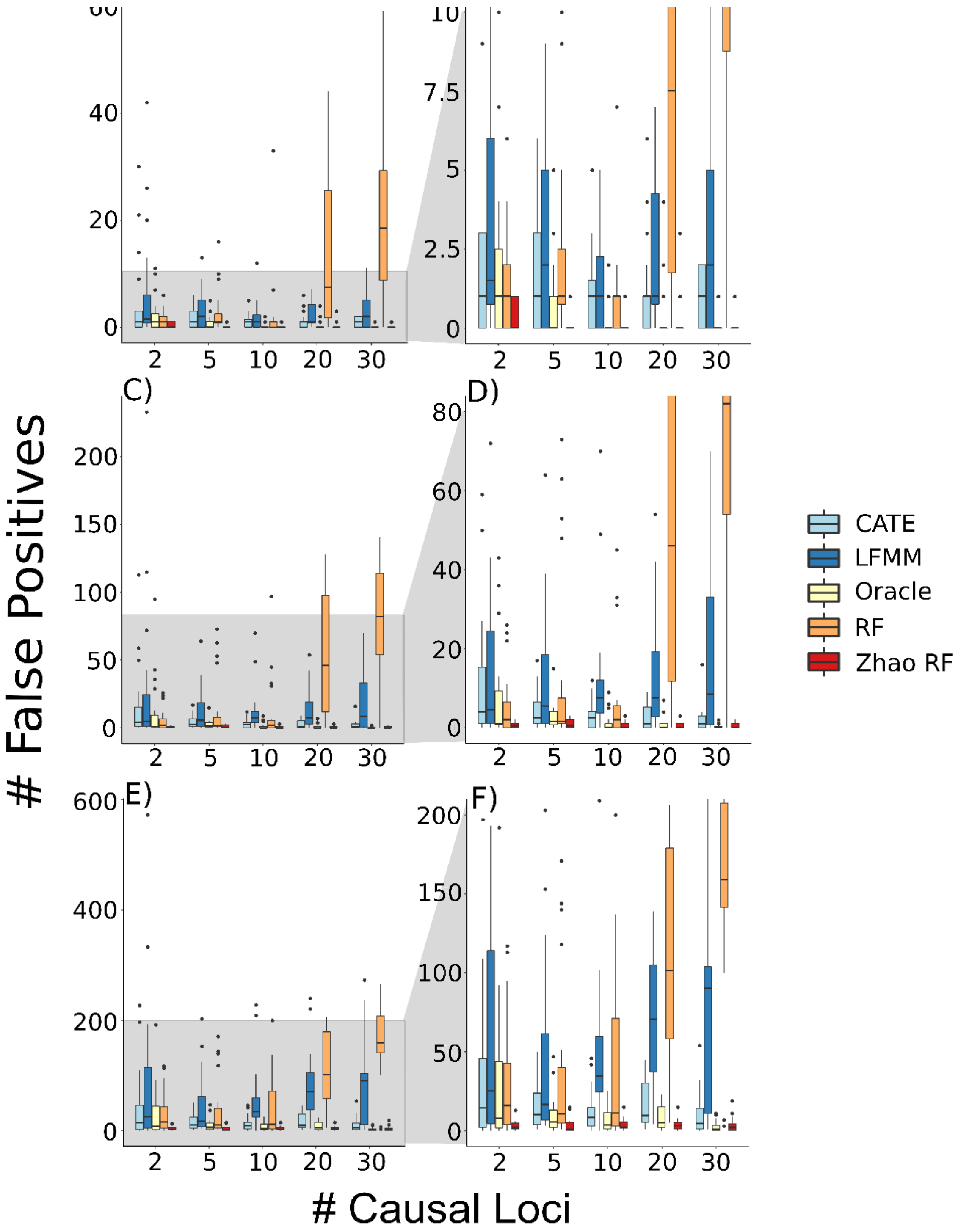


Figure 3.3. Boxplot of false positives identified by 5 association tests run on phenotypes simulated from empirical genotypes where 2, 5, 10, 20, and 30 loci were used to determine phenotype value. A) False positives when significance was corrected using sequential bonferroni, C) False positives when significance

was corrected using a false discovery rate of 0.05, E) False positives when significance was corrected using a false discovery rate of 0.2. Graphs B, D, and F depict the shaded areas of graphs A, C, and E.

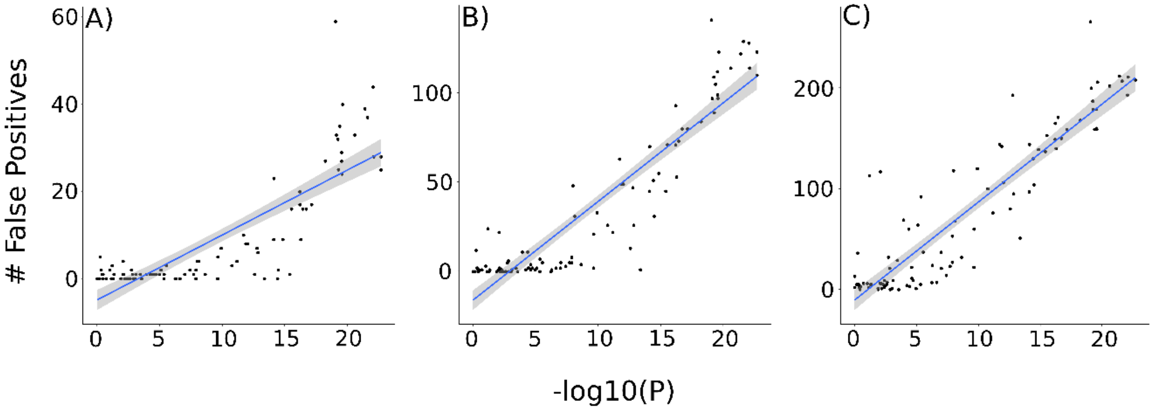


Figure 3.4. Number of false positives identified by Random Forest on 300 simulated phenotypes, plotted against the log of p-values obtained from a Kruskal-Wallace test measuring association between phenotype value and population structure. Correction for multiple testing was done using A) Sequential Bonferroni, B) False Discovery Rate with a threshold of 0.05, or C) False Discovery Rate with a threshold of 0.2.

Phenotype vs Population Structure

Our simulations showed high numbers of false positives in many Random Forest tests, but these numbers varied among simulations made with the same number of causal loci and effect sizes. We found that false positives in uncorrected Random Forest tests were positively associated with how closely phenotype distribution correlated with population structure (Figure 3.4; R=0.7, p-value = 2.2x10^{-16}), with most increases seen at KW p-values greater than 1x10^{-10}. This trend continued when we examined weakly polygenic and highly polygenic simulations separately, although correlation was weaker when phenotypes had low levels (R=0.4, p-value = 5.5x10^{-6}). By contrast, regression values for the three mixed model tests and for Zhao’s Random Forest ranged from -0.01 to 0.08 (p-values 0.003 to 0.87).

Striped Bass Condition Factor

Finally, we applied the best performing association tests in our simulations to our empirical condition factor data. Both Zhao’s Random Forest and CATE methods found no loci significantly associated with condition factor, even at a false discovery rate threshold of 0.2. Based on simulations, where Zhao’s Random Forest and CATE were able to detect at least one causal loci even in highly polygenic traits, we can conclude that condition factor is likely not strongly influenced by genetics.

Discussion

Overview

In this study, we simulated phenotypes using the observed genotypes of samples taken from three highly differentiated populations with admixed individuals, to test the ability of two mixed model and two Random Forest algorithms in accurately identifying the causal loci of simulated phenotypes. Background genetic structure was included in the simulations through principal components, and we made use of a mixed model method implemented in naturalGWAS called Oracle as a reference for a mixed model test that perfectly accounts for confounding effects included in the simulation. This represented an example of an ideal mixed model method that perfectly accounted for population structure, as a basis for comparing the other association methods. This study was the first time a straightforward method of correcting for population structure in Random Forest presented by Zhao et al. (2012) was evaluated for use on polygenic loci, and the third evaluation of its performance in general. In addition, we directly compare the performance of Random Forest analysis with more traditional mixed method analyses in identifying causal loci of highly polygenic traits. We found that all methods had low power to detect causal loci of polygenic phenotypes, but the performance Zhao's Random Forest implementation was similar to our Oracle reference test and could achieve similar power with fewer false positives than the next best performing method, CATE.

Random Forest Association

Our examination of the Random Forest correction method detailed in Zhao et al. (2012) agreed with general trends found in previous studies. This method has been evaluated twice thus far, both in simulations containing a small number of causal loci. In Zhao et al. (2012), it was tested using 100 simulated SNPs, 1 of which was causal. Zhao's correction increased the importance values of the associated SNP and decreased importance values of unrelated SNPs, although they did not directly measure false discovery rate. Brieuc et al. (2018) expanded this analysis by simulating 1000 SNPs, with 5 causal SNPs among them. Brieuc found that correction reduced the number of population divergent loci identified, but their simulations also detected a relatively large number of random SNPs identified as significantly associated with their trait. Our study expands on these tests by evaluating the performance of Zhao's Random Forest on polygenic traits with up to 30 causal loci, and explicitly investigating the influence of phenotype correlation with population structure on false positives. When both genotype and phenotype values were converted using Zhao's method, there were 35 total significant loci detected, including the 5 causal loci, 1 population structure associated locus, and 29 random loci. This was higher than the 20 total loci detected using uncorrected Random Forest in that study. In our study, we found that Zhao's Random Forest achieved the same average power as uncorrected Random Forest across all scenarios with greatly reduced false positives. Indeed, Zhao's Random Forest had the fewest false positives of any other test save the Oracle reference.

Machine learning analyses such as Random Forest have been proposed as methods for detecting polygenic loci under weak selection due to its ability to assess the performance of a locus in conjunction with other loci (Brieuc et al. 2018, Forester et al. 2018); however few studies have evaluated its power to detect loci under weak selection (Hoban et al. 2016, Forester et al. 2018). Previous comparisons have reported Random Forest has good power to detect disease-associated genes (Goldstein et al. 2010), but performs poorly when detecting SNP-SNP interactions in high dimension datasets (Winham et al. 2012, Wright et al. 2016). Most recently, Random Forest's ability to detect gene-environment interactions, which often correlate with geography and thus population structure, found that Random Forest could only reliably detect the strongest-effect loci and was dominated by false positive detections that reflect unrelated genetic structure in the dataset (Forester et al. 2018). Similarly, we found that Random Forest did not have substantially higher power to detect causal loci in highly polygenic scenarios relative to univariant methods such as LFMM and CATE, and our uncorrected Random Forest method produced high numbers of false positives relative to other methods. While Zhao's method of correcting Random Forest resulted in a drastically lower false positive rate, power to detect casual loci remained comparable to classic Random Forest.

In addition to comparing the performance of Random Forest methods to recent mixed model association tests, we evaluated the performance of empirical p-values and q-values in determining significance thresholds of Random Forest results. In both Random Forest methods, the magnitude of importance values for significant loci followed the same pattern as significance values (i.e. lower significance corresponded to higher importance). In simulations where uncorrected Random Forest produced a high number of false positives, true positives were distributed broadly throughout the range of significance values, rather than clustered at the top. Therefore, the performance of Random Forest in these cases was not due to the threshold at which importance measures were chosen as significant. Q-values used for Zhao's Random Forest performed very well in minimizing false positives while retaining power. However, calculating reliable empirical p-values for Random Forest was computationally intensive. Calculating a null distribution using 1000 permutations and models of 2000 trees took approximately 7 hours. This method may therefore be inappropriate for studies in which many different Random Forest tests need to be evaluated.

Final Recommendations

Of the methods evaluated in this study, none had high power to detect highly polygenic loci. Zhao's Random Forest had the fewest false positives in the presence of highly polygenic traits and false positives did not increase with increasing phenotype-structure correlation. When evaluating traits with few causal loci, Zhao's Random Forest and CATE performed equally well in terms of power vs false discoveries, although at different false discovery rate thresholds when correcting for multiple tests. For studies where identification of multiple low-effect loci is key, more recently proposed multivariant methods, such as a constrained ordination redundancy analyses, may hold more promise (Forester et al. 2018).

Due to the opportunistic nature of the phenotype information collected in this study, our samples included individuals of varied and unknown ages, complicating interpretation of our association results. To minimize confounding effects of environment on an individual's size, previous studies that have located significant genetic associations with condition factor and other growth-related traits have used single-age cohorts raised in controlled conditions (ex. Moghadam et al. 2007, Küttner et al. 2011). Nevertheless, the methods and laboratory protocols established in this study can be used on more precise phenotypes collected by targeted sampling, such as differences in growth-rate (Brown et al. 1998) or salinity tolerance (Cook 2010) seen at different latitudes, as well as different migratory (Gahagan et al. 2015) and spawning run (Secor et al. 2020) contingents.

Acknowledgements

We would like to thank Allen Curry and Samuel Andrews for providing tissue from Striped Bass in the Saint John River, Trevor Avery for providing scales from Striped Bass in Shubenacadie River, and Benjamin Reading for providing tissue samples from Striped Bass in Roanoke River and Cape Fear River. Sequencing services were provided by Génome Québec Innovation Centre, and funding was provided by NSERC, Canada Research Chair, and the New Brunswick Innovation Foundation.

Literature Cited

Barría, A., K. A. Christensen, G. M. Yoshida, K. Correa, A. Jedlicki, J. P. Lhorente, W. S. Davidson, and J. M. Yáñez. 2018. Genomic predictions and genome-wide association study of resistance against Piscirickettsia salmonis in coho salmon (*Oncorhynchus kisutch*) using ddRAD sequencing. G3: Genes, Genomes, Genetics 8:1183–1194.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57:289–300.

Brieuc, M. S. O., K. Ono, D. P. Drinan, and K. A. Naish. 2015. Integration of Random Forest with population-based outlier analyses provides insight on the genomic basis and evolution of run timing in Chinook salmon (*Oncorhynchus tshawytscha*). Molecular Ecology 24:2729–2746.

Brieuc, M. S. O., C. D. Waters, D. P. Drinan, and K. A. Naish. 2018. A practical introduction to Random Forest for genetic association studies in ecology and evolution. Molecular Ecology Resources 18:755–766.

Brown, J. J., A. Ehtisham, and D. O. Conover. 1998. Variation in Larval Growth Rate among Striped Bass Stocks from Different Latitudes. Transactions of the American Fisheries Society 127:598–610.

Bureau, A., J. Dupuis, B. Hayward, K. Falls, and P. Van Eerdewegh. 2003. Mapping complex traits using Random Forests. BMC genetics 4 Suppl 1:1–5.

Bureau, A., J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh. 2005. Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology 28:171–182.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22(11):3124–3140.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:7.

Cook, A. M. 2003. Growth and Survival of Age 0+ Shubenacadie River Striped Bass (Morone saxatilis) in Relation to Temperature and Salinity.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

DFO. 2014. Recovery potential assessment for the Bay of Fundy Striped Bass (*Morone saxatilis*) designatable unit. DFO Canadian Science Advisory Secretariat Science Advisory Report 2014/053.

Díaz-Uriarte, R., and S. Alvarez de Andrés. 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7:1–13.

Ekblom, R., and J. Galindo. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1–15.

Forester, B. R., J. R. Lasky, H. H. Wagner, and D. L. Urban. 2018. Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. Molecular Ecology 27:2215–2233.

François, O., and K. Caye. 2018. Naturalgwas: An R package for evaluating genomewide association methods with empirical data. Molecular Ecology Resources 18:789–797.

Frichot, E., S. D. Schoville, G. Bouchard, and O. François. 2013. Testing for associations between loci and environmental gradients using latent factor mixed models. Molecular Biology and Evolution 30:1687–1699.

Gahagan, B., D. Fox, and D. Secor. 2015. Partial migration of Striped Bass: Revisiting the contingent hypothesis. Marine Ecology Progress Series 525:185–197.

Goldstein, B. A., A. E. Hubbard, A. Cutler, and L. F. Barcellos. 2010. An application of Random Forests to a genome-wide association dataset: Methodological considerations and new findings. BMC Genetics 11.

Healy, T. M., R. S. Brennan, A. Whitehead, and P. M. Schulte. 2018. Tolerance traits related to climate change resilience are independent and polygenic. Global Change Biology 24:5348–5360.

Hoban, S., J. L. Kelley, K. E. Lotterhos, M. F. Antolin, G. Bradburd, D. B. Lowry, M. L. Poss, L. K. Reed, A. Storfer, and M. C. Whitlock. 2016. Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. American

Naturalist 188:379–397.

Hooker, G., and L. Mentch. 2019. Please stop permuting features: An explanation and alternatives. arXiv.

Korte, A., and F. Ashley. 2013. The advantages and limitations of trait analysis with GWAS: a review. Plant methods 9:29.

Küttner, E., H. K. Moghadam, S. Skúlason, R. G. Danzmann, and M. M. Ferguson. 2011. Genetic architecture of body weight, condition factor and age of sexual maturation in Icelandic Arctic charr (Salvelinus alpinus). Molecular Genetics and Genomics 286:67–79.

LeBlanc, N. M., B. I. Gahagan, S. N. Andrews, T. S. Avery, G. N. Puncher, B. J. Reading, C. F. Buhariwalla, R. A. Curry, A. R. Whiteley, and S. A. Pavey. 2020. Genomic population structure of Striped Bass (*Morone saxatilis*) from the Gulf of St. Lawrence to Cape Fear River. Evolutionary Applications:1–19.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 26:589–95.

Liaw, A., and M. Wiener. 2015. randomForest: Breiman and Cutler’s random forests for classification and regression.

Maruki, T., and M. Lynch. 2015. Genotype-frequency estimation from high-throughput sequencing data. Genetics, 201:473–486.

Maruki, T., and M. Lynch. 2017. Genotype calling from population-ge- nomic sequencing data. G3: Genes, Genomes, Genetics, 7:1393–1404.

Moghadam, H. K., J. Poissant, H. Fotherby, L. Haidle, M. M. Ferguson, and R. G. Danzmann. 2007. Quantitative trait loci for body weight, condition factor and age at sexual maturation in Arctic charr (*Salvelinus alpinus*): Comparative analysis with rainbow trout (*Oncorhynchus mykiss*) and Atlantic salmon (*Salmo salar*). Molecular Genetics and Genomics 277:647–661.

Poland, J. A., P. J. Brown, M. E. Sorrells, and J. L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7(2).

Purcell, S., and C. Chang. (n.d.). PLINK 1.9.

Rice, W. R. 1989. Analyzing Tables of Statistical Tests. Evolution 43:223–225.

Rockman, M. V. 2012. The QTN program and the alleles that matter for evolution: All that's gold does not glitter. Evolution 66:1–17.

