TRACING INVASIONS BY COMPARING NATIVE AND INTRODUCED POPULATIONS USING EMPIRICAL AND SIMULATED DATA

By

JARED BENJAMIN LEE

(Under the Direction of Rodney Mauricio)

ABSTRACT

Tracing the invasion history of introduced populations is fundamental to understanding any invasion and developing strategies to manage them. The invasion history cannot fully be developed without comparing populations from the native and introduced range. In this dissertation, I trace the invasion of the western mosquitofish, *Gambusia affinis*, in Asia and also examine the impact of missing data on tracing invasions with simulated datasets.

In Chapter 2, I examine three specific biogeographic boundaries previously described in mosquitofish (*G. holbrooki* and *G. affinis*) and examine levels of admixture across them. I demonstrate that the species boundary between *G. affinis* and *G. holbrooki* shows very little admixture. The Savannah River does not seem to be a barrier for gene flow in *G. holbrooki* but instead marks the beginning of a zone of admixture between two distinct types within the species. I also demonstrate that localities from the Mississippi River system are admixed and very different from localities farther west in Texas and Oklahoma.

In Chapter 3, I build upon the results from Chapter 2 and compare them with introduced localities throughout Asia. I also draw upon an extensive historical record and compare it to the inferences made from the genetic results. I find that most, if not all, of the localities sampled

throughout Asia can be traced back to the historical putative source locality in Seabrook, Texas. Genetic diversity was reduced throughout Asia, but very little evidence for a bottleneck was found suggesting that introductions likely occurred in large numbers or were supplemented several times.

In Chapter 4, I simulate RADseq datasets for six invasion scenarios and simulate increasing amounts of missing data in them to assess the impact of missing data on the population genetic estimates and inferences. The probability of correct population assignment was consistently high for all scenarios up to 50% missing data. Low and moderate migration scenarios performed better up to 90% missing data. The filtering process had no improvement from the random subsets tested in estimating F_{ST}, but the assignment test probabilities improved with all filtered datasets.

INDEX WORDS: mosquitofish, China, population genetics, RADseq, invasive species, assignment, southeastern United States, phylogeography

TRACING INVASIONS BY COMPARING NATIVE AND INTRODUCED POPULATIONS USING EMPIRICAL AND SIMULATED DATA

by

JARED BENJAMIN LEE

B.S., Brigham Young University, 2005

M.S., Brigham Young University, 2009

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

TRACING INVASIONS BY COMPARING NATIVE AND INTRODUCED POPULATIONS USING EMPIRICAL AND SIMULATED DATA

by

JARED BENJAMIN LEE

Major Professor:	Rodney Mauricio
Committee:	Kelly Dyer Travis Glenn John Maerz John Wares

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2014

ACKNOWLEDGEMENTS

The National Science Foundation Partnerships in International Research and Education program (Grant No. OISE 0730218) provided the funding for my field and lab work, along with my stipend for the duration of my time at the University of Georgia. The National Science Foundation East Asia and Pacific Summer Institute funded my stay in China for two months during the summer of 2011. This work was performed with the support of the Georgia Genomics Facility at the University of Georgia. This research was supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology.

I am indebted to my advisor, Rodney Mauricio, for welcoming me into his lab and giving me the freedom to pursue my research interests. My committee members (Kelly Dyer, Travis Glenn, John Maerz, and John Wares) provided critical feedback, gave needed encouragement, and answered many questions throughout each phase of my dissertation.

I am grateful for all of my labmates in the Mauricio Lab, who have supported and encouraged me over the years. Kerin Bentley started the program with me and has always been a great support through the best and worst times. Sandra Hoffberg has been a great sounding board for all of my ideas and questions. Joan West helped me get going with my lab work and answered my many questions both large and small in the lab.

The specimens that make up the bulk of my research were no trivial task to obtain. I thank the following individuals and institutions for their assistance: C.H. Chang (Academia Sinica), Y.F. Chen (Chinese Academy of Sciences), D. Dionisio, T. Dowling (Arizona State University), B. Freeman (University of Georgia), B. Kuhajda (University of Alabama), S.M. Lin

(National Taiwan Normal University), N. Onikura (Kyushu University), M. Roberts (Mississippi Museum of Natural Sciences), J. Schaefer (University of Southern Mississippi), W.C. Starnes (North Carolina Museum of Natural Sciences), W.Q. Tang (Shanghai Ocean University), C.G. Zhang (Chinese Academy of Sciences), X.B. Wu (Anhui Normal University), and Q. Zhang (Jilin University). Many of them curate large collections of museum voucher specimens whose value I consider priceless. They also provided much needed assistance on the ground in the way of equipment and personnel that made my collections possible.