Santure, A. W., and D. Garant. 2018. Wild GWAS—association mapping in natural populations. Molecular Ecology Resources 18:729–738.

Secor, D. H., M. H. P. O'Brien, B. I. Gahagan, D. A. Fox, A. L. Higgs, and J. E. Best. 2020. Multiple spawning run contingents and population consequences in migratory Striped Bass Morone saxatilis. PLoS ONE 15.

Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100:9440–9445.

Sylvester, E. V., P. Bentzen, I. R. Bradbury, M. Clément, J. Pearce, J. Horne, and R. G. Beiko. 2017. Applications of random forest feature selection for fine-scale genetic population assignment. Evolutionary Applications:1–13.

Timpson, N. J., C. M. T. Greenwood, N. Soranzo, D. J. Lawson, and J. B. Richards. 2018. Genetic architecture: The shape of the genetic contribution to human traits and disease. Nature Reviews Genetics 19:110–124.

Vilhjálmsson, B. J., and M. Nordborg. 2013. The nature of confounding in genome-wide association studies. Nature Reviews Genetics 14:1–2.

Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang. 2012. Five years of GWAS discovery. American Journal of Human Genetics 90:7–24.

Wang, J., Q. Zhao, T. Hastie, and A. B. Owen. 2017. Confounder adjustment in multiple hypothesis testing. Annals of Statistics 45:1863–1894.

Winham, S. J., C. L. Colby, R. R. Freimuth, X. Wang, M. de Andrade, M. Huebner, and J. M. Biernacka. 2012. SNP interaction detection with Random Forests in high-dimensional genetic data. BMC Bioinformatics 13.

Wright, M. N., A. Ziegler, and I. R. König. 2016. Do little interactions get lost in dark random forests? BMC bioinformatics 17:145.

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price. 2014. Advantages and pitfalls in the application of mixed model association methods. Nature Genetics 46:100–106.

Zhao, Y., F. Chen, R. Zhai, X. Lin, Z. Wang, L. Su, and D. C. Christiani. 2012. Correction for population stratification in random forest analysis. International Journal of Epidemiology 41:1798–1806.

Zhou, X., and M. Stephens. 2013. Genome-wide efficient mixed model analysis for association studies. Nature Genetics 44:821–824.

CHAPTER 4: A new automated text-mining pipeline for ecological associations

1. Nathalie M. Leblanc (corresponding author)

Department of Biological Sciences, Canadian Rivers Institute, University of New Brunswick, Saint John, NB, E2L 4L5, Canada. Email: nleblan5@unb.ca

Email: nathalie.leblanc@unb.ca

2. Christopher J. O. Baker

Department of Computer Science, University of New Brunswick, Saint John, Canada

3. Scott A. Pavey

Department of Biological Sciences, Canadian Rivers Institute, University of New Brunswick, Saint John, NB, E2L 4L5, Canada. Email: spavey@unb.ca

Abstract

A consequence of rapid advances in ecological genomics, there is an exponential increase in the quantity of literature detailing ecological function of genes. Text mining has potential to be a useful tool to acquire ecological inferences about particular genes from thousands of scientific papers. We created a rule-based text-mining pipeline in the General Architecture for Text Engineering (GATE) that scans full-text articles and extracts sentences containing gene-ecology associations. In total, we created 6 rule sets and 3 gazetteers, which we combined with an additional gazetteer containing 2744 terms taken from the existing Environment Ontology (ENVO), high level tokenization tools contained in A Nearly-New Information Extraction System (ANNIE), and two named entity recognition tools created for gene and protein annotation. When tested on 104 manually annotated papers, our pipeline had high recall (0.88) but relatively low precision (0.36.) When run using a single gene as a search target, this tool was capable of identifying a small number of potentially useful papers out of the test corpus. The rules and vocabularies created as part of this pipeline serve as a prototype pipeline for ecological text-mining that will serve as essential knowledge resources for the subsequent development of knowledge management infrastructures specific to Ecological Genomics. This will support scientists who wish to mine existing literature for relevant ecological annotations for a gene of interest, or identify genes known to be associated with an ecological trait of interest.

Introduction

As the number of publications reporting novel biological findings increases and new molecular techniques generate an increasing amount of data in a single study, there exists a need for an efficient way to organize and catalogue these data quickly and efficiently. Text-mining, the automated analysis of natural language text, has been a key tool in the biomedical field for extracting information from an exponentially expanding pool of literature (Landry and Aubin-Horth 2007, Fleuren and Alkema 2015). Many text-mining tools focus on accurately identifying specific *entities* within the text in a process called named entity recognition (Witte et al. 2007, Harmston et al. 2010). In biology, these entities comprise categories of words such as gene and protein names (Wang et al. 2018), organism names (Naderi et al. 2011), or specific mutations (Caporaso et al. 2007). Once entities have been identified and annotated they can be used to normalize multiple synonyms into standardized vocabulary, such as collecting the many disparate terms used to refer to a single protein under one standard term (Cohen 2005). Text-mining can also extract *relations* between annotated entities, in a process called relation detection (Witte and Baker 2007, Harmston et al. 2010). Text-mining tools are able to collect vast amounts of information generated about model organisms, human disease, and molecular function of genes, and techniques are continually being refined to improve both the power and the accuracy of text-mining algorithms (Zweigenbaum et al. 2007, Huang and Lu 2016, Pérez-Pérez et al. 2019). In order to evaluate a given tool, the tool is run on an assembly of papers called a corpus that has been manually annotated by one or more people to identify the terms or sentences that should be targeted by the algorithm. Performance is measured using two metrics that have long been used in information retrieval (Kent et al. 1955): precision (number of true positives, over the sum of all true and false positives) and recall (number of true positives, over the sum of all true positives and false negatives).

The increasing accessibility of next-generation sequencing tools to ecological studies has led to an explosion of genetic data generated for non-model organisms for which little functional genetic information is available. This increased amount of data has resulted in a corresponding increase in number of un-annotated genes (Pavey et al. 2012). Functional annotation of genes even in model organisms such as yeast can be challenging when genes are only present in a subset of organisms or cannot be readily observed in the lab (Peña-Castillo and Hughes 2007). In non-model organisms, genetic annotations are often not available and must be taken from the closest model organism, which may be only very distantly related and likely to have differences in the functions of its genes (Alvarez et al. 2015). At the same time, there is an increasing amount of data on the association of sequences, both annotated and un-annotated, to ecologically relevant variables that is not included in existing centralized databases (Pavey et al. 2012). Attaching or linking this information to a gene is a process known as ecological annotation, which complements functional annotation by placing the biology of a gene into the context of the environment it evolved in (Landry and Aubin-Horth 2007). There exists a need for an automated data mining tool able to identify and explore existing ecological information about sequences in order to supplement functional data and generate hypotheses (Pavey et al. 2012, Andrew et al. 2013). For example, upon identifying a potential candidate gene in an ecologically relevant species molecular ecologists should be able to quickly survey existing literature to identify associations found in previous studies and identify existing patterns.

Text-mining tools typically draw information from either one or more databases, or parse text present in abstracts or full text articles in the literature. Text-mining on abstracts and full text manuscripts must be able to deal with complex sentence construction, and large amounts of variation in terminology (Cohen et al. 2010). Abstracts have traditionally been used for literature-based text-mining, due to both their greater accessibility and their information-rich nature (Cohen et al. 2010, Westergaard et al. 2018), but use of full text articles has increased as open-access articles become more common as they allow for the retrieval of information that is simply not present in abstracts (Cohen et al. 2010). Moreover, abstracts often contain only a subset of key findings in a study, and those findings are summarized. In ecotoxicology, for example, abstracts examined for a text-mining pipeline often did not directly mention species or toxins under study or did not mention all species or toxins used in the study (Baker and Riazanov n.d.). A biomedical text-mining tool that can be used on abstracts or full text found that recall increased from 45% to 95% when full text was searched (Müller et al. 2004). A recent study that demonstrated extraction of protein-compartment associations performed better on full text articles than on abstracts only (Westergaard et al. 2018).

Outside of the biomedical field, text-mining tools have only recently begun to be developed for use with ecological data, mining both databases and natural language texts. Existing tools have recently been leveraged to conduct systemic reviews, such as a study that used two text-mining tools designed to identify geographic and taxonomic entities to identify wide-ranging trends in the distribution of taxa across the globe (Millard et al. 2020). Some early examples of text-mining tools developed specifically for ecology include Seqenv (Sinclair et al. 2016) and its expansion (Ijaz et al. 2017), which matches user-given sequences to the NCBI database and normalizes their location metadata using the Environment Ontology. A more recent pipeline used 50 aquatic food web studies to design a tool that extracted sentences associating a specific taxa name to a body size measurement (Compson et al. 2018). To date, there is no existing text-mining resource intended to mine full-text articles for associations between genes or sequences and ecological concepts commonly involved in adaptive divergence, inclusive of not only geographic location but also climate variables, sex, movement patterns and behavior.

In this paper, we present new tools for the extraction of ecologically relevant information relating to genes and proteins. These tools represent a first step toward a text-mining pipeline that can extract and automatically annotate genes with gene-environment associations, similar to tools in the biomedical field that aid in functional annotation of genomes (Aerts et al. 2008, Haeussler et al. 2011). We demonstrate that standard text-mining methods can be used to successfully identify and extract over 85% of target sentences containing gene-ecological associations. To build and test these tools, we created a corpus of 137 peer-reviewed articles, manually annotated for sentences that describe associations between specific genes or gene families and an ecological variable. Using this corpus, we designed a pipeline comprising of existing entity-recognition tools, subject specific gazetteers and rules to specifically recognize a wide range of ecologically relevant variables and common ‘connector’ terms related to gene presence or expression.

Methods

Corpus Development

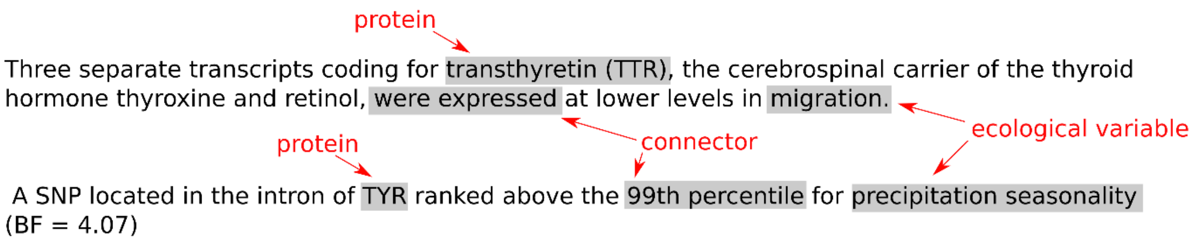
While there exist multiple gold standard corpora for testing new biomedical text-mining tools (Wissler et al. 2014), there are none available for ecological adaptation. In order to develop our text-mining pipeline we first compiled a corpus of peer-reviewed articles containing associations between genes and ecological variables. The primary step was to review publications from the Journal Molecular Ecology for relevant titles and abstracts indicating an association between DNA and some ecological variable. We then expanded our search using citations within those articles to develop a list of keywords for use in querying Google Scholar. Articles were chosen if they contained at least one sentence where all of the following components co-occur: a gene or protein, an ecological variable, and a ‘connector’ term, which was defined as any word that indicates an association between the former two annotations. This included terms indicating transcription differences, regulation, outlier patterns of differentiation, and codon differences of a gene in the presence of an ecological variable (Figure 4.1). Chosen articles were downloaded as PDFs and then converted into plain text files using pdftotext v4.02 (FooLabs 2014), and filtered further with a custom bash file. Our corpus included both gene expression and DNA outlier studies. When sorting articles into training and test datasets, we endeavored to ensure an equal composition of papers focusing on RNA vs

DNA, as well as different types of ecological variables. All articles were manually annotated by identifying sentences containing these three elements (Figure 4.1).

Due to time constraints, our initial corpus consisted of 26 articles. These articles were used to construct custom lists of words and terms (gazetteers) as well as rules detailing patterns of letters and words (JAPE rules; Cunningham et al. 2000), both of which are used to annotate relevant terms within the text. While relatively small, this corpus contained articles covering a wide range of ecological variables commonly investigated in molecular ecology, ranging from transgenic agricultural studies, studies testing genetic association with stress, migration, bleaching, or pollution, and outlier analyses of local adaptation to habitat. An additional corpus of 104 articles was then compiled to test recall and precision of our pipeline.

Pipeline Development – Annotation Groups

Annotations were grouped into three main categories: gene or protein names, ecological or environmental variables, and ‘connector’ terms relating the first two categories to each other, and annotation was done within the General Architecture for Text Engineering (GATE) v. 8.4 (Cunningham 2002). Each category of annotation was identified and annotated separately to allow for a modular implementation of the resulting gazetteers and rules. The first annotation group, gene and protein names, have been targeted in many other text-mining tools, while the latter two categories have not previously been targeted by text-mining tools. **[mention connector ontology]** It was these last two categories that we focused on when creating our new tools.



Three separate transcripts coding for transthyretin (TTR), the cerebrospinal carrier of the thyroid hormone thyroxine and retinol, were expressed at lower levels in migration.

A SNP located in the intron of TYR ranked above the 99th percentile for precipitation seasonality (BF = 4.07)

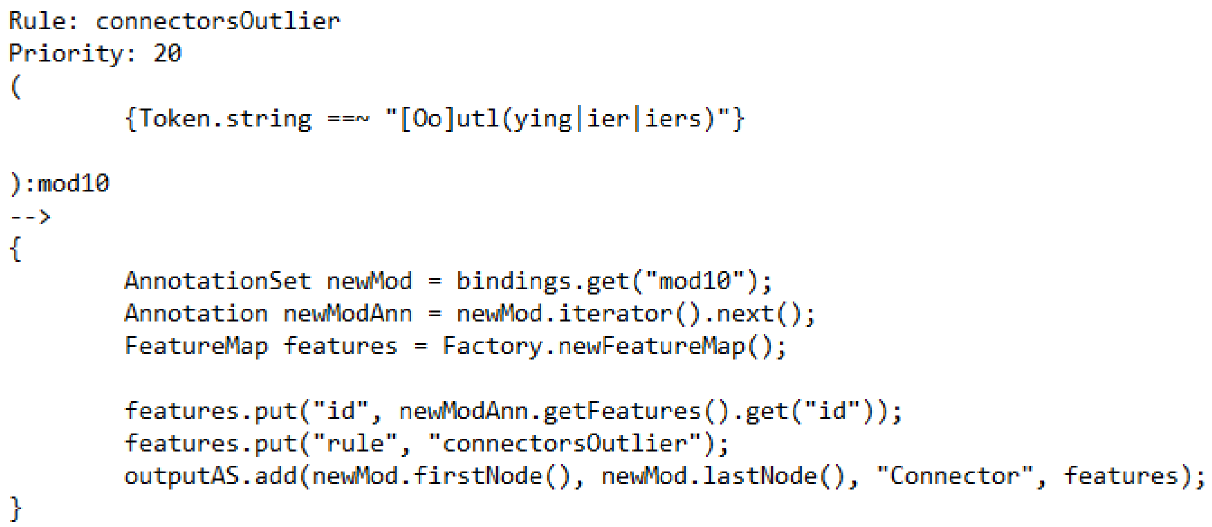
Figure 4.1. Two sentences identified during manual curation of articles, with annotation targets shaded in grey and labelled as one of three annotation categories.

The annotation of genes and proteins via named entity recognition is an ongoing process for which a number of tools have already been developed and are constantly being refined (Wang et al. 2018). While these tools were created for use in biomedical articles, gene names remain the same in medicine and ecology, and so we adopted two such named entity recognition tools for use within GATE: ABNER (Settles 2005) and Penn's BioTagger (McDonald and Pereira 2005). When used on biomedical corpora, ABNER has a recall of 77.8% and a precision of 68.1% (Settles 2005) and Penn's BioTagger has a recall of 78.7% and a precision of 86.4% (McDonald and Pereira 2005). While these tools are both considered very good, we wanted to maximize the number of successfully annotated genes in our corpora. Annotations from both taggers were therefore combined to increase total number of annotated putative gene names and combined with additional annotation rules downstream.

Using manually annotated sentences in our test corpus, we initially created gazetteers of ecological variables (e.g. hypoxia, migration, precipitation) and connectors (e.g. associated, outlier, expressed, percentile) used in the articles along with synonyms and related terms. Based on common variations and patterns seen in gazetteer vocabulary, we created a series of rules using the Java Annotation Patterns Engine (JAPE) implemented in GATE to capture as many variations of terms as possible (see Figure 4.2). We refined our gazetteers and JAPE rules iteratively by running the existing pipeline on our training corpus and calculating precision, recall, and F-Score. In text-mining, recall represents the number of true positive annotations made by the text-mining algorithm is divided by the total number of manual annotations, and precision represents the number of true positive annotations is divided by the total number of automatically extracted annotations. The relative performance of both these metrics can be described using the overall harmonic mean of precision and recall values, known as the F-Score (Witte and Baker 2007, Sagayam et al. 2012), which is calculated using the following equation:

$$F - Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

We then ran our pipeline on our second, independent test corpus and calculated precision and recall again to determine the performance of our pipeline on literature it was not specifically designed on.



```
Rule: connectorsOutlier
Priority: 20
(
	{Token.string ==~ "[Oo]utl(ying|ier|iers)"}

):mod10
-->
{
	AnnotationSet newMod = bindings.get("mod10");
	Annotation newModAnn = newMod.iterator().next();
	FeatureMap features = Factory.newFeatureMap();

	features.put("id", newModAnn.getFeatures().get("id"));
	features.put("rule", "connectorsOutlier");
	outputAS.add(newMod.firstNode(), newMod.lastNode(), "Connector", features);
}
```

Figure 4.2. JAPE rule matching one common connector term, targetting outlier loci. The “right” side of the rule is located within the first set of brackets and tells the pipeline to look at the content of each word and find one that matches “outlying”, “outlier” or “outliers”. When found, the word is captured and passed to the ”left” side of the rule, located within the second set of brackets. Here, matching words are retrieved and saved into an annotation called “Connector”.

Pipeline Implementation

Briefly, the pipeline had the following steps (Figure 4.3). PDFs were converted to unstructured text using pdftotext, and periods contained in the terms ‘Fig.’ and ‘et al.’ were removed. We use GATE’s high-order annotation tools, A Nearly-New Information Extraction System (ANNIE), for initial processing of texts: annotating words and punctuation as tokens for later processing, identifying the beginning and end of sentences while attempting to distinguish between abbreviation periods and sentence-ending periods, and flagging tokens by type of word or text. The gazetteer **Units** extracted common units of measurement (e.g. ppb, pptr, mg/kg, µg/kg). Gene and protein names were annotated using machine-learning approaches implemented in ABNER and Penn BioTagger. The file **potential references** contained JAPE rules that annotated any case-sensitive instance of the word “References” and all text between that and the end of the text file, and **reference finalizer** kept the last such annotation, if there was more than one. This successfully captured the section of an article containing only citation information while leaving the previous sections unannotated, even in text files converted from pdfs in two column format. The JAPE transducer **Lookup Transducer** consolidated gazetteer-based annotations into consistent annotation types, while the JAPE transducer **Protein Finalizer** consolidated annotations from both protein taggers, removing any that consisted of more than 4 single letters separated by spaces, or that contained the word “Fig.” Additionally, this transducer contained rules that annotated any terms containing words with all capitals or mixed capitals followed by a number, and terms that consist of “hsp” followed by a number. For all protein annotations, terms were only annotated if they were not contained within a recognized ‘references’ section. We also tagged any reference to the term “gene” or “protein” to minimize number of false negatives due to protein annotation, as the focus of this project was the development of annotation rules for association terms and ecological variables. Connector terms were identified using the JAPE transducer **Connectors**, which contained 17 rules identifying words commonly used to describe an association between a genetic locus and an ecological variable (e.g. “expressed”, “troughed”, “outlier”, “ranked”). Finally, the JAPE transducer **Environment Variables** contained 4 rules matching common root words in ecological traits measured in molecular ecology, along with possible variations (ex. Any term ending in stress/stressed/stresses/stressor).

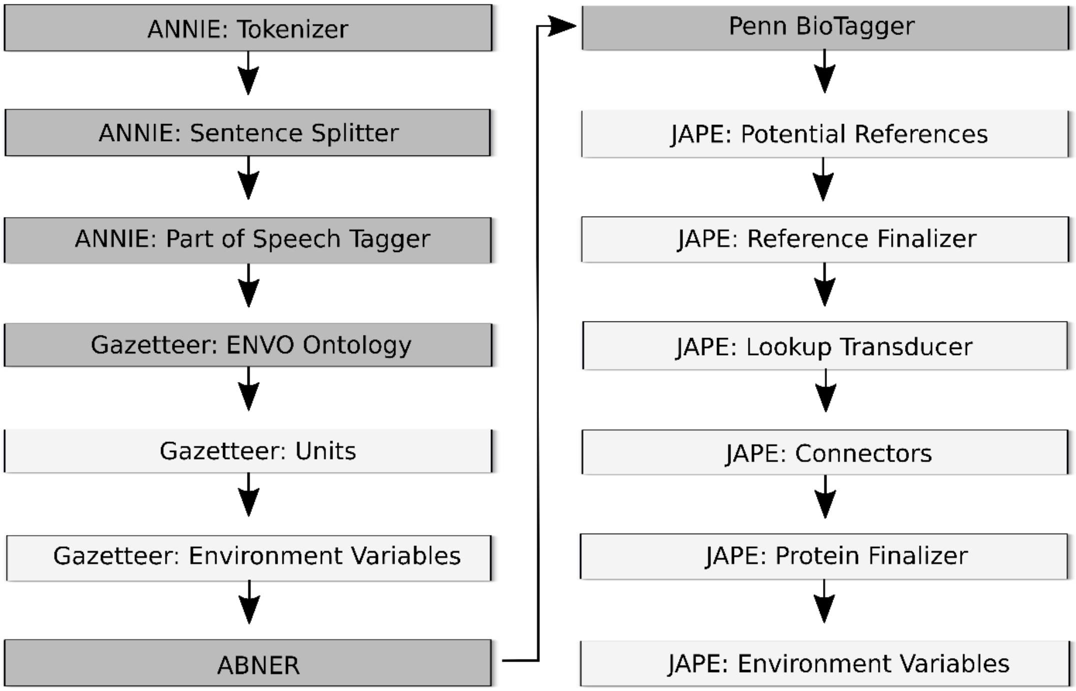


Figure 4.3. Visualization of an ecological annotation pipeline workflow, as implemented during our final run. Components developed during this project are represented as white boxes, while existing components used during the pipeline are represented as grey boxes.

In real use-case scenarios, scientists will likely be using our pipeline to search for ecological associations with a particular gene or gene family, or for genes associated with particular ecological variables. We tested our pipeline's performance searching for specific genes by modifying gene and protein annotation to search for a list of synonyms for the gene '*Per3*' (Period Circadian Protein Homolog 3; see supplementary file for full list), using the same corpus of 104 articles, and recorded number of extracted sentences with complete association data, as well as total number of extracted sentences.

Results

Our custom pipeline extracted 6369 sentences from the 104 test articles. We recorded number of extracted sentences that were present in manual annotations compared to total number of extracted sentences. When run on the 104 paper test corpus, our pipeline extracted 88% of manually annotated sentences, giving it a recall of 0.88. Precision was much lower: 17% of extracted sentences were manually annotated target sentences, giving a precision of 0.17. Within the remaining extracted sentences, we identified a number of relevant sentences that were not found by manual annotators. These additional extracted sentences were considered positive annotations if, out of context, they provided some useful information about expression patterns of a gene, including specific gene or family names and their relation to specific environmental conditions, tissue locations, or other specific associated genes. When these sentences were added to the list of true positive annotations, precision rose to 0.36. The remaining extracted sentences consisted of sentences with language too vague to interpret out of context, methods sentences, speculative sentences found in the introduction, tables, table or figure captions, and headers or metadata. The recall and precision values reported here correspond to an F-score of 0.51.

When our pipeline was used to search for ecological associations specific to the gene ‘Per3’, we obtained a total of 33 extracted sentences taken from 4 of our 104 articles (Table 4.1). Of these sentences, 17 (52%) from 2 articles contained association information connecting Per3 to an ecologically relevant trait, and 15 of these sentences were from a single article. Using these extracted sentences, the user is able to quickly identify 1 or 2 articles in our corpus to read in-depth and is given an outline of what information is contained in each of these articles relating to the gene of interest.

Table 4.1. 33 sentences extracted from 104 articles, when our ecological association text-mining pipeline was run with the gene ‘Per3’ specified. Sentences with association data in them are marked as ‘good’.

Article	*Good*	*Extracted Sentence*
Brandstätter2001		In mammals, the master clock controlling circadian rhythmicity is located in the hypothalamic suprachiasmatic nuclei (SCN) which are characterized by distinct anatomical, physiological and neurochemical features [4,5] as well as the presence of speci(R)c transcription factors and clock genes, particularly three homologues (Per1, Per2, Per3) of the Drosophila clock gene period (per) [6].
Helfer2006		To investigate avian clock gene expression, partial cDNA sequences of six mammalian clock gene homologs (Bmal1, Clock, Per2, Per3, Cry1, and Cry2) and a novel avian cryptochrome gene (Cry4) were cloned from the house sparrow, a model system in circadian research.
Helfer2006		Keywords Birds, Circadian clock, Clock genes (Bmal1, Clock, Per2, Per3, Cry1, and Cry2), Circadian rhythm, Bird brain INTRODUCTION The discovery of genes that are directly associated with circadian clock function has lead to great progress in our understanding of the basic molecular mechanisms of circadian oscillations.
Helfer2006		In the pineal gland, the eye and the hypothalamus of both chicken and quail Per3 mRNA levels also exhibit robust circadian rhythms, with highest expression during the late night and a trough during the early day (Yoshimura et al., 2000; Yamamoto et al., 2001; Yasuo et al., 2003).
Helfer2006		Temporal Expression Patterns of Clock Genes in the House Sparrow Brain All genes examined in this study (Bmal1, Clock, Per2, Per3, Cry1, Cry2, Cry4, Gapdh, b-Actin) were expressed in the house sparrow brain.
Helfer2006	Yes	All genes believed to be part of the central circadian rhythm-generating mechanism (Bmal1, Clock, Per2, Per3, Cry1, and Cry2) showed pronounced 24 h rhythms of mRNA expression (Figures 2 and 3).
Helfer2006	Yes	Per3 mRNA (Figure 2D) peaked at ZT 22.5, i.e., 2 h phase-advanced to Per2 (one-way ANOVA, p , 0.00001).
Helfer2006		However, it was 122 G. Helfer et al FIGURE 2 Temporal expression profiles of Bmal1, Clock, Per2, Per3, Gapdh, and b-Actin in the house sparrow brain in birds kept in LD 12:12.
Helfer2006		In mammals, three different Per homologs (Per1, Per2, and Per3) have been described, of which only two are believed to be part of the core circadian transcriptional/translational feedback loop (Shearman et al., 2000; Bae et al., 2001).
Simonneaux2004	Yes	Consistently, injection of the h-adrenergic antagonist propranolol at nighttime markedly reduced the level of Aa-nat, Per1 and Cry2 mRNA but not that of Per3 and Cry1 mRNA (Fig 5).
Simonneaux2004	Yes	In contrast, Per3 and Cry1 mRNA levels were not significantly altered by the adrenergic agonist (Fig 4).

Simonneaux2004		Effect of daytime injection of a h-adrenergic agonist on Aa-nat, Per1, Per3, Cry2 and Cry1 mRNA levels in the rat pineal gland.
Simonneaux2004	Yes	Strengthening the previous observations made with the adrenergic antagonist, light exposure at night markedly reduced Aa-nat, Per1 and Cry2 gene expression, but not that of Per3 and Cry1 (Fig 6).
Simonneaux2004		Daily rhythm in Aa-nat, Per1, Per3, Cry2 and Cry1 mRNA levels in the rat pineal gland.
Simonneaux2004		Briefly, Aa-nat, Per1, Per3, Cry1 and Cry2 sense and antisense riboprobes were synthesized from 1 Ag of linearized plasmid (2 h at 37jC) using either T7, T3 or SP6, as appropriate, RNA-polymerase (MAXIscript transcription kit, Ambion; a[35S]-UTP, 1250 Ci/mmol, NEN-Dupond, Zaventem, Belgium), then purified following ethanol and ammonium acetate treatment.
Simonneaux2004		Circadian variation in Aa-nat, Per1, Per3, Cry2 and Cry1 mRNA levels in the rat pineal gland.
Simonneaux2004	Yes	Daily and circadian variations in pineal gene expression The mRNA level of the four clock genes Per1, Per3, Cry1 and Cry2 displayed significant daily variation with higher nocturnal values, the largest increase being observed for Per1 mRNA and the smallest for Cry1 mRNA (Fig 2).
Simonneaux2004		The nocturnal increase in Per3 mRNA may therefore be induced by another intracellular pathway triggered by one of the other neurotransmitters present in the rat pineal gland (Refs.
Simonneaux2004	Yes	Per3 mRNA levels are not altered by light exposure or adrenergic antagonist given at nighttime or adrenergic agonist administrated at daytime.
Simonneaux2004	Yes	Although analyses on low levels of mRNA are not easy, our results suggest that like Per3, the nocturnal increase may not be regulated by the clock-controlled norepinephrine nor by light.
Simonneaux2004	Yes	A marked daily and circadian rhythm of Per2 gene expression was observed in the rat pineal gland [14,39] but in vivo injection of an adrenergic agonist [39] or in vitro application of norepinephrine did not increase Per2 mRNA level similar to our Per3 results [15].
Simonneaux2004	Yes	Additionally, we showed that transcription of Per3, Cry1 and Cry2 is increased during the night, as already reported for Per1 and Per2 gene.
Simonneaux2004	Yes	In this study, we demonstrated that Per3 gene is also expressed in the rat pineal gland.
Simonneaux2004	Yes	Our study reports that Per3 gene as well is highly expressed in the rat pineal gland with a 5- 10-fold increase during the night.
Simonneaux2004		Effect of nighttime injection of a h-adrenergic antagonist on Aa-nat, Per1, Per3, Cry2 and Cry1 mRNA levels in the rat pineal gland.
Simonneaux2004		Effect of light exposure at night on Aa-nat, Per1, Per3, Cry2 and Cry1 mRNA levels in the rat pineal gland.
Simonneaux2004	Yes	The expression of Per3 and Cry1 displays a daily rhythm not regulated by norepinephrine, suggesting the involvement of another day/night regulated transmitter(s).
Simonneaux2004		The endogenous rhythmicity of the hypothalamic pacemaker relies upon genetic regulation involving a set of ''clock genes: three Period genes (Per1, Per2, Per3), two Cryptochrome genes (Cry1 * Corresponding author.
Simonneaux2004	Yes	In the present study, we report that the other clock genes, Per3, Cry1 and Cry2, are expressed in the rat pineal gland.
Simonneaux2004		In addition, we have analyzed, in parallel to that of Aa-nat, how expression of Per1, Per3, Cry1 and Cry2 coding genes are regulated by circadian, adrenergic and light components.
Simonneaux2004	Yes	Per1, Per3, Cry2 and Cry1 clock genes are expressed in the pineal gland and their transcription is increased during the night.
Simonneaux2004	Yes	In contrast, Per3 and Cry1 day and night mRNA levels are not responsive to adrenergic ligands (as previously reported for Per2) and daily expression of Per3 and Cry1 appears strongly damped or abolished in constant darkness.
Yoshimura2000	Yes	It seems that three PER proteins (PER1, PER2, PER3) also somehow negatively regulate CLOCK-BMAL1 s transcriptional activity [18,12,15,24].

New JAPE rules and gazetteers have been cleaned and commented and will be freely available on Dryad once published. We have re-organized our rules in archived versions of files to improve modularity, so our broader rules encompassing non-specific ‘gene’ terms have been separated into their own file. Additionally, we have created separate files for connector terms used in gene expression vs DNA studies, for downstream applications focused on only one type of study. We have also uploaded manual annotations for each of our corpus articles, along with formatted citation information for each, in excel files on Dryad.

Discussion

Our pipeline represents the first attempt to annotate ecologically relevant traits for use in ecological association studies across multiple taxa and environments. Currently, our tools export single sentences in spreadsheet format that researchers can examine for information relevant to their search. As implemented, the spreadsheet can be used to identify existing ecological associations with a given protein, or existing protein associations with a given ecological trait. This can be used to supplement searches for functional annotation of a given gene, or to identify possible candidate genes for follow-up studies of ecologically important genes. This text-mining framework can easily be adapted for use in scenarios where researchers are searching for specific genes or ecological variables, similar to relation-based search engines such as Textpresso and Pharmspresso that display sentences containing relations between genes, diseases, or drugs (Müller et al. 2004, Garten and Altman 2009).

Our preliminary tests demonstrate that ecological associations can be successfully extracted from both gene expression and outlier studies with high recall, despite challenges inherent in anticipating both the breadth of possible ecological traits, and variations in terminology used for those traits. The tools created in this study represent initial steps and proof of concept of a text-mining pipeline able to extract a broad range of genetic ecological associations, even with limited integration of ontologies. We have also identified five key areas of improvement needed to develop this initial pipeline into one ready for general use, described below.

User Interface

The present study focused on the development of text-mining resources for use in an ecological association pipeline, but a complete pipeline also requires a command-line interface to allow users to input desired search terms or parameters. An immediate next step before the release of our pipeline is the implementation of command-line parameters that allows users to specify target gene terms and/or target ecological variables. Our pipeline will be able to substitute user-provided gene or ecology terms, loaded as gazetteers, for more general gene and ecological variable annotation tools as required.

Corpora

As the pipeline created in this study develops, a larger corpus will be assembled to both train and test improvements. In addition, the corpora used in this study were assembled by expert and non-expert curators, and have not passed through the expensive and time-consuming process required to obtain high quality, gold standard certification (Wissler et al. 2014). Identification of relevant sentences in articles that were not found during manual curation highlights the challenges inherent in creating comprehensive annotations of new corpora. Further improvements can be made to our manually curated corpora, and an expanded corpus would allow for greater coverage when searching for patterns among “false positive” annotations. Additionally, the larger the corpus used for recall and precision evaluation, the more representative these values are to different terminology present in the literature.

Improving recall by expanding coverage of ecological variables

The high variability in type and terminology of ecological variables seen in literature is a challenging aspect of information extraction in natural language documents, particularly when combined with the increased sentence complexity found in full texts compared with abstracts (Cohen et al. 2010). Dictionary and rule-based text-mining tools in particular are unable to reliably capture all possible variations of terms or adapt to changing vocabulary or novel entities (Wang et al. 2018). The rules and gazetteers written for our pipeline were able to identify ecological variables in 104 articles with high recall, but ensuring our pipeline remains robust to the broadest range of possible ecological variables remains a priority. The limitations of discrete rules in identifying unforeseen terminology are primarily addressed in two ways: incorporating comprehensive ontologies of concepts within a field (Spasic et al. 2005), and training machine-learning algorithms to recognize target concepts without having to predict all possible terms beforehand (Wang et al. 2018).

Ontologies, networks of concepts with defined relationships to each, are a common method of leveraging a wide range of standardized terminology and synonyms (Spasic et al. 2005). Ontologies allow for the detection of different abbreviations and synonyms of a concept and the collection of these different terms under a single, unambiguous descriptor (normalization; Witte et al. 2007). In text-mining, ontologies are crucial both in a starting point for concept matching and terminology standardization, and in enabling higher-order relations between concepts to be included in a text-mining pipeline (Spasic et al. 2005). Initial text-mining tools developed for ecotoxicology made use of chemical identifiers taken from the Chemical Entities of Biological Interest (ChEBI) ontology (Degtyarenko et al. 2008) for text annotation (Baker and Riazanov n.d.), or functional information from gene ontology to supplement annotations (Chepelev et al. 2011). In lipodomics, an ontology was created to consolidate the wide range of non-standard lipid name synonyms with existing databases, and then instantiated that ontology on a corpus of articles (Baker et al. 2007). Each sentence containing a lipid-protein association was stored as a relation to the corresponding lipid ('lipid' appears in 'sentence') in the ontology. Unfortunately, while there are numerous ontologies collecting terms in biomedical research (Spasic et al. 2005) and for physio-chemical and toxicity entities (Hardy et al. 2012), ontologies are a new development in the field of ecology.

Interest in adapting ontologies as a method of centralizing and standardizing ecological data began about a decade ago (Baird et al. 2011, Rubach et al. 2011, Pavey et al. 2012), and ontologies have been successfully created and utilized in text-mining applications in ecology to date. A crowd-sourced ontology of geographic and physical objects exists in the form of ENVO (Buttigieg et al. 2016), a tool that this pipeline made use of to expand ecological annotations; however this ontology is limited in scope to physical location descriptors. In botany, an ontology was recently developed that attempts to provide a comprehensive collection of ecological concepts used to characterize plants, particularly individual plant phenotypes and physical parts (Garnier et al. 2017). In future iterations of this pipeline, more extensive integration of domain-specific ontologies, where they exist, could further improve recall. Based on the range of ecological variables in all 137 articles examined in this initial study, we have identified both the CheBi and TOP ontologies as containing terminology used in sentences that contain gene-ecology association data. In addition, ontologies of relationship terms such as the Semanticscience Integrated Ontology (Dumontier et al. 2014) will be tested as a supplement to our existing dictionary of connector terms. This will also produce a pipeline that is more robust when testing is expanded to larger corpora.