Megan Behringer, Adam Bewick, Ryan Johnson, Katie Pieper, and Brian Whigham helped me write the scripts that made Chapter 4 possible. I am grateful to have the support of a stellar group of graduate students in the Genetics Department, especially Emily Bewick and Sarah Sander who were always available to review a manuscript, interpret results, or just chat about ideas over ice cream. Peter Unmack trained me in the lab so many years ago and has constantly provided input on my projects over the years.

The most recognition goes to Heather Lee, my wife, who has stood by me and supported me throughout my graduate career. She let me work at all hours of the day and night, at home, school, and abroad. I am better because of her and look forward to our next adventures together.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

Invasive species are a threat biological diversity around the globe. It is estimated that approximately 42% of species listed as threatened or endangered are at risk primarily to invasive species (Pimentel *et al.* 2005). Introductions of the Nile perch and the brown tree snake are common examples of invasive species that have led to the extinction of many native species (Ogutu-Ohwayo 1990; Wiles *et al.* 2003). Furthermore, rising CO_2 levels, warmer temperatures, and altered precipitation patterns due to global climate change has the potential to exacerbate the threat by facilitating the spread and persistence of invasive species (Bradley *et al.* 2010; Rahel & Olden 2008). Thus, understanding biological invasions is important for preserving biological diversity.

Biological invasions occur when organisms are introduced, establish, and rapidly spread outside of their native range (Elton 1958). Depending on the niche of that organism in the introduced range, there will be a spectrum of environmental impacts ranging from relatively minor to extremely damaging. Species with greater detrimental effects tend to attract more attention and are the focus of much research (Lowe *et al.* 2000). The movement of organisms into new ranges is also a natural ecological phenomenon (Vermeij 1991). For example, the closing of the isthmus of Panama led to the Great Biotic Interchange where the flora and fauna of North America and South America came into contact with one another after being separated for millions of years (Marshall 1988). These natural invasions are different from biological invasions because they usually occur over thousands to millions of years, whereas a biological invasion can occur over a period of a few centuries or less. However, biological invasions are also different from natural range expansion because the species often overcome major geographic barriers through human-mediated transport. For example, the zebra mussel is native

to the Black and Caspian seas but has established itself in many European and North American waterways by being transported in ballast water of transoceanic vessels (May *et al.* 2006). Humans also move organisms around as a food source, which is the case for bullfrogs (Culley 1981), Nile perch (Pringle 2005), and crayfish (Yan *et al.* 2001). Thus, human-mediated dispersal is a key component to biological invasions since they overcome natural barriers in shorter periods of time than would occur naturally.

It is important to study invasive species in their native and introduced range in order to test hypotheses regarding the introduction, potential causes for invasiveness, and the impact of invasions (Hierro *et al.* 2005). Invasive species are known to undergo a kind of accelerated evolution and adapt to their environment in just a few generations (Cox 2004) and attempting to demonstrate specific adaptations enabling the success of an organism's invasion proves to be challenging (Keller & Taylor 2008). Accurate knowledge of the invasion history allows studies to be designed in which native source populations are compared with introduced populations for potentially adaptive traits (Estoup & Guillemaud 2010). Introduced populations of the brown anole in Florida contained unique combinations of mitochondrial haplotypes that did not occur together in the native range (Kolbe *et al.* 2004). They further found other introduced populations of the brown anole were derived from the Florida introductions. The subsequent introductions from Florida thus contain more genetic diversity than those in the native range and have greater evolutionary potential. Research from the native and introduced range for many invasive species has provided valuable information on studying the accelerated evolution of invasive species (Ascunce *et al.* 2011; Blum *et al.* 2007; Brown *et al.* 2007; Caldera *et al.* 2008; Estoup *et al.* 2004; Estoup *et al.* 2001; Tsutsui & Suarez 2000, 2001).

Tracing invasion routes involves describing how the introduction took place and what routes were taken. The number of introductions can vary from a single introduction to repeated introductions of individuals (Kolbe *et al.* 2004; Tsutsui & Suarez 2001). The routes of the introduction can also be determined, as in the case for brown anole, which was introduced many times from different parts of the native range in Cuba into Florida and then subsequently introduce from Florida to other parts of the world (Kolbe *et al.* 2004). Thus, tracing invasions will involve comparing populations from the native range and all introduced ranges of interest.