The second method of improving flexibility of a text-mining approach is known as machine learning, where algorithms are trained to recognize entities in already annotated articles and then apply the patterns detected to new, un-annotated articles (Basaldella et al. 2017, Wang et al. 2018). Machine learning methods are popular in gene and protein entity recognition in biomedicine, as they allow for recognition of gene names that are not present in gazetteers or ontologies and can adapt to changing terminology (Wang et al. 2018). However, machine learning methods are dependent on the availability of large, well-annotated training corpora (Basaldella et al. 2017). While these large corpora are available for gene recognition, gold standard corpora still need to be developed in ecological fields. In the absence of gold standard corpora, a 'silver standard' corpus can be created using existing text-mining tools and then used to train machine-learning algorithms (Kang et al. 2012). The CALBC Silver Standard Corpus, for example, was created using the harmonized results of 7 different text-mining pipelines, annotating genes, diseases, chemical entities, and species in over 100,000 abstracts (Rebholz-Schuhmann et al. 2011). The corpora developed over the course of this study represent the first steps toward the eventual creation of a machine-learning algorithm for recognizing ecological variables, and the annotations created using our pipeline can be used as training data toward this goal.

Improving precision through section identification and refinement of rules

The highest precision our pipeline was able to achieve was 0.36, a value that is quite low even when precision is not the main goal of the pipeline and which results in a large number of irrelevant annotations a user must sift through. Improvements to our pipeline to raise precision to at least 0.6–0.7 are needed to improve usability. A large proportion of the false positives that led to this relatively low precision score were sentences extracted from the methods section, and therefore the identification and exclusion of article sections is of primary importance in improving precision of the current pipeline.

This type of section identification is not possible with the current PDF to text conversion strategy used, which extracts pdf content into unstructured text files, similar to the format present in many existing full text repositories of scientific literature such as CiteSeerX (Tkaczyk et al. 2012). However, in recent years several PDF to XML converters have been developed in an attempt to preserve embedded text grouping such as Introduction and Methods sections (Bast and Korzen 2017). Several smaller datasets have been developed that provide articles in text format with hierarchal section structures retained, such as Grotoap2 (Tkaczyk et al. 2014) and ACL Anthology Reference Corpus (Bird et al. 2008). Given the large number of false positive sentences extracted from the methods section in our test, and the tendency of methods sentences to contain the same terminology and structure used in results, incorporation of these new XML conversion methods should significantly increase precision of our pipeline.

An additional type of false positive extracted using our pipeline was sentences taken from results that do not include all target information (i.e. specific genes and ecological variables within the sentence), due to the omission of specific gene names or treatment details (ex. statements referring to "at the 4 h mark" without specifying that samples were examined after temperature stress), raising the possibility that extracting text on either side of these sentences may provide the needed context for these to be useful. The sentences examined during this study represent only text where all three annotation types were found within the same sentence. When we expanded our sentence window to look at annotations found within 2 or 3 sentences, we found that precision was lowered without adding significant numbers of true positives (34 out of 393 sentences extracted only when a 3 sentence window was used). However, our pipeline could be modified to keep our 1 sentence window when matching annotations, but optionally report the sentence immediately before and after the extracted sentences. For now, these false positive sentences may be relevant to a user when considered alongside other annotations that provide the missing information.

Further refinement of rules developed in this study will also improve precision measures prior to public dissemination of our pipeline. In particular, the primary goal of this study was the development of tools to capture ecological variables and association-based "connector" terms in molecular ecology articles. We therefore expanded gene and protein annotation in our study to minimize the number of missed annotations due to failed gene annotation, as described in our methods by annotating any instance of the word "gene" or "transcript". Accordingly, removing rules matching any instance of "gene" or "transcript" will increase precision of the pipeline. In addition, more targeted uses of this pipeline, such as searching for a specific gene or gene family, will greatly reduce false positives due to gene NER annotation.

Remaining false positives have a variety of causes. For example, some papers contained instances where acronyms (often described only in methods) obscured the identity of an ecological variable (ex. "In addition, LL resulted in a damped rhythm in the qCry1 mRNA, owing mainly to the elevation of the trough values"). Other sentences contained relevant words but were speculatory (ex. "The enhanced tolerance to Cd in transgenic plants is possibly due to the regulatory properties of the OsMSR3 protein"), or described physiological functions of a gene (ex. “TT8 controls its own expression in a feedback regulation involving TTG1 and homologous MYB and bHLH factors, allowing a strong and cell-specific accumulation of flavonoids in Arabidopsis thaliana”). Finally, a small number of sentences were affected by the pdf to text conversion process, obscuring relevant information that may have been contained in them (ex. “rBeo3t.ryBtoitsryetlilsipetlilcipat-icaan- danBdoBtroytrtyistiscicninereereaa--iinnooccuullaatteeddssuusscecpeptibtilbelLe.Lfo.rfmoromlonogloinlignieliSninenSaipnan laliplyal lily showing disease progression at different time points”).

Supplemental Data

While the text-mining pipeline created here focuses on the extraction of natural language text from article bodies, we have determined that in order to access the majority of ecological association data present in the literature, an additional tool capable of parsing supplemental data files is required. Currently, most ecological annotation information is stored in the text of articles and in supplementary spreadsheets, since there is no searchable archive of ecological association data like there is for gene function data. Databases such as ArrayExpress store information about tissue, cell types, and disease states in which transcripts are found, while Gene Ontology is the largest of several ontologies aiming to consolidate existing information on molecular function, cellular location, and activities of gene products (Ashburner et al. 2000). Ecological studies typically only deposit raw sequence data into databases such as Sequence Read Archive, and store ecological association data in excel files in Dryad or as supplementary information. For example, of the 26 articles used in our training corpus, 11 (41%) had not archived data into any existing database. The most common methods of storing data were Dryad (37%), which is simply an archive of files, and BioProject (22%), which stores only raw sequence data. In addition, 15 of the 26 articles (56%) stored data on actual associated genes and loci in supplemental excel files and figures. There is therefore a large amount of information about possible ecological annotations of genes and loci that is stored in places not accessible to text-mining algorithms targeted at databases. The large amount of data available only in supplemental tables and spreadsheets highlights the need for a way to identify and extract gene-specific association data from spreadsheets recognized as having inconsistent formatting and nomenclature (Schneider et al. 2019). This ability would allow for the most thorough collection of data from published literature, especially as the amount of data within a typical molecular ecology study continues to grow.

The retrieval of supplementary files for text-mining is similar to the automated retrieval of fulltext articles themselves, another feature that will be implemented into our pipeline in the future. Currently, our text-mining pipeline runs on a corpus of previously downloaded articles. Automated processing of new articles from online databases is possible through APIs (Application Programming Interfaces) attached to some databases. These allow for the retrieval of information from the database through commands in a programming language instead of the graphical website interface. Extraction of new fulltext articles can easily be implemented at the same time as extraction of supplemental data if APIs such as Europe PubMed Central's RESTful Web Service (http://europepmc.org/restfulwebservice) are used. Supplementary files were successfully extracted using this API in a study by Kafkas et al. (2015), who found that the majority of database citations in supplementary files are not present in the fulltext of biomedical articles. Kafkas' study highlighted the importance of including supplementary files in text-mining efforts, as well as the lack of standardization and content guidelines that hinder the ability of text-mining protocols to accurately extract information. Nevertheless, extension of our current pipeline to allow for processing even of supplementary tables will greatly increase both the amount of data extracted and the usefulness of the pipeline.

Finally, a future goal assessed during this study was the possibility of using text-mining to annotate novel sequences of putative genes or functional regions that do not have an existing gene name associated with them. The end-product of such a tool would allow scientists to input a DNA sequence and obtain any ecological associations that have previously been found for similar sequences. Similar tools have been created for annotated genes, and the process of matching text entities from articles to database information (such as archived sequences) is known as grounding (Witte et al. 2007). Matching identification terms with sequences is more difficult when they are unannotated. While some studies simply don't report results for unannotated or uncharacterized sequences, others assign them study-specific IDs. Ideally, these study-specific IDs match their corresponding fastq sequence headers, such as those seen in Churcher et al. (2015). In this study, uncharacterized RNA transcripts are reported alongside annotated transcripts and given IDs in the format 'Aa_OE_[number]'. The sequence corresponding to each ID can be found in a fastq text file deposited in Dryad, where each sequence is labelled ‘Aa_OE_1’ to ‘Aa_OE_417669’. Identification of ecological annotations for these loci therefore requires a way to search for a sequence in fastq files, extract the corresponding header ID, and locate instances of that ID in text or supplementary files. The text-mining pipeline created in this paper can easily be adapted to the last step in this process, and sequences have successfully been extracted from supplementary files in the biomedical field (Haeussler et al. 2011), although this study excluded next-generation sequencing files. What remains is an efficient way to search these archived fastq files for unannotated sequences of interest.

Literature Cited

Aerts, S., M. Haeussler, S. van Vooren, O. L. Griffith, P. Hulpiau, S. J. M. Jones, S. B. Montgomery, and C. M. Bergman. 2008. Text-mining assisted regulatory annotation. Genome Biology 9:1–13.

Alvarez, M., A. W. Schrey, and C. L. Richards. 2015. Ten years of transcriptomics in wild populations: What have we learned about their ecology and evolution? Molecular Ecology 24:710–725.

Andrew, R. L., L. Bernatchez, A. Bonin, C. A. Buerkle, B. C. Carstens, B. C. Emerson, D. Garant, T. Giraud, N. C. Kane, S. M. Rogers, J. Slate, H. Smith, V. L. Sork, G. N. Stone, T. H. Vines, L. Waits, A. Widmer, and L. H. Rieseberg. 2013. A road map for molecular ecology. Molecular Ecology 22:2605–2626.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, and J. T. Eppig. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25:25–29.

Baird, D. J., C. J. O. Baker, R. B. Brua, M. Hajibabaei, K. McNicol, T. J. Pascoe, and D. de Zwart. 2011. Toward a knowledge infrastructure for traits-based ecological risk assessment. Integrated Environmental Assessment and Management 7:209–215.

Baker, C. J. O., R. Kanagasabai, W. T. Ang, A. Veeramani, H. S. Low, and M. R. Wenk. 2007. Towards ontology-driven navigation of the lipid bibliosphere. Asia Pacific Bioinformatics Network (APBioNet) 6th International Conference on Bioinformatics, InCoB 2007 - Proceedings 9:1–11.

Baker, C. J. O., and A. Riazanov. (n.d.). Aquatic Toxicity Test Information Extraction.

Basaldella, M., L. Furrer, C. Tasso, and F. Rinaldi. 2017. Entity recognition in the biomedical domain using a hybrid approach. Journal of Biomedical Semantics 8:1–14.

Bast, H., and C. Korzen. 2017. A Benchmark and Evaluation for Text Extraction from PDF. 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL):1–10.

Bird, S., R. Dale, B. J. Dorr, B. Gibson, M. T. Joseph, M. Y. Kan, D. Lee, B. Powley, D. R. Radev, and Y. F. Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008:1755–1759.

Buttigieg, P. L., E. Pafilis, S. E. Lewis, M. P. Schildhauer, R. L. Walls, and C. J. Mungall. 2016. The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation. Journal of Biomedical Semantics 7:1–12.

Caporaso, J. G., W. A. Baumgartner, D. A. Randolph, K. B. Cohen, and L. Hunter. 2007. MutationFinder: A high-performance system for extracting point mutation mentions from text. Bioinformatics 23:1862–1865.

Chepelev, L. L., A. Riazanov, A. Kouznetsov, H. S. Low, M. Dumontier, and C. J. O. Baker. 2011. Prototype semantic infrastructure for automated small molecule classification and annotation in lipidomics. BMC Bioinformatics 12.

Churcher, A. M., P. C. Hubbard, J. P. Marques, A. V. M. Can??rio, and M. Huertas. 2015. Deep sequencing of the olfactory epithelium reveals specific chemosensory receptors are expressed at sexual maturity in the European eel Anguilla anguilla. Molecular Ecology 24:822–834.

Cohen, A. M. 2005. Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. Proceedings of the acl-ismb workshop on linking biological literature, ontologies and databases: Mining biological semantics. 17–24.

Cohen, K. B., H. L. Johnson, K. Verspoor, C. Roeder, and L. E. Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC Bioinformatics 11.

Compson, Z. G., W. A. Monk, C. J. Curry, D. Gravel, A. Bush, C. J. O. Baker, M. S. Al Manir, A. Riazanov, M. Hajibabaei, S. Shokralla, J. F. Gibson, S. Stefani, M. T. G. Wright, and D. J. Baird. 2018. Linking DNA Metabarcoding and Text Mining to Create Network-Based Biomonitoring Tools: A Case Study on Boreal Wetland Macroinvertebrate Communities. Page Advances in Ecological Research. First edition. Elsevier Ltd.

Cunningham, H., D. Maynard, and V. Tablan. 2000. JAPE: a Java Annotation Patterns Engine (Second Edition). Page Technical Report CS--00--10.

Cunningham, H. 2002. GATE, a General Architecture for Text Engineering. Computers and the Humanities 36:223–254.

Degtyarenko, K., P. De matos, M. Ennis, J. Hastings, M. Zbinden, A. Mcnaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. 2008. ChEBI: A database and ontology for chemical entities of biological interest. Nucleic Acids Research 36:344–350.

Dumontier, M., C. J. O. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath, D. Klassen, J. P. McCusker, N. Queralt-

Rosinach, M. Samwald, N. Villanueva-Rosales, M. D. Wilkinson, and R. Hoehndorf. 2014. The semanticscience integrated ontology (SIO) for biomedical research and knowledge discovery. Journal of Biomedical Semantics 5:1–11.

Fleuren, W. W. M., and W. Alkema. 2015. Application of text mining in the biomedical domain. Methods 74:97–106.

FooLabs. 2014. Xpdf: A PDF Viewer for X.

Garnier, E., U. Stahl, M. A. Laporte, J. Kattge, I. Mougenot, I. Kühn, B. Laporte, B. Amiaud, F. S. Ahrestani, G. Bönisch, D. E. Bunker, J. H. C. Cornelissen, S. Díaz, B. J. Enquist, S. Gachet, P. Jaureguiberry, M. Kleyer, S. Lavorel, L. Maicher, N. Pérez-Harguindeguy, H. Poorter, M. Schildhauer, B. Shipley, C. Violle, E. Weiher, C. Wirth, I. J. Wright, and S. Klotz. 2017. Towards a thesaurus of plant characteristics: an ecological contribution. Journal of Ecology 105:298–309.

Garten, Y., and R. B. Altman. 2009. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC bioinformatics 10 Suppl 2:1–9.

Haeussler, M., M. Gerner, and C. M. Bergman. 2011. Annotating genes and genomes with DNA sequences extracted from biomedical articles. Bioinformatics 27:980–986.

Hardy, B., G. Apic, P. Carthew, D. Clark, D. Cook, I. Dix, S. Escher, J. Hastings, D. J. Heard, N. Jeliazkova, P. Judson, S. Matis-Mitchell, D. Mitic, G. Myatt, I. Shah, O. Spjuth, O. Tcheremenskaia, L. Toldo, D. Watson, A. White, and C. Yang. 2012. Toxicology ontology perspectives. Altex 29:139–156.

Harmston, N., W. Filsell, and M. P. H. Stumpf. 2010. What the papers say: Text mining for genomics and systems biology. Human Genomics 5:17–29.

Huang, C. C., and Z. Lu. 2016. Community challenges in biomedical text mining over 10 years: Success, failure and the future. Briefings in Bioinformatics 17:132–144.

Ijaz, A. Z., T. C. Jeffries, U. Z. Ijaz, K. Hamonts, and B. K. Singh. 2017. Extending SEQenv: A taxa-centric approach to environmental annotations of 16S rDNA sequences. PeerJ 2017:1–25.

Kafkas, Ş., J. H. Kim, X. Pi, and J. R. McEntyre. 2015. Database citation in supplementary data linked to Europe PubMed Central full text biomedical articles. Journal of Biomedical Semantics 6:1–7.

Kang, N., E. M. van Mulligen, and J. A. Kors. 2012. Training text chunkers on a silver standard corpus: Can silver replace gold? BMC Bioinformatics 13:0–5.

Kent, A., M. M. Berry, F. U. J. Luehrs, and J. W. Perry. 1955. Machine literature searching VIII. Operational criteria for designing information retrieval systems. American documentation 6:93–101.

Landry, C. R., and N. Aubin-Horth. 2007. Ecological anotation of genes and genomes through ecological genomics. Molecular Ecology 16:4419–4421.

McDonald, R., and F. Pereira. 2005. Identifying gene and protein mentions in text using conditional random fields. BMC Bioinformatics 6:1–7.

Millard, J. W., R. Freeman, and T. Newbold. 2020. Text-analysis reveals taxonomic and geographic disparities in animal pollination literature. Ecography 43:44–59.

Müller, H. M., E. E. Kenny, and P. W. Sternberg. 2004. Textpresso: An ontology-based information retrieval and extraction system for biological literature. PLoS Biology 2.

Naderi, N., T. Kappler, C. J. O. Baker, and R. Witte. 2011. Organismtagger: Detection, normalization and grounding of organism entities in biomedical documents. Bioinformatics 27:2721–2729.

Pavey, S. A., L. Bernatchez, N. Aubin-Horth, and C. R. Landry. 2012. What is needed for next-generation ecological and evolutionary genomics? Trends in Ecology & Evolution 27:673–678.

Peña-Castillo, L., and T. R. Hughes. 2007. Why are there still over 1000 uncharacterized yeast genes? Genetics 176:7–14.

Pérez-Pérez, M., G. Pérez-Rodríguez, A. Blanco-Míguez, F. Fdez-Riverola, A. Valencia, M. Krallinger, and A. Lourenço. 2019. Next generation community assessment of biomedical entity recognition web servers: Metrics, performance, interoperability aspects of BeCalm.

Rebholz-Schuhmann, D., A. J. Yepes, C. Li, S. Kafkas, I. Lewin, N. Kang, P. Corbett, D. Milward, E. Buyko, E. Beisswanger, K. Hornbostel, A. Kouznetsov, R. Witte, J. B. Laurila, C. J. O. Baker, C. J. Kuo, S. Clematide, F. Rinaldi, R. Farkas, G. Móra, K. Hara, L. I. Furlong, M. Rautschka, M. L. Neves, A. Pascual-Montano, Q. Wei, N. Collier, M. F. M. Chowdhury, A. Lavelli, R. Berlanga, R. Morante, V. Van Asch, W. Daelemans, J. L. Marina, E. van Mulligen, J. Kors, and U. Hahn. 2011. Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. Journal of Biomedical Semantics 2.

Rubach, M. N., R. Ashauer, D. B. Buchwalter, H. J. De Lange, M. Hamer, T. G. Preuss, K. Töpke, and S. J. Maund. 2011. Framework for traits-based assessment in ecotoxicology. Integrated Environmental Assessment and Management 7:172–186.

Sagayam, R., S. Srinivasan, and S. Roshni. 2012. A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques. International Journal Of Computational Engineering Research (ijceronline.com) 2:2250–3005.

Schneider, F. D., D. Fichtmueller, M. M. Gossner, A. Güntsch, M. Jochum, B. König-Ries, G. Le Provost, P. Manning, A. Ostrowski, C. Penone, and N. K. Simons. 2019. Towards an ecological trait-data standard. Methods in Ecology and Evolution 10:2006–2019.

Settles, B. 2005. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 21:3191–3192.

Sinclair, L., U. Z. Ijaz, L. J. Jensen, M. J. L. Coolen, C. Gubry-Rangin, A. Chroňáková, A. Oulas, C. Pavloudi, J. Schnetzer, A. Weimann, A. Ijaz, A. Eiler, C. Quince, and E. Pafilis. 2016. Seqenv: Linking sequences to environments through text mining. PeerJ 2016:1–17.

Spasic, I., S. Ananiadou, J. McNaught, and A. Kumar. 2005. Text mining and ontologies in biomedicine: Making sense of raw text. Briefings in Bioinformatics 6:239–251.

Tkaczyk, D., A. Czeczko, K. Rusek, Ł. Bolikowski, and R. Bogacewicz. 2012. GROTOAP: Ground truth for open access publications. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries:381–382.

Tkaczyk, D., P. Szostek, and L. Bolikowski. 2014. GROTOAP2 - The methodology of creating a large ground truth dataset of scientific articles. D-Lib Magazine 20:1–14.

Wang, X., C. Yang, and R. Guan. 2018. A comparative study for biomedical named entity recognition. International Journal of Machine Learning and Cybernetics 9:373–382.

Westergaard, D., H. H. Stærfeldt, C. Tønsberg, L. J. Jensen, and S. Brunak. 2018. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Computational Biology 14:1–16.

Wissler, L., M. Almashraee, M. Dagmar, and A. Paschke. 2014. The Gold Standard in Corpus Annotation. Journal of Political Economy 7:551.

Witte, R., and C. J. O. Baker. 2007. Towards a systematic evaluation of protein mutation extraction systems. Journal of Bioinformatics and Computational Biology 5:1339–1359.

Witte, R., T. Kappler, and C. J. O. Baker. 2007. Ontology Design for Biomedical Text Mining. Pages 281–313 Semantic Web. Springer, Boston, MA.

Zweigenbaum, P., D. Demner-fushman, H. Yu, and K. B. Cohen. 2007. Frontiers of biomedical text mining: Current progress. Briefings in Bioinformatics 8:358–375.

GENERAL DISCUSSION

The four studies conducted over the course of this PhD thesis span a wide range of topics at the heart of molecular ecology in the era of next generation sequencing. The first three chapters represent the first application of next-generation sequencing to study genetic characteristics of Striped Bass, including an examination of the performance of current genetic association tests on our reduced representation SNP library. I characterize the first cohort of juvenile Striped Bass seen in the Saint John River since the disappearance of the local population decades ago, establishing evidence of a genetically divergent spawning population of Striped Bass still present in the Saint John River. Chapter 2 includes the first published genetic characterization of a little-studied area of the Striped Bass range along the eastern coast of Nova Scotia as part of one of the most comprehensive genetic structure analyses of the Striped Bass species to date. Following decades of difficulty in identifying Roanoke River Striped Bass in coastal stocks, I demonstrate that SNP markers can successfully discriminate among Striped Bass from the three major origin regions of coastal migratory stocks of Striped Bass: Hudson River, Chesapeake Bay, and Roanoke River. The last chapter details the development of a preliminary set of tools enabling the automated extraction of associations between proteins and a wide range of ecological variables, the first text-mining pipeline to target associations between these two concepts.

Here I will summarize the key findings of each chapter of this thesis and discuss next steps in each area of study covered by my thesis. These will include several questions remaining regarding Striped Bass movements and distribution, discussion of projects already underway to help fill in knowledge gaps, and upcoming plans for further genetic characterization of the Striped Bass range (chapter 1 and 2). Next steps for genetic association mapping in Striped Bass include an examination of possible genetic associations with different migratory patterns within a single population of Striped Bass, and determining whether there is a genetic basis for sex determination within this species (chapter 3). I will also review key goals in future developments of my automated text-mining pipeline (chapter 4).

Genetic population structure of Striped Bass

Chapter 1 – Evidence of a genetically distinct population of Striped Bass (*Morone saxatilis*) within the Saint John River, New Brunswick, Canada

In my first chapter, I examine the genetic origin of 21 Striped Bass aged 1 to 3 collected in the Saint John River in 2014–2016, the first juvenile Striped Bass caught in the river since the 1970s. I also use tissue from individuals collected in Shubenacadie River (n = 21), Hudson River (n = 23), and the head of Chesapeake Bay (n = 19). I use a double digest restriction-site associated DNA (ddRAD) technique to find and genotype 4700 SNP markers in all samples and examined genetic divergence among groups and individuals using both population differentiation statistics (F_{ST}) and genetic clustering techniques. This represents the first time juvenile Striped Bass have been genetically characterized in the Saint John River, and the first SNP library that has been constructed for Striped Bass. I found that all four groups of samples were significantly different from each other genetically, with a global F_{ST} of 0.101 ($P < 0.001$). Shubenacadie River Striped Bass were most divergent (F_{ST} = 0.115–0.16), and Hudson River Striped Bass were genetically very similar to Chesapeake Bay Striped Bass (F_{ST} = 0.019). The Saint John River juveniles were much more divergent from all other populations (FST = 0.076–0.115) than Hudson River and Chesapeake Bay were from each other. Genetic clustering analyses grouped Striped Bass into three distinct genetic clusters: Shubenacadie River, both US populations combined, and the Saint John River juveniles. When individual assignment to these three clusters was examined, 15 Saint John River juveniles belonged entirely to the Saint John River genetic cluster, while the remaining 6 were admixed and showed ancestry to both the Saint John River cluster and either the US cluster or Shubenacadie River. The genetic divergence seen within the Saint John River juveniles presents strong evidence of a surviving local population of Striped Bass, while the admixture seen in a third of the juveniles provides the first evidence that migrant adults occasionally spawn in the river. Direct evidence of spawning remains to be collected; however it is unlikely these juvenile Striped Bass are from a different, previously unknown population.

Chapter 2 – Genomic Population Structure of Striped Bass (*Morone saxatilis*) from the Gulf of St. Lawrence to Cape Fear River

In chapter 2, I expand my population genetics study to include 477 individuals from 15 different locations along the Striped Bass range, from the Gulf of St. Lawrence to

Roanoke and Cape Fear River in North Carolina. I genotyped 1256 SNP markers that passed quality filtering on this expanded sample set. Individuals in this expanded sample set clustered into 7 distinct genetic clusters, representing the Gulf of St. Lawrence, the Shubenacadie River, the Saint John River, the Kennebec River and Hudson River, the Delaware River and Chesapeake Bay system, and North Carolina rivers. The 3 Canadian clusters were markedly more divergent from each other and from all US clusters, while all 3 US genetic clusters were relatively similar to each other. When these 6 genetic clusters were used as reporting groups for mixed stock analysis, we were able to assign 99% of individuals to one of the 6 groups. An expanded sample set of 32 Saint John River juveniles further confirmed results from chapter 1 and were shown to be genetically divergent from Gulf of St. Lawrence, Kennebec River, and Roanoke River Striped Bass, all unlikely but possible genetic sources for the juveniles found in the river. In addition, we found that the majority of Striped Bass collected in Mira River on the east coast of Nova Scotia were matched the genotype of Gulf of St. Lawrence bass, but a small number were of US origin. Finally, we detected a small number of significant differences among rivers within the Chesapeake Bay, with very low levels of differentiation, possibly reflecting non-uniform migration patterns among tributaries within the bay.

This study demonstrates for the first time that genotyping-by-sequencing allows for highly accurate mixed stock analysis of Striped Bass along the Atlantic Coast. The ability to accurately discriminate between Roanoke River and Chesapeake Bay Striped Bass in particular is crucial in light of the apparent increase in migratory bass from Roanoke

River in the last 30 years (Callihan et al. 2014), in order to measure the extent to which Roanoke River now contribute to different coastal stocks of Striped Bass. Accurate mixed stock analysis is also important for tracking shifts in distribution as a result of climate change, particularly if these shifts result in increased mixing of previously isolated spawning populations.

Future Research Focus – Population connectivity of Striped Bass

The survival and persistence of the native population of Striped Bass in the Saint John River has been a contentious topic over the past several decades, due in part to the failure of surveys to find evidence of eggs or larvae within the Saint John River (Andrews et al. 2017) and the lack of high-quality tissue samples of individuals taken before the population's collapse and disappearance. Because I cannot directly compare the genome of these juveniles with the genome of the historical native population within the Saint John River, I explored the likelihood of alternative explanations for the origin of these juveniles. These include the possibility that these juveniles represent migratory fish from an unknown and unsampled population. To test this hypothesis, I compared their genome with that of the three populations known to contribute migrants to the Saint John River: Shubenacadie River, Hudson River, and Chesapeake Bay. In chapter 2, we also compared them to bass from the Gulf of St. Lawrence, the recently re-established population in the Kennebec River, and Roanoke River bass, which have become more migratory in the last 10 years (Callihan et al. 2014). None of these populations displayed a genetic signature similar to that seen in Saint John River juveniles. I also explored the possibility that these juveniles could be the result of a founder effect when the river was recolonized by migrant bass. When populations are founded by a very small number of breeders and are isolated from gene flow, this can result in very low genetic diversity and rapid divergence of the new population from the population the founders originated from, in what is known as a founder effect (Templeton 1980). To determine what type of genetic pattern would be seen in the SJR population if such a founder effect occurred in a long-lived species that had been recolonized only a couple of generations ago, I examined previous studies that had been performed on known founder events that have occurred in the last five to fifty generations. In some studies, no genetic difference was observed between the new population and the origin population of breeders (Clegg et al. 2002, Malaney et al. 2018). In others, differences found were relatively small (F_{ST} = 0.02–0.05; Eales et al. 2008; Hawley et al. 2006), comparable to that seen between Striped Bass from Hudson River and Chesapeake Bay. I found no evidence that such a recent founder effect could produce the deep divergence seen between the Saint John River genetic cluster and all other Striped Bass populations, particularly considering that the presence of admixed individuals means that this population is not isolated from its neighbors.

In the last twenty years, three different sets of genetic markers have been used to examine SJR Striped Bass. Bentzen and Paterson (2008) were the first to detect a divergent genetic group of Striped Bass in the SJR using 11 microsatellites, comprised of large, old adults thought to have been spawned in the last 1990s. None of these individuals were available for use in my SNP studies, but the genetic cluster found by my SNP markers in both juvenile and adult Striped Bass caught within the river bears many similarities. In contrast, the set of microsatellites used by Wirgin et al. (2020) found that all Saint John

River individuals analyzed appeared admixed between Shubenacadie River and US genetic clusters, and F_{ST} values were much lower though still significant, implying a local spawning population descended largely from migrants. A direct comparison of these three datasets is difficult without the presence of reference samples used in multiple datasets or, in the case of microsatellites, shared loci between datasets. Is the genetic cluster detected in my thesis the same as the cluster found by Bentzen and Paterson in 2008? Is there a difference in the Striped Bass samples used in each study? Are the markers used in Bentzen and Paterson's study and my thesis detecting genetic differences present in only some areas of the genome, that are not represented in Wirgin et al.'s dataset? While microsatellites can contain a great deal of information per locus, the small number of markers in a dataset represent a much smaller section of a genome than a large SNP dataset. To begin to address these questions, we provided several tissue samples representing both pure Saint John River bass and admixed bass found within the Saint John River to Dr. Bentzen, to be included in an upcoming, larger genetic survey of Striped Bass within the Bay of Fundy. The inclusion of samples that have been genotyped both by my SNP dataset and Bentzen and Paterson's microsatellites will shed light on whether the recent juveniles present in the Saint John River resemble the older 1990s cohort of Striped Bass discovered over a decade ago. Inclusion of the Annapolis River individuals from Wirgin et al. (2020) in this dataset could also shed further light on the genetic origins of Striped Bass found in this river, as well as a comparison of Saint John River samples used in each study to determine whether any overlap exists.

The study in chapter 2 is the first documented detection of US Striped Bass on the east coast of Nova Scotia (n = 3), but unpublished microsatellite genetic data collected by the Eskasoni Fish and Wildlife Commission have also discovered individuals of apparent US origin recently (P. Bentzen, pers. comm.), corroborating our findings. These genetic detections stand in contrast to numerous mark-recapture and acoustic telemetry studies on Striped Bass that have largely failed to find evidence of movement between the Gulf of St. Lawrence and the southern Striped Bass range (Hogans and Melvin 1984, Waldman et al. 1990, Rulifson and Dadswell 1995, Richards and Rago 1999, Douglas et al. 2003, DFO 2010, Pautzke et al. 2010, Broome 2014, Callihan et al. 2015, Gahagan et al. 2015, Andrews et al. 2018). There is increasing interest in better characterizing the movements of Striped Bass in this little-studied part of their native range (Andrews et al. 2019b, 2019a), particularly following telemetry evidence of freshwater migrations during spawning season in Mira River (Andrews et al. 2019b). Further studies are required to determine whether these detections point to an expanded migratory range of US bass, possibly due to a northward range shift in response to climate change, or rare migrants that remain and spawn on the Nova Scotia coast.

The low genetic divergence among US Striped Bass in this and previous genetic studies (Brown et al. 2005, Gauthier et al. 2013, Wirgin et al. 2020) points to a greater amount of gene flow among regions in the central part of the Striped Bass range. Additionally, the failure of a panel of >1,000 SNPs to consistently differentiate between Striped Bass from different tributaries within the Chesapeake Bay and the apparent genetic homogeny between upper Chesapeake Bay and Delaware River indicate that spawning populations within this area are highly connected. Recent acoustic telemetry work on Chesapeake Bay and Hudson River Striped Bass corroborates these findings. In both studies, a small number of Striped Bass were documented as straying between Hudson River and Chesapeake Bay (Secor et al. 2020a, 2020b). Additionally, over half of Potomac River Striped Bass less than 10 years of age were recorded in other tributaries within Chesapeake Bay during the spawning period, and 10-27% of tagged individuals were observed exiting the bay via the Chesapeake-Delaware canal that connects the upper Chesapeake Bay with the Delaware River (Secor et al. 2020a). This follows previous observations of adult Striped Bass using the Chesapeake-Delaware canal during spawning season (Kneebone et al., 2014; Koo and Wilson, 1972; Nichols and Miller, 1967). Given these numbers and the apparent genetic homogeny of spawning populations within the Delaware River and Chesapeake Bay, this region seems to function as a single large metapopulation rather than separate populations.

While the population genetic analysis presented in chapter 2 is one of the most comprehensive in literature in terms of breadth of the range covered, it does not include every known spawning population in the Striped Bass range. In North Carolina, the Neuse River supports a small population of Striped Bass that is thought to have natural spawning (Morris et al. 2003). Both the Cape Fear River and the Neuse River were heavily stocked with Roanoke River bass until 2010, and evidence of a surviving, genetically distinct group is inconsistent and sparse (Anderson et al. 2014). South of the Cape Fear River, resident spawning populations of Striped Bass are known to reside in the following watersheds: the Pee Dee River, the Santee-Cooper river system and the connected rivers Congaree and Wateree, the combined Combahee, Ashepoo, and Edisto rivers (the ACE Basin), and the Savannah River (Anderson et al. 2014). Like North Carolina, these rivers have a long history of within-state stocking as bass from the Santee-Cooper system were transferred to the other watersheds until 1997 (Bulak et al. 2004). A recent microsatellite study found that genetic divergence among South Carolina watersheds has increased following the cessation of stocking (Anderson et al. 2014), and so a comprehensive genetic structure should include samples from all watersheds. We have secured 18 tissue samples collected from Pee Dee River in South Carolina by the South Carolina Department of Natural Resources in coordination with Ben Gahagan. We aim to include all major watersheds in South Carolina to our dataset, allowing for both a comprehensive coverage of the native range and comparisons with populations that have been stocked with Santee-Cooper Striped Bass.

Finally, a complete genetic profile of the Striped Bass native range includes a survey of the current status of populations within the Gulf of Mexico. The status of Striped Bass in this area is much more tenuous than on the eastern coast. Small historical populations are thought to have extended to the Mississippi River (Wooley and Crateau 1983), but most such populations are considered extirpated today. There remains a spawning population of Striped Bass in the ACF river system, comprising the Apalachicola, Chattahoochee, and Flint rivers. While most of the native spawning populations of Striped Bass along the Texas coast have been extirpated, there exists a small spawning population in Trinity River just below Lake Livingston Dam, Texas (Kurzawski and Maddux 1991, Smith and Buckmeier 2016). Located on the extreme southern end of Striped Bass' temperature tolerance, the population is thought to depend on a regular influx of escapees from Lake Livingston (Smith and Buckmeier 2016), which was stocked with bass largely from Santee-Cooper and Kerr Reservoir from 1967-1985 (Fries et al. 2004). We currently have 92 samples from individuals collected from the Trinity River for use in future studies.

Association studies in Striped Bass

Chapter 3 – Comparing mixed models and Random Forest association tests using naturalGWAS and a Striped Bass SNP dataset

In chapter 3, I focused on a dataset containing 7319 SNPs and 171 Striped Bass from three regions that had length and weight measurements available, for which I calculated condition factor of each bass. The R package *naturalGWAS* to test the performance of four current methods of detecting genotype-phenotype associations. We investigated the effect of level of polygeny on power and false positives, as well as the impact of phenotype values correlating with population stratification. Random Forest displayed low power and high false positives even at moderate population structure correlation. In contrast, Zhao's Random Forest method had remarkably low levels of false positives, and comparable power to confounder adjusted multiple testing, the best performing of the mixed model methods. When CATE and Zhao's Random Forest were used to analyze real condition factor data, both tests found no associated loci. While specific to my Striped Bass dataset, this study is nevertheless the first time Zhao's method of correcting for population structure in Random Forests has been evaluated for use on polygenic phenotypes. It also represents the third account of this test's performance in general, and demonstrates that both power and false positives are comparable to or better than recent mixed model methods of detecting association.