In order to trace invasions, studies often employ direct and indirect methods for ascertaining the source and mode of introduction (Austin *et al.* 2011; Estoup & Guillemaud 2010). Direct methods may include published accounts and records of introductions (Suarez & Tsutsui 2004), whereas indirect methods involve looking at genetic patterns in both the native and introduced range (Pascual *et al.* 2007). Direct methods may suffer from inaccuracy or incompleteness because recorded accounts are anecdotal, lack details, or are second-hand accounts. However, some records may contain extensive detail regarding the introduction, as may be the case for a biological control agent sponsored by a government agency. Indirect methods typically rely on population genetics to estimate demographic parameters like the number of founders and the geographical source of the invasion. The reliability of these methods can vary depending on the number localities sampled, the number of markers used, and the amount of genetic variability in the introduced range (Fitzpatrick *et al.* 2012). Together both methods can complement one another in providing a clearer picture regarding the invasion. For example, recorded introductions of mosquitofish into Europe guided sampling efforts in the native range to identify source populations (Vidal *et al.* 2009). Moreover, with very little

historical data, population genetics determined bullfrogs were introduced to Europe six times (Ficetola *et al.* 2008).

The population genetic and phylogeographic methods used for tracing invasions, where samples from multiple individuals in multiple populations are scored for a suite of genetic markers for analysis, have a certain 'forensic' aspect to them since they attempt to reconstruct past events based on current data (Dlugosch & Parker 2008; Wares *et al.* 2005). These methods are used for a broad range of applications including understanding the origin of humans (Ayala 1995), identifying illegal ivory trade (Wasser *et al.* 2004), and excluding suspects in criminal proceedings (Metzker *et al.* 2002) to name just a few. For invasive species, some examples of indirect methods addressing important questions include the geographical source of invasions (Caldera *et al.* 2008; Kolbe *et al.* 2004; Tsutsui & Suarez 2001), the number of invasions (Dlugosch & Parker 2008; Holland *et al.* 2004), and genetic diversity within introduced populations (Tsutsui & Suarez 2000).

The distribution of genetic diversity across the native range can broadly be referred to as population structure and can provide useful insights when tracing invasions. Native populations containing highly structured populations would exhibit distinct genetic signatures across the range. An excellent example of this is the brown anole in its native range in Cuba. Kolbe *et al.* (2004) found that introduced haplotypes in Florida came from eight distinct clades in the native range. Due to the high degree of population structure, the identity of the source populations and occurrence of multuple introductions was easy to detect. However, when native ranges exhibit lower levels of population structure tracing invasions becomes more challenging. In the zebra mussel, only two mitochondrial haplotypes were found in introduced European and North American populations. These two haplotypes were distributed across many localities in the

native range. Part of the native range was excluded as being the source, but the lower levels of population structure made it difficult to identify a specific source locality (May *et al.* 2006).

Phylogeographic and population genetic studies on native freshwater fishes of the southeastern United States have described several broad patterns of population structure that are concordant across species suggesting a shared history (Bermingham & Avise 1986; Soltis *et al.* 2006). Many species have shown distinct Atlantic and Gulf Coast lineages with a break occurring somewhere on the Florida peninsula (Bowen & Avise 1990; Gold & Richardson 1998; Gold *et al.* 1999; Keeney *et al.* 2005; Wirgin *et al.* 2002). Another major pattern found in fishes is an east-west split at the Apalachicola River in Florida (Kristmundsdóttir & Gold 1996; Philipp *et al.* 1983; Wooten & Lydeard 1990). Species distributed across the Mississippi River have aso shown an east-west split in population structure (Near *et al.* 2001). The Ozark and Appalachian mountains also influence the population structure of fish species in the region (Gonzalez-Vilasenor & Powers 1990; Strange & Burr 1997). These major patterns in the southeastern United States can also be found in many other taxa besides fish (Soltis *et al.* 2006). However, not all species show these same patterns and some fish species show no population structure at all (Buonaccorsi *et al.* 2001; Turner *et al.* 1996; Zatcoff *et al.* 2004). In a review of phylogeographic patterns found in the southeastern United States, Soltis *et al.* (2006) observed that current patterns described above are often explained using Pleistocene refugia models, but some of the lineages they reviewed may be older suggesting a Pliocene divergence. However, many species in the region have their own distinct phylogeograhpic patterns that may be the result of other mechanisms (Near & Keck 2005; Scott *et al.* 2009), but their similarity to the patterns described above may lead to erroneous conclusions (Soltis *et al.* 2006).