Future Research Focus – Association studies

Association studies in molecular ecology aim to find genetic loci that influence or cause various phenotypic traits, such as physiology, appearance, or behavior. While great strides have been made in our ability to detect connections between phenotypes that are controlled primarily by one or two loci, many phenotypes are influenced by dozens or even hundreds of genetic loci, each one having a relatively small magnitude of effect, and these loci are much more difficult to detect (Rockman 2012). In recent years, Random Forest analyses have gained attention as a means of addressing this challenge, as the algorithm is capable of assessing the performance of a locus in conjunction with other loci (Brieuc et al. 2018, Forester et al. 2018). In contrast to earlier studies that found Random Forest performed well in detecting loci associated with disease (Goldstein et al. 2010), my simulations in this chapter found that neither Random Forest analysis had higher power to detect genetic loci exhibiting only small effects on a phenotype, and largely detected the same number of causal loci as the mixed model analyses. This agrees with another recent study that looked at Random Forest's performance finding gene-environment associations. In this study, as with ours, Random Forest detected only one or two genetic loci that had the strongest correlations with the environment (Forester et al. 2018). When Zhao's method of correcting for population structure was applied, we did see drastically lower numbers of false positive loci compared to simulations run with uncorrected Random Forest and to false positives seen in Forester et al. (2018), but overall power was comparable to mixed model analyses.

While I began the search for phenotype-genotype associations using condition factor as the most readily available phenotype information available within my samples, I have also been involved in tracking phenotype information on migratory behavior of this species. There is increasing evidence of distinct groups within a Striped Bass population in terms of migratory behavior (Secor et al. 2001, Gahagan et al. 2015, Andrews et al. 2017); similar migratory contingents have been shown to have a genetic basis in species such as Rainbow Trout (*Oncorhynchus mykiss*; Hecht et al. 2013). While the majority of Striped Bass appear to show differential migration, transitioning to a migratory lifestyle by 10 years of age, a substantial number of bass appear to stay within their natal river their entire life (Secor and Piccoli 1996, Zlokovitz et al. 2003, Secor et al. 2020a). Currently, the Pavey lab has 321 samples from acoustically tagged Striped Bass collected in the Saint John River (n = 77) and the coast of Massechussetts (n = 244), the latter of which originate in Hudson River and Chesapeake Bay. In order to properly assess genetic differences between migratory and resident Striped Bass, we need tissue from confirmed resident Striped Bass from the same populations as confirmed ocean migrants. Efforts to acquire funding for directed sampling of resident Hudson River Striped Bass are underway.

Finally, during my PhD program I began work on determining whether there is a genetic component to Striped Bass sex determination. Sex determination in Striped Bass, like in many fish species, is also unknown, although the closely related European Sea Bass is thought to make use of a polygenic method of sex determination, with environmental influence (Vandeputte et al. 2007). DNA has been isolated and sequenced for 124 male and 79 female Striped Bass, collected from Shubenacadie River, the Atlantic coast, Cape Fear River, Roanoke River, Trinity River, and Florida broodstocks. Initial association tests regarding sex determination with these samples have revealed no obvious associations with sex, and examination of heterozygosities and genotyping rates have revealed no sex-specific loci (Appendix H). However, while almost 200 individuals have been sequenced thus far, within-population sample sizes are relatively small, particularly after sample sizes are adjusted to balance sex ratios. Further sampling is required before a comprehensive report can be published on evidence for genetic sex determination in Striped Bass.

Automated text-mining in molecular ecology

Chapter 4 – A new automated text-mining pipeline for ecological associations

In chapter 4, I created a total of 6 rule sets and 3 gazetteers designed to recognize a wide range of ecological variables and terms that denote an associative relationship between those variables and something else. I combined these novel tools with existing named entity recognition tools to annotate protein names within text and created a pipeline that could extract associations between proteins and ecological variables. I tested this pipeline against a manually annotated corpus of 104 molecular ecology articles and successfully annotated 88% of all target manually annotated sentences. Of the 6369 sentences extracted by the automated pipeline, 36% contained useful information about protein-ecological associations. The tools created for this pipeline are the first text-mining resources able to annotate ecological associations in literature and will be integral to the creation of a user-friendly resource allowing scientists to quickly and efficiently find ecologically relevant associations with a protein of interest or to identify potential candidate genes for ecologically important traits.

Future Research Focus – Text-mining

The initial pipeline I created in chapter 4 is complete in the sense that it is able to move from pdf article to the final step of extracting relevant sentences, but it is not quite ready to be used by the average scientist. In its current state, my pipeline can be used by anyone with sufficient knowledge of its programming language to make minor changes to the underlying code required to incorporate specific search terms (i.e. searching for a specific gene). For example, in order to restrict my search to the ‘Per3’ gene, I constructed a gazetteer with a list of gene synonyms, and then I modified the pipeline to load this gazetteer instead of Abner and PennBio gene taggers and turned off filtering rules that look specifically for Abner and PennBio annotations for quality filtering. The most immediate next step, therefore, is to incorporate “command line parameters” into the code. These give a user control over the types of input used when the pipeline runs without having to modify any code. For example, a user would be able to supply a list of

‘Per3’ synonyms as an input file to the command-line program, and my pipeline will be able to recognize that this input has been provided, automatically flag the appropriate sections of code to load the input file as a gazetteer, and ignore code used for more general gene annotation by Abner and PennBio taggers. This process is fairly straightforward to implement, but will require some time to test and streamline.

Additionally, general use of this pipeline will rely on the ability to run it on a large and up to date corpus of articles. Therefore, the second key aspect needed to make this pipeline accessible to scientists for general use is the ability to retrieve articles outside of my pre-selected corpora. This could be accomplished in a number of ways. Many centralized article search engines, such as PubMed Central or Web of Science, possess an Application Programming Interface (API) that allows third party programs to access documents and metadata stored in the search engine. The Europe PMC RESTful API gives programs access to full text open source articles in XML format as well as any attached supplemental data files. Alternatively, we could create our own searchable database of ecological peer-reviewed literature and host it on UNB’s ACENet.

In chapter 4, I describe four areas of possible improvement to the text-mining algorithm itself, with key improvements involving a reduction in the number of false positive sentences extracted by my pipeline, thereby improving the precision. Implementation of an API that allows for the automatic download of articles in XML format offers an additional opportunity for filtering out unwanted extracted sentences. Such XML text files typically contain section identifiers for standard sections in scientific articles such as methods and results. By adding an additional filtering step in our pipeline that identifies and excludes text found in the introduction and methods section of articles, I could eliminate a large proportion of false positive sentences and thereby improve both the precision and efficiency of my pipeline.

Ultimately, the Pavey Lab plans to construct an Ecological Association Ontology (Spasic et al. 2005), a network of concepts connected to each other with defined relationships, that will stand in parallel to more physiology-focused ontologies such as the widely used Gene Ontology. This is no small feat: The Gene Ontology Consortium includes 36 different groups and organizations and over 300 individual contributors, and is managed by a dozen principal investigators and project leads. However, once created the existence of such an ontology will feed into the text-mining pipeline I have already created, further expanding and adding flexibility to the specific ecological variables able to be searched for. An Ecological-based ontology will, crucially, organize ecologically important variables that are of interest in association studies into connected, overlapping categories. This will make it easier both to encompass a wide breadth of ecological variables in a text-mining search, and allow users to search for broad categories of ecological traits that will automatically include terms for traits within that category.

Conclusion

The results of this thesis present a variety of new tools for the management of Striped Bass along the Atlantic coast of North America and for fisheries ecology in the age of next-generation sequencing. I report genetic information on the connectivity of Striped Bass populations through the majority of their range, including the first genetic analysis of juvenile Striped Bass in the Saint John River and of the little-known aggregations of Striped Bass located along the eastern coast of Nova Scotia. The SNP markers generated over the course of this thesis, as well as the library preparation protocol I helped to refine, pave the way for the creation of a powerful genetic tool to identify the source population of mixed stock Striped Bass, and to fill knowledge gaps surrounding genotype associations with characteristics and behavior in Striped Bass. Additionally, I present the first steps in the creation of the first automated text-mining pipeline targeted at ecologically important genomic associations and ecological annotation of non-model organisms.

Literature Cited

Anderson, A. P., M. R. Denson, and T. L. Darden. 2014. Genetic Structure of Striped Bass in the southeastern United States and effects from stock enhancement. North American Journal of Fisheries Management 34:653–667.

Andrews, S. N., T. Linnansaari, R. A. Curry, and M. J. Dadswell. 2017. The Misunderstood Striped Bass of the Saint John River, New Brunswick: Past, Present, and Future. North American Journal of Fisheries Management 37:235–254.

Andrews, S. N., B. Wallace, M. Gautreau, T. Linnansaari, and R. A. Curry. 2018. Seasonal movements of Striped Bass *Morone saxatilis* in a large tidal and hydropower regulated river. Environmental Biology of Fishes 101:1549–1558.

Andrews, S. N., C. F. Buhariwalla, B. Fleet-Pardy, M. J. Dadswell, T. Linnansaari, and R. A. Curry. 2019a. Left out in the cold: the understudied overwintering ecology of Striped Bass in Canada. Environmental Biology of Fishes 102:499–518.

Andrews, S. N., M. J. Dadswell, C. F. Buhariwalla, T. Linnansaari, and R. A. Curry. 2019b. Looking for Striped Bass in Atlantic Canada: The Reconciliation of Local, Scientific, and Historical Knowledge. Northeastern Naturalist 26:1–30.

Bentzen, P., and I. G. Paterson. 2008. Report: genetic analysis of Striped Bass collected by Kingsclear First Nation in the Saint John River, New Brunswick. Report to the Department of Fisheries and Oceans, Dartmouth, Nova Scotia. p. 1-22.

Brieuc, M. S. O., C. D. Waters, D. P. Drinan, and K. A. Naish. 2018. A practical introduction to Random Forest for genetic association studies in ecology and evolution. Molecular Ecology Resources 18:755–766.

Broome, J. E. 2014. Population characteristics of Striped Bass (*Morone saxatilis*

Walbaum, 1792) in Minas Basin and patterns of acoustically detected movements within Minas Passage (Master's thesis). Acadia University, Wolfville, NS, Canada.

Brown, K. M., G. A. Baltazar, and M. B. Hamilton. 2005. Reconciling nuclear microsatellite and mitochondrial marker estimates of population structure: breeding population structure of Chesapeake Bay Striped Bass (*Morone saxatilis*). Heredity 94:606–15.

Bulak, J. S., C. S. Thomason, K. Han, and B. Ely. 2004. Genetic Variation and Management of Striped Bass Populations in the Coastal Rivers of South Carolina. North American Journal of Fisheries Management 24:1322–1329.

Callihan, J. L., C. H. Godwin, and J. A. Buckel. 2014. Effect of demography on spatial distribution: Movement patterns of the Albemarle Sound-Roanoke River stock of Striped Bass (*Morone saxatilis*) in relation to their recovery. Fishery Bulletin 112:131–143.

Callihan, J. L., J. E. Harris, and J. E. Hightower. 2015. Coastal migration and homing of Roanoke River Striped Bass. Marine and Coastal Fisheries 7:301–315.

Clegg, S. M., S. M. Degnan, J. Kikkawa, C. Moritz, A. Estoup, and I. P. F. Owens. 2002. Genetic consequences of sequential founder events by an island-colonizing bird. Proceedings of the National Academy of Sciences 99:8127–8132.

DFO. 2010. Assessment of the habitat quality and habitat use by the Striped Bass population of the St. Lawrence Estuary, Quebec. DFO Canadian Science Advisory Secretariat. Science Advisory Report. 2010/069.

Douglas, S. G., R. G. Bradford, and G. Chaput. 2003. Assessment of Striped Bass (*Morone saxatilis*) in the Maritime provinces in the context of species at risk.

Fisheries and Oceans Canada, Center for Science Advice (CSA), Gulf Region, Dartmouth, NS, Canada.

Eales, J., R. S. Thorpe, and A. Malhotra. 2008. Weak founder effect signal in a recent introduction of Caribbean Anolis. Molecular Ecology 17:1416–1426.

Forester, B. R., J. R. Lasky, H. H. Wagner, and D. L. Urban. 2018. Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. Molecular Ecology 27:2215–2233.

Fries, L. T., J. N. Fries, B. T. Hysmith, and J. S. Bulak. 2004. Analysis of Fluctuating Asymmetry In Three Populations Of Striped Bass.

Gahagan, B., D. Fox, and D. Secor. 2015. Partial migration of Striped Bass: Revisiting the contingent hypothesis. Marine Ecology Progress Series 525:185–197.

Gauthier, D. T., C. A. Audemard, J. E. L. Carlsson, T. L. Darden, M. R. Denson, K. S. Reece, and J. Carlsson. 2013. Genetic population structure of US Atlantic Coastal Striped Bass (*Morone saxatilis*). Journal of Heredity 104:510–520.

Goldstein, B. A., A. E. Hubbard, A. Cutler, and L. F. Barcellos. 2010. An application of Random Forests to a genome-wide association dataset: Methodological considerations and new findings. BMC Genetics 11.

Hawley, D. M., D. Hanley, A. A. Dhondt, and I. J. Lovette. 2006. Molecular evidence for a founder effect in invasive house finch (*Carpodacus mexicanus*) populations experiencing an emergent disease epidemic. Molecular Ecology 15:263–275.

Hecht, B. C., N. R. Campbell, D. E. Holecek, and S. R. Narum. 2013. Genome-wide association reveals genetic basis for the propensity to migrate in wild populations of rainbow and steelhead trout. Molecular Ecology 22:3061–3076.

Hogans, W., and G. Melvin. 1984. Kouchibouguac National Park Striped Bass (*Morone saxatilis*) fishery survey, New Brunswick. Aquatic Industries Limited. St Andrew's.

Kneebone, J., W. S. Hoffman, M. J. Dean, D. A. Fox, and M. P. Armstrong. 2014. Movement patterns and stock composition of adult Striped Bass tagged in Massachusetts coastal waters. Transactions of the American Fisheries Society 143:1115–1129.

Koo, T. S. Y., and J. S. Wilson. 1972. Sonic tracking Striped Bass in the Chesapeake and Delaware Canal. Transactions of the American Fisheries Society 101:453–462.

Kurzawski, K. F., and H. R. Maddux. 1991. Striped Bass Spawning in the Lower Trinity River, Texas. Page Fisheries Management Data Series.

Malaney, J. L., C. W. Lackey, J. P. Beckmann, and M. D. Matocq. 2018. Natural rewilding of the Great Basin: Genetic consequences of recolonization by black bears (*Ursus americanus*). Diversity and Distributions 24:168–178.

Morris, J. A., R. A. Rulifson, and L. H. Toburen. 2003. Life history strategies of Striped Bass, *Morone saxatilis,* populations inferred from otolith microchemistry. Fisheries Research 62:53–63.

Nichols, P. R., and R. V. Miller. 1967. Seasonal movements of Striped Bass, Roccus saxatilis (Walbaum), tagged and released in the Potomac River, Maryland, 1959-61. Chesapeake Science 8:102.

Pautzke, S. M., M. E. Mather, J. T. Finn, L. A. Deegan, and R. M. Muth. 2010. Seasonal use of a New England estuary by foraging contingents of migratory Striped Bass. Transactions of the American Fisheries Society 139:257–269.

Richards, R. A., and P. J. Rago. 1999. A case history of effective fishery management:

Chesapeake Bay Striped Bass. North American Journal of Fisheries Management 19:356–375.

Rockman, M. V. 2012. The QTN program and the alleles that matter for evolution: All that’s gold does not glitter. Evolution 66:1–17.

Rulifson, R. A., and M. J. Dadswell. 1995. Life history and population characteristics of Striped Bass in Atlantic Canada. Transactions of the American Fisheries Society 124:477–507.

Secor, D. H., and P. M. P. Piccoli. 1996. Age- and sex-dependent migrations of Striped Bass in the Hudson River as determined by chemical microanalysis of otoliths. Estuaries 19:778–793.

Secor, D. H., J. R. Rooker, E. Zlokovitz, and V. S. Zdanowicz. 2001. Identification of riverine, estuarine, and coastal contingents of Hudson River Striped Bass based upon otolith elemental fingerprints. Marine Ecology Progress Series 211:245–253.

Secor, D. H., M. H. P. O’Brien, B. I. Gahagan, J. Carter Watterson, and D. A. Fox. 2020a. Differential migration in Chesapeake Bay Striped Bass. PLoS ONE 15:1–19.

Secor, D. H., M. H. P. O’Brien, B. I. Gahagan, D. A. Fox, A. L. Higgs, and J. E. Best. 2020b. Multiple spawning run contingents and population consequences in migratory Striped Bass *Morone saxatilis*. PLoS ONE 15.

Smith, N. G., and D. L. Buckmeier. 2016. Living on the edge: persistence of a fringe Striped Bass population. Journal of the Southeastern Association of Fish and Wildlife Agencies 3:50–56.

Spasic, I., S. Ananiadou, J. McNaught, and A. Kumar. 2005. Text mining and ontologies in biomedicine: Making sense of raw text. Briefings in Bioinformatics 6:239–251.

Templeton, A. R. 1980. The theory of speciation via the founder principle. Genetics 94:1011–1038.

Vandeputte, M., M. Dupont-Nivet, H. Chavanne, and B. Chatain. 2007. A polygenic hypothesis for sex determination in the European Sea Bass *Dicentrarchus labrax*. Genetics 176:1049–1057.

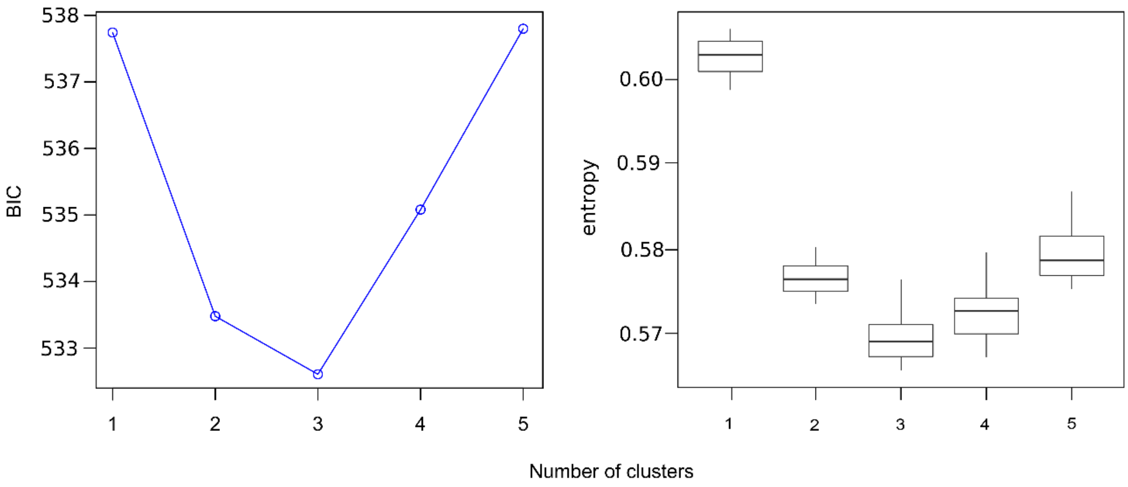
Waldman, J. R., D. J. Dunning, Q. E. Ross, and M. T. Mattson. 1990. Range dynamics of Hudson River Striped Bass along the Atlantic Coast. Transactions of the American Fisheries Society 119:910–919.

Wirgin, I., L. Maceda, M. Tozer, J. Stabile, and J. Waldman. 2020. Atlantic coastwide population structure of Striped Bass *Morone saxatilis* using microsatellite DNA analysis. Fisheries Research 226:105506.

Wooley, C. M., and E. J. Crateau. 1983. Biology, population estimates, and movement of, native and introduced Striped Bass, Apalachicola River, Florida. North American Journal of Fisheries Management 3:383–394.

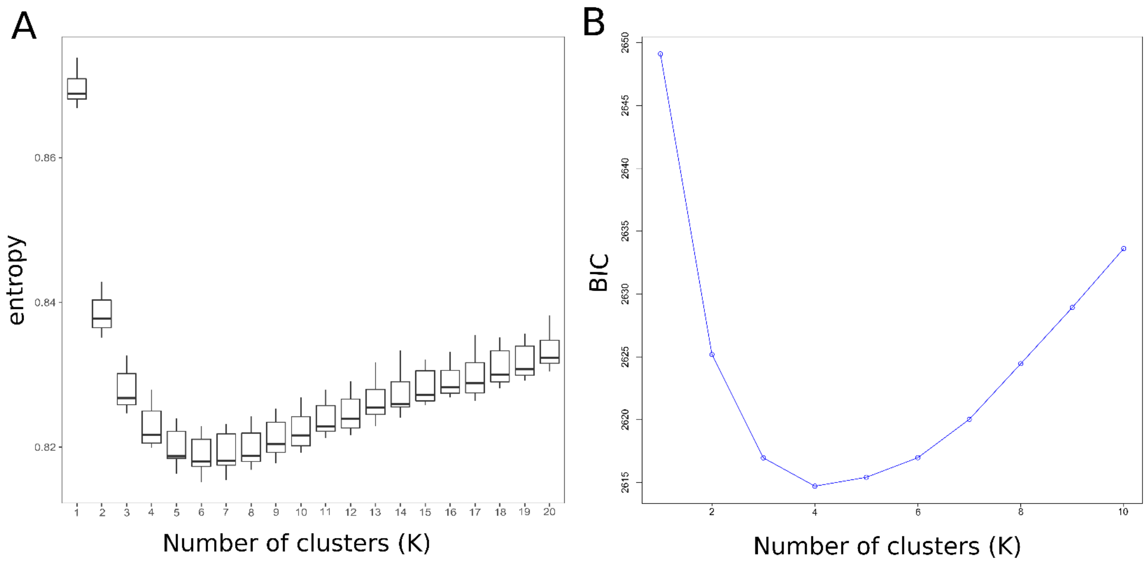
Zlokovitz, E. R., D. H. Secor, and P. M. Piccoli. 2003. Patterns of migration in Hudson River Striped Bass as determined by otolith microchemistry. Fisheries Research 63:245–259.

Appendix A: Entropy Values from Chapter 1



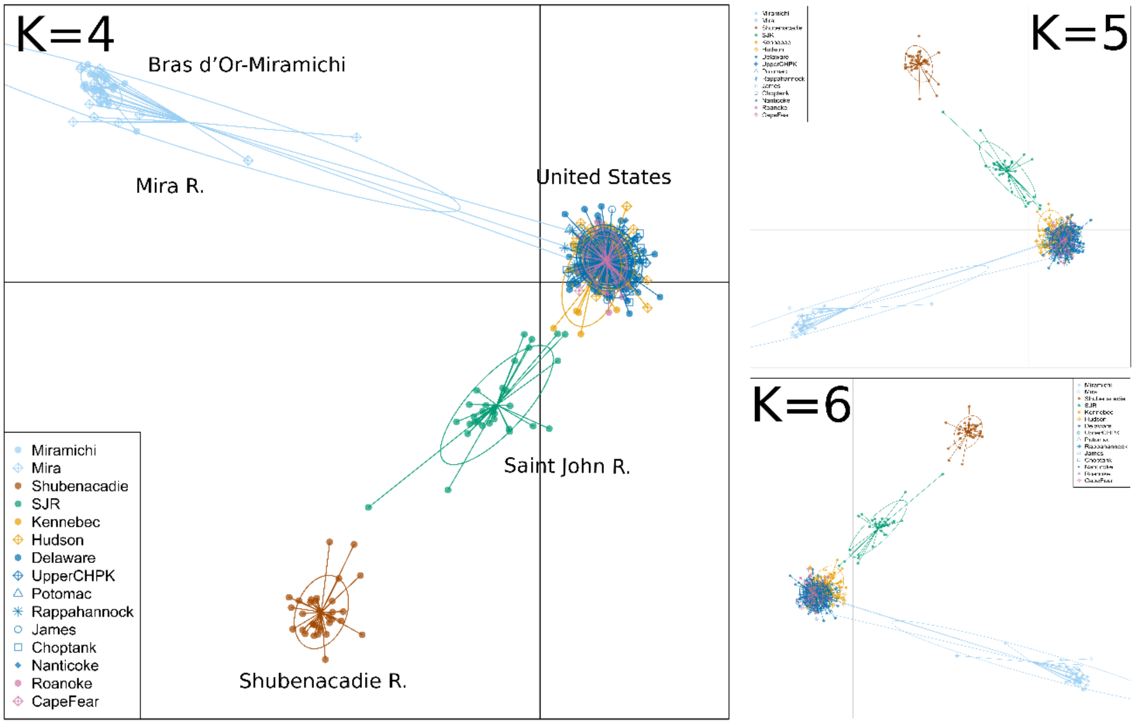
A) Bayesian Information Criterion values for most likely number of groups using DAPC. B) Entropy values calculated using the cross-entropy criterion for genetic clusters inferred from Striped Bass, *Morone saxatilis*, populations collected in four sites (Chesapeake Bay, Hudson River, Shubenacadie River, and Saint John River), using 4,700 SNP loci.

Appendix B: Entropy values from Chapter 2



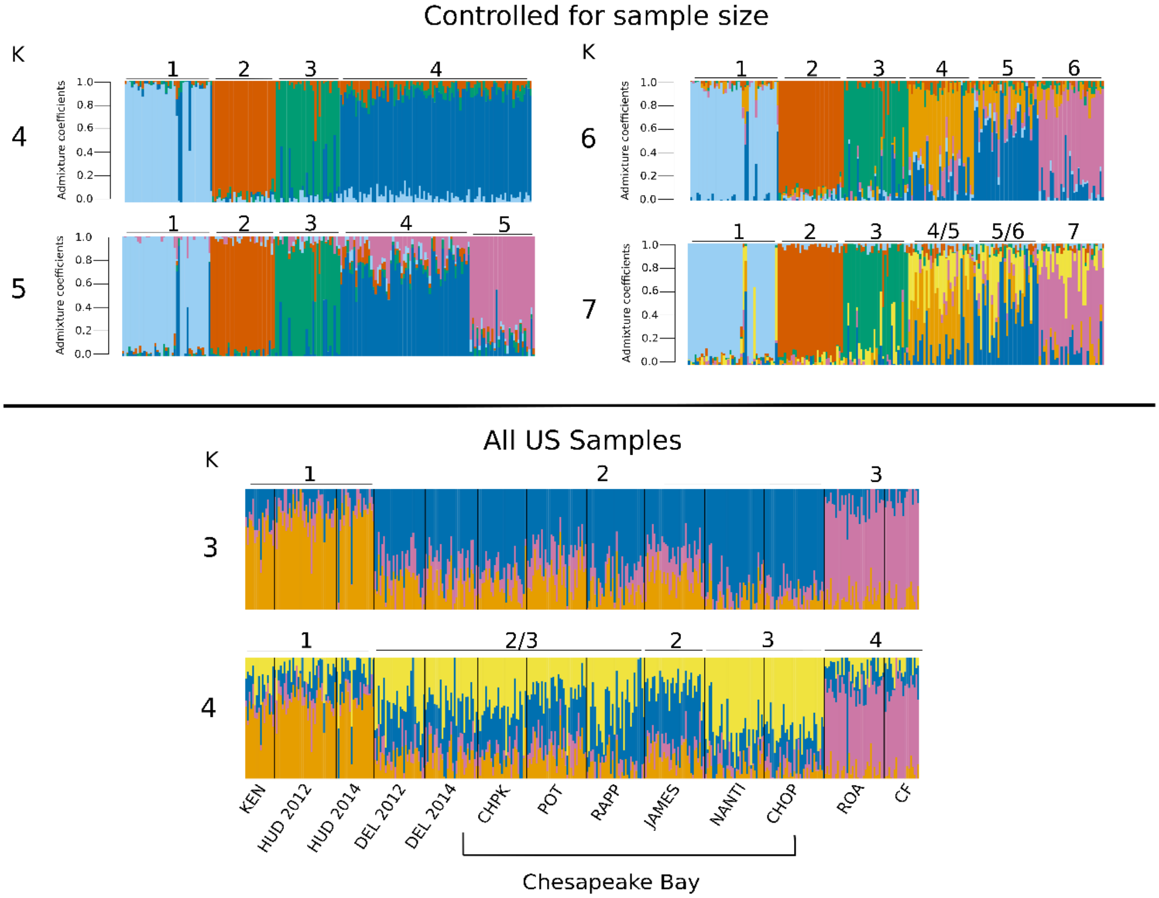
A) Entropy values calculated using the cross-entropy criterion for genetic clusters implemented in LEA, inferred from Striped Bass (*Morone saxatilis*) populations collected in 15 locations, using 1,256 putatively neutral SNP loci. B) Bayesian Information Criterion values for most likely number of clusters in DAPC, run with the same samples.

Appendix C: DAPC graphs at 4-6 clusters



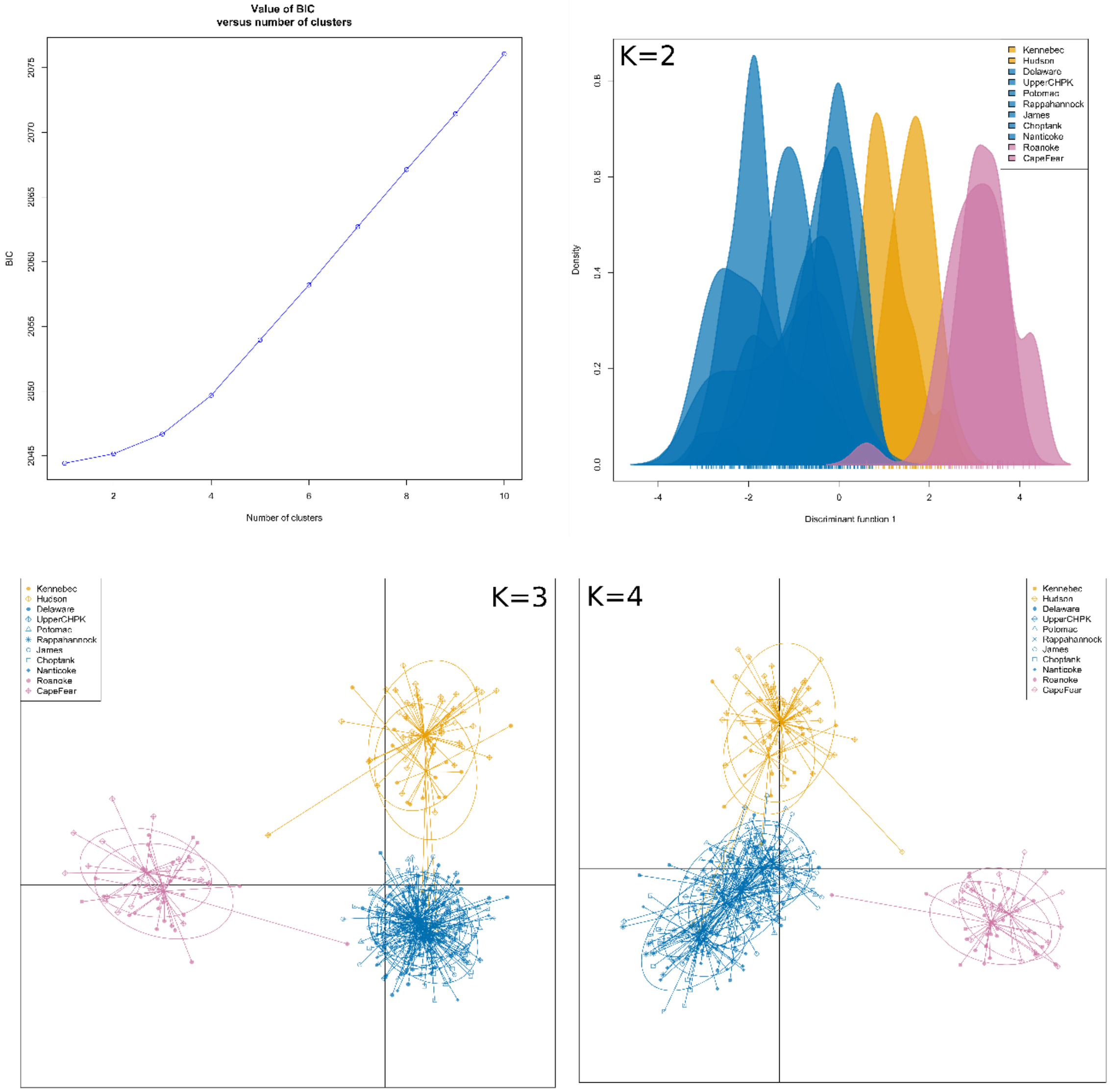
DAPC plots of Striped Bass, *Morone saxatilis*, populations collected in 15 locations, constructed using 1,256 putatively neutral SNP loci and calculated assuming 1) 4 groups, B) 5 groups, C) 6 groups. Individual Striped Bass are represented by symbols depicted in the legend, and a line connects the dot to the site it was sampled in. Distance between dots corresponds to genetic distance along two discriminant functions. Major groupings are labelled according to which populations are contained within.

Appendix D: Additional LEA graphs



On the top, individual admixture coefficients of Striped Bass controlled for sample size, using clusters found at K=6. All results calculated using 1,256 putatively neutral SNP loci. On the bottom, individual admixture coefficients of 374 Striped Bass, *Morone saxatilis*, collected at 11 locations to 3 and 4 genetic clusters. Individual Striped Bass are represented by vertical bars, with percent genotype similarity to each cluster represented by colours. Clusters are numbered and locations are labelled with the cluster they most resemble. Population shorthands are as follows: KEN = Kennebec River, HUD = Hudson River, DEL = Delaware River, CHPK = Upper Chesapeake Bay, POT = Potomac River, RAPP = Rappahannock River, JAMES = James River, CHOP = Choptank River, NANTI = Nanticoke River, ROA = Roanoke River, CF = Cape Fear.

Appendix E: DAPC with Canadian pops removed



1) Bayesian information criterion estimating the most likely number of groups from which Striped Bass collected at 11 locations come, where lower values indicate higher likelihood. 2-4) DAPC plots of Striped Bass, *Morone saxatilis*, populations collected at 11 locations, constructed using 1,256 SNPs and calculated assuming 2) 2 groups, 3) 3 groups, 4) 4 groups. Individual Striped Bass are represented by symbols depicted in the legend, and a line connects the dot to the site it was sampled in. Distance between dots corresponds to genetic distance along two discriminant functions. Major groupings are labelled according to which populations are contained within.

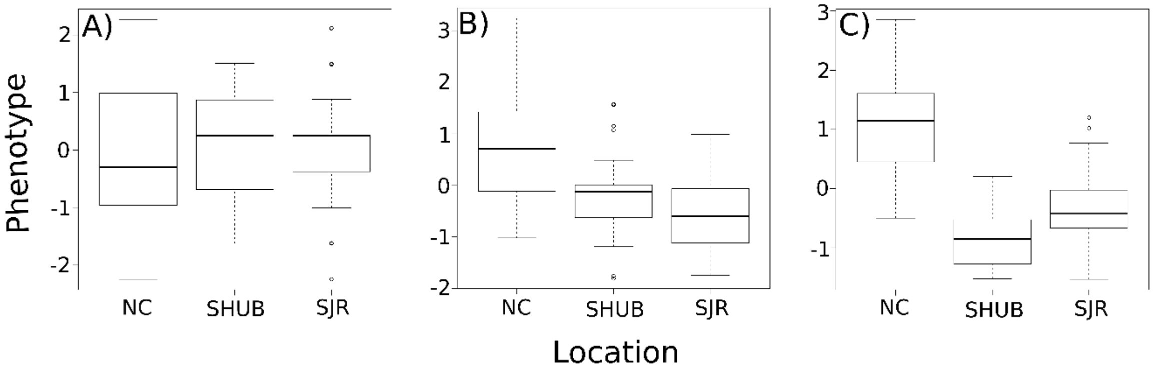
Appendix F: Assignment accuracy to 15 locations

Self-assignment of Striped Bass samples from 15 locations in GeneClass2 using 1,256 putatively neutral SNP loci. Individuals were considered to belong to a reporting group if they were assigned with a confidence score of 80% or more.