Mosquitofish, *Gambusia affinis* and *G. holbrooki*, are two species native to the southeastern United States but introduced around the world (Pyke 2008). They are the most widely distributed species in the genus, which is the largest genus in the family Poeciliidae (Pyke 2005). With a broad distribution across much of the southeastern United States, mosquitofish provide an excellent system to study the phyolgeographic patterns in the southeastern United States and use those patterns to trace the invasion of mosquitofish around the world.

In the chapters that follow, I utilize fundamental population genetic and phylogeographic methods to compare native and introduced populations with empirical and simulated data. In Chapter 2, I test three proposed genetic breaks in the native range of mosquitofish. This chapter allows me to explore the genetic diversity and population structure in the native range. While previous studies have explored the genetic diversity of these species in a descriptive way, I revisit the proposed genetic breaks and test them with a unique set of molecular markers, broader geographical sampling, and modern analytical methods.

In Chapter 3, I reconstruct the invasion route of *G. affinis* for populations throughout Asia. While invasions of *G. holbrooki* have been thoroughly explored (Ayres *et al.* 2012; Ayres *et al.* 2010; Vidal *et al.* 2009), Asian introductions of *G. affinis* are among some of the earliest recorded and provide a parallel comparison between the two species. I draw upon an extensive historical record (direct methods) and compare it with the results from genetic markers (indirect methods). The results from the native range (Chapter 2) make conclusions easier since both ranges are sampled thoroughly with the same markers.

Finally, in Chapter 4, I simulate RADseq datasets for several invasion scenarios and explore how missing data in these datasets impacts the parameter estimates and potentially alters the conclusions made. In Chapters 2 and 3, I use conventional sequencing and genotyping

methods to generate the data, however, in recent years new technology has provided the ability to generate data for hundreds of individuals for thousands of loci. Since the generation of these large, genome-wide datasets is still new, it is an ideal time to explore how missing data influences their analysis, particularly for invasion scenarios. This project allows me to look at how next-generation sequencing (NGS) technology can impact the field of invasion biology and also explore how researchers using RADseq datasets can best utilize the tools available for analyzing them.

All of these chapters emphasize the theme that studying invasions requires the comparison of native and introduced populations. Often in the literature, studies on invasive species will sample one range more than the other (usually the introduced range), which can impact the ability to exclude hypotheses regarding the invasion route. I demonstrate that by comparing both the native and introduced populations we gain a clear picture of the invasion history.

References

Ascunce M, Yang C, Oakey J, Calcaterra L (2011) Global invasion history of the fire ant Solenopsis invicta. *Science* **331**, 1066-1068.

Austin CC, Rittmeyer EN, Oliver La*, et al.* (2011) The bioinvasion of Guam: inferring geographic origin, pace, pattern and process of an invasive lizard (*Carlia*) in the Pacific using multi-locus genomic data. *Biological Invasions* **13**, 1951-1967.

Ayala F (1995) The Myth of Eve: Molecular Biology and Human Origins. *Science* **270**, 1930-1936.

Ayres R, Pettigrove V, Hoffmann A (2012) Genetic structure and diversity of introduced eastern mosquitofish (*Gambusia holbrooki)* in south-eastern Australia. *Marine and Freshwater Research* **63**, 1206-1214.

Ayres RM, Pettigrove VJ, Hoffmann Aa (2010) Low diversity and high levels of population genetic structuring in introduced eastern mosquitofish (*Gambusia holbrooki*) in the greater Melbourne area, Australia. *Biological Invasions* **12**, 3727-3744.

Bermingham E, Avise JC (1986) Molecular zoogeography of freshwater fishes in the southeastern United States. *Genetics* **113**, 939-965.

Blum MJ, Jun Bando K, Katz M, Strong DR (2007) Geographic structure, genetic diversity and source tracking of Spartina alterniflora. *Journal of Biogeography* **34**, 2055-2069.

Bowen B, Avise J (1990) Genetic structure of Atlantic and Gulf of Mexico populations of sea bass, menhaden, and sturgeon: influence of zoogeographic factors and life-history patterns. *Marine Biology* **107**, 371-381.

Bradley Ba, Blumenthal DM, Wilcove DS, Ziska LH (2010) Predicting plant invasions in an era of global change. *Trends in Ecology & Evolution* **25**, 310-318.

Brown GP, Shilton C, Phillips BL, Shine R (2007) Invasion, stress, and spinal arthritis in cane toads. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 17698-17700.

Buonaccorsi V, Starkey E, Graves J (2001) Mitochondrial and nuclear DNA analysis of population subdivision among young-of-the-year Spanish mackerel (Scomberomorus maculatus) from the western Atlantic and Gulf of Mexico. *Marine Biology* **138**, 37-45.

Caldera EJ, Ross KG, DeHeer CJ, Shoemaker DD (2008) Putative native source of the invasive fire ant *Solenopsis invicta* in the USA. *Biological Invasions* **10**, 1457-1479.

Cox GW (2004) *Alien Species and Evolution* Island Press, Washington DC.

Culley D (1981) Have we turned the corner on bullfrog culture? *Aquaculture magazine*.

Dlugosch KM, Parker IM (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology* **17**, 431-449.

Elton CS (1958) *The Ecology of Invasions by Animals and Plants* John Wiley & Sons, Inc., New York.

Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet J-M (2004) Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, Bufo marinus. *Evolution; international journal of organic evolution* **58**, 2021-2036.

Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data: why, how and so what? *Molecular Ecology* **19**, 4113-4130.

Estoup a, Wilson IJ, Sullivan C, Cornuet JM, Moritz C (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, Bufo marinus. *Genetics* **159**, 1671-1687.

Ficetola GF, Bonin A, Miaud C (2008) Population genetics reveals origin and number of founders in a biological invasion. *Molecular Ecology* **17**, 773-782.

Fitzpatrick BM, Fordyce Ja, Niemiller ML, Reynolds RG (2012) What can DNA tell us about biological invasions? *Biological Invasions* **14**, 245-253.

Gold J, Richardson L (1998) Mitochondrial DNA diversification and population structure in fishes from the Gulf of Mexico and western Atlantic. *Journal of Heredity* **89**, 404-414.

Gold J, Richardson L, Turner T (1999) Temporal stability and spatial divergence of mitochondrial DNA haplotype frequencies in red drum (Sciaenops ocellatus) from coastal regions of the western Atlantic Ocean and Gulf of Mexico. *Marine Biology* **133**, 593-602.

Gonzalez-Vilasenor LI, Powers DA (1990) Mitochondrial-DNA restriction-site polymorphisms in the teleost Fundulus heteroclitus support secondary intergradation. *Evolution*, 27-37.

Hierro J, Maron J, Callaway R (2005) A biogeographical approach to plant invasions: the importance of studying exotics in their introduced and native range. *Journal of Ecology* **93**, 5-15.

Holland BS, Dawson MN, Crow GL, Hofmann DK (2004) Global phylogeography of *Cassiopea* (Scyphozoa: Rhizostomeae): molecular evidence for cryptic species and multiple invasions of the Hawaiian Islands. *Marine Biology* **145**, 1119-1128.

Keeney D, Heupel M, Hueter R, Heist E (2005) Microsatellite and mitochondrial DNA analyses of the genetic structure of blacktip shark (Carcharhinus limbatus) nurseries in the northwestern Atlantic, Gulf of Mexico, and Caribbean Sea. *Molecular Ecology* **14**, 1911-1923.

Keller SR, Taylor DR (2008) History, chance and adaptation during biological invasion: separating stochastic phenotypic evolution from response to selection. *Ecology Letters* **11**, 852-866.

Kolbe JJ, Glor RE, Rodríguez Schettino L*, et al.* (2004) Genetic variation increases during biological invasion by a Cuban lizard. *Nature* **431**, 177-181.

Kristmundsdóttir ÁÝ, Gold JR (1996) Systematics of the Blacktail Shiner (Cyprinella venusta) inferred from analysis of mitochondrial DNA. *Copeia*, 773-783.

Lowe S, Browne M, Boudjelas S, De Poorter M (2000) 100 of the world's worst invasive species. A selection from the Global Invasive Species Database. The Invasive Species Specialist Group (ISSG) a sepcialist group of the Species Survival Commission (SSC) of the World Conservation Union (IUCN).

Marshall L (1988) Land Mammals and the Great American Interchange. *American Scientist* **76**, 380-388.