Location	N	# Assigned	# Inc. Assigned	# Not Assigned	"Incorrect" assignments
BD-MICHI	19	18	1	0	Assigned to MIRA
MIRA	22	1	20	1	18 BD-MICHI, 1 NANTI, 1 POT
SHUB	33	33	0	0	
SJR	32	29	2	1	2 DEL
KEN	16	3	11	2	2 DEL | 9 HUD
HUD	55	52	3	0	2 CHOP | 1 ROA
DEL	57	29	16	12	4 CHOP | 2 U CHPK | 1 JAMES | 3 NANTI | 3 POT | 3 RAPP
CHPK	27	2	21	4	1 CHOP | 14 DEL | 3 NANTI | 1 POT | 2 RAPP
POT	33	2	22	9	1 NANTI | 5 JAMES | 1 HUD | 1 U CHPK | 14 DEL
RAPP	32	2	25	5	4 CHOP | 1 U CHPK | 12 DEL | 2 JAMES | 4 NANTI | 2 POT
JAMES	33	12	12	9	1 U CHPK | 7 DEL | 4 POT
CHOP	33	8	19	6	11 DEL | 6 NANTI | 1 POT | 1 RAPP
NANTI	33	11	16	6	9 CHOP | 6 DEL | 1 RAPP
ROA	30	29	1	0	1 DEL
CF	22	8	13	1	13 ROA

Abbreviations: N = Number of individuals. BD-MICHI = Bras d'Or-Miramichi, MIRA = Mira River, SHUB = Shubenacadie River, SJR = Saint John River, KEN = Kennebec River, HUD = Hudson River, DEL = Delaware River, CHPK = Upper Chesapeake Bay, POT = Potomac River, RAPP = Rappahannock River, JAMES = James River, CHOP = Choptank River, NANTI = Nanticoke River, ROA = Roanoke River, CF = Cape Fear.

Appendix G: Phenotype-Genetic structure correlations



Boxplots showing three simulated phenotype values of individuals in each of North Carolina (NC), Shubenacadie River (SHUB), and Saint John River (SJR), with A) low correlation with sampling location (KW p-value < 0.99), B) middling correlation (KW p-value < 2.30×10^{-9}, and C) high correlation (KW p-value < 9.1×10^{-23}).

Appendix H: Sex determination in Striped Bass

Teleost fish are known to possess a remarkably wide range of genetic and environmental sex determination systems (Mank et al. 2006, Godwin and Roberts 2018). Perhaps because of this diversity and the lack of morphologically distinct sex chromosomes in most fish species, sex determination mechanisms are still unknown for many species of fish (Gamble 2016), including Striped Bass. The European Sea Bass (*Dicentrarchus labrax*), a close relative of Striped Bass, is known to use a mixture of environmental and polygenic influences to control sex determination (Palaiokostas et al. 2015, Faggion et al. 2019). However, sex determination systems in fish are known to vary within a genus, and even sometimes within a species (Mank and Avise 2009). While a comprehensive study into sex determination in Striped Bass usually involves careful breeding and pedigree analysis, reduced representation SNP libraries such as the one I have constructed for Striped Bass can provide a valuable initial examination (Gamble 2016).

A subset of the samples collected during over the course of my thesis in chapters 1-3 contained information on the sex of the individual (Shubenacadie River n = 33; Roanoke River n = 45; Cape Fear River n = 26; Upper Chesapeake Bay = 33). I also obtained tissue samples taken from Trinity River, Texas (n = 70), a small number of samples collected along the Atlantic coast that did not have a known origin population (n = 18), and a small number of individuals from a Florida hatchery (n = 10), all of which had sex information recorded. Using these 233 samples, I investigated whether I could detect any underlying genetic component to sex determination in Striped Bass.

Association

Two different stacks protocols were used to process SNP data for this analysis. For association studies, samples were processed alongside my chapter 3 samples in stacks 2, using the same parameters and quality filtering thresholds. I applied an initial missing data threshold of 50% and then excluded loci if they had an Fis value < -0.3 or an observed heterozygosity > 0.6 in all populations, using statistics output by the populations module. Using VCFTools, I excluded loci with a minor allele frequency less than 0.05 and loci with more than two alleles, loci that had more than 30% missing data, and genotypes with a read depth less than 7. I used the VCFTools command –missing-indv to examine missing data per individual after this filtering, and excluded any samples with greater than 30% missing data. Then, I used the VCFTools command --relatedness to exclude all but one of any full sibling groups in both of these datasets.

With this dataset, I looked for sex-specific heterozygosity patterns as well as general association with sex as a phenotype. For samples with unknown population origin, I ran a genetic clustering analysis using sparse Non-Negative Matrix Factorization (sNMF) implemented in the R package *LEA* v2.8.0 (Frichot and François 2015) and used ancestral assignment values to determine which of the three major US regions the samples came from. In order to examine whether any loci or chromosomes were exclusively heterozygous in one sex, I used VCFTool's '--hardy' function to extract heterozygosities and looked for loci that were heterozygous in all males but in 2 or fewer females, or loci that were heterozygous in all females but only 2 or fewer males. To identify any genotype-phenotype associations related to sex, I used Bayescan (Foll and Gaggiotti 2008) and an implementation of Random Forest described by Zhao et al. (2012). Bayescan has been used to successfully detect sex-specific loci when association is strong enough to influence population structure analyses (Maroso et al. 2016, Benestan et al. 2017). I used Bayescan to test for loci with unusual patterns of differentiation between male and female individuals, using a burn-in of 50,000 followed by 100,000 iterations, a thinning interval of 10, and a sample size of 5k. Prior odds were set to 1,000. In addition to this, I used Random Forest to build a model differentiating male and female individuals. To correct for population structure in Random Forest, I used the R package *randomforest* v4.6-14 (Liaw and Wiener 2015) and ran a regression analysis based on population assignment and recoded genotypes using the regression residuals (Zhao et al. 2012). Models were run on several ntree and mtry values to determine which resulted in the lowest out-of-bag error, and final values were ntree = 125 and mtry = 72. I first trained a model on a subset of approximately 60% (n = 147) of the dataset, sorted such that an equal number of male and female individuals were included from each location. I then tested the model on the remaining 40% (n = 86) of samples, called holdout samples, to remove upward assignment bias (Anderson 2010). I determined whether the Random Forest model contained genotypes informative of sex if it was able to correctly identify the sex of the test subset of individuals. Bayescan and Random Forest were run on all sampling locations, as well as on individuals from each sampling location separately.

Sex-specific loci

Stacks 2 does not record sample-specific depths of monomorphic loci when run with a reference genome, and so I also processed samples using stacks 1.46 using the same parameters detailed in chapter 2. Using stacks 1.46 I was able to obtain a *matches.tsv* file for each sample that details raw number of reads for each DNA fragment amplified for that individual. Any individual that was excluded in the stacks 2 filtering process due relatedness was also excluded here.

I used an R script created and described in Fowler and Buonaccorsi (2016) and available in Dryad (https://doi.org/10.5061/dryad.3c8s8) which uses raw reads compiled in *matches.tsv* files produced by the stacks pipeline to test for the absence of loci in one sex but not the other. Using this script, I identified loci genotyped at a stack depth of 5 or more in different proportions of male and female Striped Bass to look for sex-specific genes or chromosomes. I first looked for loci present in at least 66% of male individuals but no female individuals, and in 66% of females but no males. I then lowered thresholds to 50% and then to 33% in one sex but not the other. Finally, I looked for loci present in at least 66% of one sex but no more than 1 or 2 individuals in the other.

Results

Two Texas individuals were removed from downstream analyses due to sibling relation. In my stacks 2 dataset 7834 polymorphic SNPs passed quality filtering, largely spread over 24 linkage groups. None of these SNPs were heterozygous in one sex but not the other. Bayescan qvalues were 0.9 or higher for all loci, indicating no outliers related to sex. The Random Forest model was tested on 86 holdout samples (72 male and 14 female), and correctly sexed 55/72 (76%) male and 6/14 (43%) female samples. While overall sexing accuracy was 71%, performance on female samples is particularly poor. In addition, out of bag error rates were approximately 45% across ntree and mtry values, indicating poor overall model performance.

In the stacks 1 dataset I examined 670,168 SNP loci, 106,009 of which were present in at least 1/3 of samples. I found no loci that were present in only male or only female individuals, and no loci that were present in 66% of one sex but only 1 or 2 of the other sex. While a higher-coverage panel of SNPs may be able to detect more subtle genetic influences on sex in Striped Bass, based on my current SNP library I could not find any evidence of genetic sex determination in Striped Bass.

Literature Cited

Anderson, E. C. 2010. Assessing the power of informative subsets of loci for population assignment: Standard methods are upwardly biased. Molecular Ecology Resources 10:701–710.

Benestan, L., J. S. Moore, B. J. G. Sutherland, J. Le Luyer, H. Maaroufi, C. Rougeux, E. Normandeau, N. Rycroft, J. Atema, L. N. Harris, R. F. Tallman, S. J. Greenwood, F. K. Clark, and L. Bernatchez. 2017. Sex matters in massive parallel sequencing: Evidence for biases in genetic parameter estimation and investigation of sex determination systems. Molecular Ecology 26:6767–6783.

Faggion, S., M. Vandeputte, B. Chatain, P. A. Gagnaire, and F. Allal. 2019. Population-specific variations of the genetic architecture of sex determination in wild European sea bass Dicentrarchus labrax L. Heredity 122:612–621.

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

Fowler, B. L. S., and V. P. Buonaccorsi. 2016. Genomic characterization of sex-identification markers in Sebastes carnatus and Sebastes chrysomelas rockfishes. Molecular Ecology 25:2165–2175.

Frichot, E., and O. François. 2015. LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution 6:925–929.

Gamble, T. 2016. Using RAD-seq to recognize sex-specific markers and sex chromosome systems. Molecular Ecology 25:2114–2116.

Godwin, J., and R. Roberts. 2018. Environmental and Genetic Sex Determining Mechanisms in Fishes. Page Transitions Between Sexual Systems.

Liaw, A., and M. Wiener. 2015. randomForest: Breiman and Cutler’s random forests for classification and regression.

Mank, J. E., and J. C. Avise. 2009. Evolutionary diversity and turn-over of sex determination in teleost fishes. Sexual Development 3:60–67.

Mank, J. E., D. E. L. Promislow, and J. C. Avise. 2006. Evolution of alternative sex-determining mechanisms in teleost fishes. Biological Journal of the Linnean Society 87:83–93.

Maroso, F., R. Franch, G. Dalla Rovere, M. Arculeo, and L. Bargelloni. 2016. RAD SNP markers as a tool for conservation of dolphinfish Coryphaena hippurus in the

Mediterranean Sea: Identification of subtle genetic structure and assessment of populations sex-ratios. Marine Genomics 28:57–62.

Palaiokostas, C., M. Bekaert, J. B. Taggart, K. Gharbi, B. J. McAndrew, B. Chatain, D. J. Penman, and M. Vandeputte. 2015. A new SNP-based vision of the genetics of sex determination in European sea bass (Dicentrarchus labrax). Genetics Selection Evolution 47:1–10.

Zhao, Y., F. Chen, R. Zhai, X. Lin, Z. Wang, L. Su, and D. C. Christiani. 2012. Correction for population stratification in random forest analysis. International Journal of Epidemiology 41:1798–1806.

Appendix I: Permission to Reprint from Journals

North American Journal of Fisheries Management

LeBlanc, N. M., S. N. Andrews, T. S. Avery, G. N. Puncher, B. I. Gahagan, A. R. Whiteley, R. A. Curry, and S. A. Pavey. 2018. Evidence of a genetically distinct population of Striped Bass within the Saint John River, New Brunswick, Canada. North American Journal of Fisheries Management 38:1339–1349.

License number 5015080026704, issued Feb 23, 2021 grants permission to reprint in a dissertation or thesis.

Evolutionary Applications

LeBlanc, N. M., B. I. Gahagan, S. N. Andrews, T. S. Avery, G. N. Puncher, B. J. Reading, C. F. Buhariwalla, R. A. Curry, A. R. Whiteley, and S. A. Pavey. 2020. Genomic Population Structure of Striped Bass (Morone saxatilis) from the Gulf of St. Lawrence to Cape Fear River. Evolutionary Applications:1–19.

Permission not required for reproduction in academic theses.

Curriculum Vitae

Nathalie M LeBlanc

PhD Biology (2016-2021)
University of New Brunswick, Saint John, New Brunswick
Thesis: Population Genomics and Association of Striped Bass using double-digest RAD-Seq
Supervisor: Dr. Scott Pavey

Masters Biology (2013-2015)
Acadia University, Wolfville, Nova Scotia
Thesis: Phylogeographic analysis of purple sandpipers (Calidris maritima) as revealed by mitochondrial DNA and microsatellites
Supervisors: Dr. Don Stewart, Dr. Mark Mallory

BSc Biology (Co-op, Honours) and Computer Science (2008-2013)
Acadia University, Wolfville, Nova Scotia
Thesis: Phylogeographic Analysis of Genetic Structure Present in Purple Sandpipers
Supervisors: Dr. Don Stewart, Dr. Mark Mallory

Publications:

Andrews, S. N., Linnansaari, T., Curry, R. A., **LeBlanc, N. M.,** Pavey, S. A. (2020). Winter ecology of Striped Bass (*Morone saxatilis*) near its northern limit of distribution in the Saint John River, New Brunswick. Environmental Biology of Fishes. 103: 1343-1358.

Andrews, S., Linnansaari, T., **Leblanc, N.,** Pavey, S., Curry, R. (2020). Movements of juvenile and sub-adult Striped Bass *Morone saxatilis* in the Saint John River, New Brunswick, Canada. Endangered Species Research, 43, 281–289. doi: 10.3354/esr01074

LeBlanc, N. M., B. I. Gahagan, S. N. Andrews, T. S. Avery, G. N. Puncher, B. J. Reading, C. F. Buhariwalla, R. A. Curry, A. R. Whiteley, S. A. Pavey (2020). Genomic Population Structure of Striped Bass (*Morone saxatilis*) from the Gulf of St. Lawrence to Cape Fear River. Evolutionary Applications, (November 2019), 1–19.

Puncher, G. N., S. Rowe, G. A. Rose, **N. M. LeBlanc**, G. J. Parent, Y. Wang, S. A. Pavey (2019). Chromosomal inversions in the Atlantic cod genome: Implications for management of Canada's Northern cod stock. Fisheries Research, 216, 29–40.

Andrews, S. N., T. Linnansaari, **N. LeBlanc**, S. Pavey, R. A. Curry (2019). Interannual variation in spawning success of Striped Bass (*Morone saxatilis*) in the Saint John River, New Brunswick. River Research and Applications, 36(1), 13–24.

LeBlanc, N. M., S. N. Andrews, T. S. Avery, G. N. Puncher, B. I. Gahagan, A. R. Whiteley, R. A. Curry, and S. A. Pavey. 2018. Evidence of a Genetically Distinct Population of Striped Bass within the Saint John River, New Brunswick, Canada. North American Journal of Fisheries Management 38(6): 1339-1349.

LeBlanc, N. M., D. T. Stewart, S. Pálsson, M. F. Elderkin, G. Mittelhauser, S. Mockford, J. Paquet, G. J. Robertson, R. W. Summers, L. Tudor, M. L. Mallory. 2017. Population structure of Purple Sandpipers (*Calidris maritima*) as revealed by mitochondrial DNA and microsatellites. Ecology and Evolution. 7(9): 3225-3242.

Power, J. W. B., **N. LeBlanc**, S. Bondrup-Nielsen, M. J. Boudreau, M. S. O'Brien, and D. T. Stewart. 2015. Spatial Genetic and Body Size Trends in Atlantic Canada Coyote (*Canis latrans*) Populations. Northeastern Naturalist 22(3): 598-612.

Manuscripts Under Review:

LeBlanc, N. M., S. A. Pavey. Comparing mixed models and Random Forest association tests using naturalGWAS and a Striped Bass SNP dataset. Molecular Ecology Resources. Under review.

Puncher, G. N., Y. Wang, R. Martin, G. DeCelles, S. X. Cadrin, D. Zemeckis, S. Rowe, **N. M. Leblanc**, G. J. Parent, S. A. Pavey. Transborder gene flow between Canada and the U.S.A. and fine-scale population structure of Atlantic cod in the broader Gulf of Maine Region. Transactions of the American Fisheries Society. Accepted.

Conference Presentations:

Genomic Population Structure of Striped Bass (*Morone saxatilis*) from the Gulf of St. Lawrence to Cape Fear River. 2019. Canadian Society for Ecology and Evolution 2019 Annual Meeting. Fredericton, New Brunswick, August 18th – 21st | **Presentation**

Genomic Population Structure of Striped Bass (*Morone saxatilis*) from the Gulf of St. Lawrence to Cape Fear River. 2018. 44th Annual Meeting of the Atlantic International Chapter of the AFS. Vergennes, Vermont, Sept 23-25th | **Presentation**

Genomic Investigation into a breeding population of Striped Bass within the Saint John River. 2018. Canadian Society for Ecology and Evolution 2018 Annual Meeting. Guelph, Ontario, July 18-21 | **Presentation**

Genomic Investigation into a breeding population of Striped Bass within the Saint John River. 2017. 43rd Annual Meeting of the Atlantic International Chapter of the AFS. White Point Resort, Nova Scotia, Sept 17-19th | **Presentation**

Population genomics of Striped Bass using next-generation sequencing. 2016. 42nd Annual Meeting of the Atlantic International Chapter of the AFS. Holderness, New Hampshire, Sept 11-13th | **Presentation**

Population genomics and mixed stock analysis of Striped Bass (*Morone saxatilis*) using ddRAD markers. 2016. Canadian Society for Ecology and Evolution 2016 Annual Meeting. St. John's, Newfoundland. July 7-10th | **Poster**

Phylogeographic analysis of Purple Sandpipers (*Calidris maritima*) as revealed by mitochondrial DNA and microsatellites. 2015. AFO/WOS/SCO Ornithological Conference 2015. Wolfville, Nova Scotia, July 15-19th | **Poster**

Phylogeographic analysis of Purple Sandpipers (*Calidris maritima*) as revealed by mitochondrial DNA and microsatellites. 2015. Canadian Society for Ecology and Evolution 2015 Annual Meeting. Saskatoon, Saskatchewan. May 21-25th | **Presentation**

Phylogeographic analysis of Purple Sandpipers (*Calidris maritima*) as revealed by mitochondrial DNA and microsatellites. 2013. Mersey Tobeatic Research Institute Science Conference. Bridgewater, Nova Scotia, November 1st | **Poster**

Phylogeographic analysis of genetic structure present in Purple Sandpipers. 2013. Science Atlantic Environment Conference. Wolfville, Nova Scotia, March 15-17th | **Poster**

Phylogeographic analysis of Purple Sandpipers (*Calidris maritima*) as revealed by mitochondrial DNA and microsatellites. 2013. Biofeedback, Acadia University. Wolfville, Nova Scotia, Feb | **Presentation**

Evaluation of the Influence of Thrombin on FVIII-expressing Blood Outgrowth Endothelial Cells (BOECs) Implanted Into the Omentum of Hemophilia A Dogs. 2010. Annual General Stem Cell Network Conference. Calgary, Alberta. Nov 22-24th | **Poster**