May GE, Gelembiuk GW, Panov VE, Orlova MI, Lee CE (2006) Molecular ecology of zebra mussel invasions. *Molecular Ecology* **15**, 1021-1031.

Metzker ML, Mindell DP, Liu X-M*, et al.* (2002) Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 14292-14297.

Near TJ, Keck BP (2005) Dispersal, vicariance, and timing of diversification in Nothonotus darters. *Molecular Ecology* **14**, 3485-3496.

Near TJ, Page LM, Mayden RL (2001) Intraspecific phylogeography of Percina evides (Percidae: Etheostomatinae): an additional test of the Central Highlands pre‐Pleistocene vicariance hypothesis. *Molecular Ecology* **10**, 2235-2240.

Ogutu-Ohwayo R (1990) The decline of the native fishes of lakes Victoria and Kyoga (East Africa) and the impact of introduced species, especially the Nile perch, Lates niloticus, and the nile tilapia, Oreochromis niloticus. *Environmental biology of fishes*.

Pascual M, Chapuis MP, Mestres F*, et al.* (2007) Introduction history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods. *Molecular Ecology* **16**, 3069-3083.

Philipp DP, Childers WF, Whitt GS (1983) A biochemical genetic evaluation of the northern and Florida subspecies of largemouth bass. *Transactions of the American Fisheries Society* **112**, 1-20.

Pimentel D, Zuniga R, Morrison D (2005) Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics* **52**, 273-288.

Pringle RM (2005) The Origins of the Nile Perch in Lake Victoria. *BioScience* **55**, 780.

Pyke GH (2005) A Review of the Biology of *Gambusia affinis* and *G. holbrooki*. *Reviews in Fish Biology and Fisheries* **15**, 339-365.

Pyke GH (2008) Plague Minnow or Mosquito Fish? A Review of the Biology and Impacts of Introduced *Gambusia* Species. *Annual Review of Ecology, Evolution, and Systematics* **39**, 171-191.

Rahel FJ, Olden JD (2008) Assessing the effects of climate change on aquatic invasive species. *Conservation Biology* **22**, 521-533.

Scott CH, Cashner M, Grossman GD, Wares JP (2009) An awkward introduction: phylogeography of Notropis lutipinnis in its 'native' range and the Little Tennessee River. *Ecology of Freshwater Fish* **18**, 538-549.

Soltis DED, Morris ABA, McLachlan JS, Manos PS, Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology* **15**, 4261-4293.

Strange RM, Burr BM (1997) Intraspecific phylogeography of North American highland fishes: a test of the Pleistocene vicariance hypothesis. *Evolution*, 885-897.

Suarez AV, Tsutsui ND (2004) The Value of Museum Collections for Research and Society. *BioScience* **54**, 66.

Tsutsui N, Suarez A (2000) Reduced genetic variation and the success of an invasive species. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5948-5953.

Tsutsui N, Suarez A (2001) Relationships among native and introduced populations of the Argentine ant (*Linepithema humile*) and the source of introduced populations. *Molecular Ecology* **10**, 2151-2161.

Turner TF, Trexler JC, Kuhn DN, Robison HW (1996) Life-history variation and comparative phylogeography of darters (Pisces: Percidae) from the North American central highlands. *Evolution* **50**, 2023-2036.

Vermeij GJ (1991) When biotas meet: understanding biotic interchange. *Science* **253**, 1099-1104.

Vidal O, García-Berthou E, Tedesco Pa, García-Marín J-L (2009) Origin and genetic diversity of mosquitofish (*Gambusia holbrooki*) introduced to Europe. *Biological Invasions* **12**, 841-851.

Wares JP, Hughes AR, Grosberg RK (2005) Mechanisms that Drive Evolutionary Change. In: *Species Invasions: Insights from Species Introduction and Invasions*, pp. 229-257.

Wasser SK, Shedlock AM, Comstock K, *et al.* (2004) Assigning African elephant DNA to geographic region of origin: applications to the ivory trade. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 14847-14852.

Wiles G, Bart J, Beck R, Aguon C (2003) Impacts of the brown tree snake: patterns of decline and species persistence in Guam's avifauna. *Conservation Biology* **17**, 1350-1360.

Wirgin I, Waldman J, Stabile J, Lubinski B, King T (2002) Comparison of mitochondrial DNA control region sequence and microsatellite DNA analyses in estimating population structure and gene flow rates in Atlantic sturgeon Acipenser oxyrinchus. *Journal of Applied Ichthyology* **18**, 313-319.

Wooten M, Lydeard C (1990) Allozyme variation in a natural contact zone between Gambusia affinis and Gambusia holbrooki. *Biochemical systematics and ecology* **18**, 169-173.

Yan X, Zhenyu L, Gregg W, Dianmo L (2001) Invasive species in China—an overview. *Biodiversity & Conservation* **10**, 1317-1341.

Zatcoff M, Ball A, Sedberry G (2004) Population genetic analysis of red grouper, Epinephelus morio, and scamp, Mycteroperca phenax, from the southeastern US Atlantic and Gulf of Mexico. *Marine Biology* **144**, 769-777.

CHAPTER 2: PHYLOGEOGRAPHY AND POPULATION GENETICS OF NATIVE MOSQUITOFISH (GAMBUSIA AFFINIS AND GAMBUSIA HOLBROOKI): TESTING GENETIC BREAKS WITH MULTIPLE LOCI[1]

[1] Lee JB and Mauricio R. To be submitted to *Journal of Biogeography*.

Abstract

Phylogeography has grown as a field over the last 25 years and has provided a broad range of results for natural populations around the globe. Early phylogeographic studies throughout the southeastern United States revealed a number of breaks that were concordant for many species. However, despite analytical advancements in phylogeography and population genetics, few studies have revisited this region to reexamine some of the early results in this field. We use two closely related livebearing fish, *Gambusia affinis* and *Gambusia holbrooki,* native to the southeastern United States to examine three previously described genetic breaks in the region: (1) the species boundary between *G. affinis* and *G. holbrooki* in Alabama, (2) the Savannah River, and (3) the Mississippi River. We genotyped 18 microsatellite markers and sequenced a mitochondrial DNA fragment in 42 localities across the range of both species. We observed very little gene flow across the species boundary between these two taxa and add further detail to the species boundary. *Gambusia holbrooki* localities did not exhibit a strong genetic break at the Savannah River. While populations north and south of the Savannah River are different, localities in South Carolina show a great deal of admixture between the two groups. The evidence did not indicate the Mississippi River as a barrier for dispersal, instead all localities within the Mississippi River clustered together while localities west of the Mississippi drainage were a unique group. Our results are largely concordant with previous studies, but provide valuable information from more extensive geographic sampling. Since this species has been introduced around the world, we also discuss how this study can help out with future studies of mosquitofish invasions.

Introduction

The diversification of populations from one another is an important focus of evolutionary biology. Isolated populations have the potential to evolve on independent trajectories and given enough time can form new species (Coyne & Orr 2004). This process can begin within a species as barriers to gene flow begin to isolate populations leading to population structure. For over 25 years, phylogeography has described intraspecific patterns of genetic diversity, gene flow, and demography with much of the emphasis in North America and Europe (Beheregaray 2008; Soltis *et al.* 2006; Taberlet *et al.* 1998). As a result we have recognized many phylogeographic breaks that help structure populations. During the same time, technical and analytical advances in population genetics have enabled us to probe deeper into the demographic factors behind population structure and test more complex scenarios of population subdivision (Hickerson *et al.* 2010).

Some of the early studies of phylogeography described a number of patterns in the southeastern United States with a variety of taxa (Avise *et al.* 1987; Bermingham & Avise 1986). These studies revealed several genetic breaks corresponding to geographic features that in some cases were concordant across multiple taxa. This suggests a shared history often explained by glaciation cycles and Pleistocene refugia models. While not all species are concordant, studies have found similar genetic breaks across a wide range of taxa in the southeastern United States, (Soltis *et al.* 2006).

The closely related livebearing fishes *Gambusia affinis* and *Gambusia holbrooki* were the subject of early phylogeographic studies using both allozymes and mitochondria RLFPs describing population structure, gene flow, and genetic diversity patterns in the zone of sympatry between the two species (Scribner & Avise 1993; Wooten *et al.* 1988). Both species are

widespread throughout the southeastern United States and belong to the largest genus of livebearing fish (Poeciliidae). They are also the only two species of this genus that have successfully been introduced outside their native range. Commonly referred to as mosquitofish, *G. affinis* and *G. holbrooki* were widely introduced around the world in the early 20th century in an effort to control mosquito populations (Krumholz 1948). Today, due to their introductions, they are the most widely distributed freshwater fish established on all continents except Antarctica (Pyke 2005, 2008). Among the many negative environmental impacts of mosquitofish, they are known to prey upon and eliminate native larvae and juveniles of a variety of invertebrates, fish, and amphibians (Stockwell & Henkanaththegedara 2011) and are considered one of the worst invasive species in the world due to their worldwide distribution and high fecundity (Lowe *et al.* 2000; Pyke 2008). Given the environmental concerns regarding these species, it is necessary to reconstruct the invasion history of mosquitofish in an attempt to identify source populations and the number of introductions in order to better inform management strategies (Estoup & Guillemaud 2010). Researchers have attempted to reconstruct the invasion history of mosquitofish in Europe, Australia, and New Zealand (Ayres *et al.* 2010; Purcell *et al.* 2012; Vidal *et al.* 2009). However, comparisons with the native range are difficult without adequate sampling and common markers.

This study provides a firm knowledge of the genetic diversity of mosquitofish, which can then be used to compare with genetic patterns in the various introduced ranges and reconstruct invasion histories. We use mitochondrial sequence data (mtDNA) and microsatellite markers to understand the genetic diversity and population structure of *G. affinis* and *G. holbrooki* throughout their native range. Specifically, we test for three different genetic breaks across the range of these two species: (A) the species boundary in Alabama and western Georgia (Scribner

& Avise 1993; Wooten *et al.* 1988); (B) a break at the Savannah River in *G. holbrooki* resulting in two distinct types on either side (Wooten *et al.* 1988); and (C) a break at the Mississippi River in *G. affinis* dividing the range (Soltis *et al.* 2006). Aside from testing these specific patterns, we also were careful to observe any other unexpected patterns as no other study has assayed these two species to the geographical extent we present here. While the species boundary between the two species is clear, the other two breaks do not pose major barriers for gene flow.

Materials & Methods

Study system

Western and eastern mosquitofish (*G. affinis* and *G. holbrooki*, respectively) are native to the southeastern United States. The western mosquitofish's range extends from northern Mexico up through Oklahoma and eastward to northern Georgia with its northern limit extending through Missouri. The eastern mosquitofish's range starts in southern Florida and moves north through much of the Atlantic seaboard states including Maryland and New Jersey, but only goes west as far as the Appalachian mountains and into Alabama. Both species inhabit slow moving water in lakes, ponds, and rivers feeding on a broad diet. The two species are often considered together due to their similar biology, use as a mosquito control agent, and taxonomic confusion (Pyke 2005, 2008). Early studies into the patterns of population structure of these two species based on a suite of allozyme loci and mitochondrial RFLPs revealed a zone of sympatry in Alabama extending into western Georgia (Scribner & Avise 1993; Wooten *et al.* 1988). Furthermore, two distinct forms of *G. holbrooki* were observed seemingly divided by the Savannah River (Wooten *et al.* 1988). Interestingly, these genetic breaks are not the same as other fish species in the region (Bermingham & Avise 1986). However, despite its importance as an invasive species, no

recent study has examined the population structure extensively throughout the range of both species.

Sampling & Laboratory protocols

We sampled 42 localities of mosquitofish (*G. affinis* = 24, *G. holbrooki* = 18) from the majority of the range for both species (Table 2.1, Figure 2.1). Many samples were obtained from alcohol preserved museum voucher specimens and we further supplemented these with direct field sampling. Fish were caught with a dip net and immediately preserved in 100% alcohol. We identified the species by examining the morphology of the gonopodium on all mature males in a locality (Rauchenberger 1989). We extracted genomic DNA from muscle tissue from each specimen using a modified phenol-chloroform protocol (Hillis *et al.* 1996).

We used polymerase chain reaction (PCR) to amplify a fragment of the mitochondrial gene cytochrome *b* (cyt *b*) for 10 individuals per locality (21 and 23 excluded) using the primers *cytb516F* (5' YGCCACCTTAACTCGCTTCT 3') and *Thr23R* (5' CGGTTTACAAGACCGACGCT 3'), which were designed for this study. PCR amplifications had a 25 μl volume [10mM Tris-HCl, 50 mM KCl, 0.1% Triton X-100, 0.4 μM of each primer, 2.0 mM $MgCl_2$, 0.4 mM dNTPs, 0.5 units EconoTaq DNA Polymerase (Lucigen), and ~25 ng DNA template] and were carried out using the following thermal profile: initial denaturation for 180s followed by 30 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 30 s with a final extension for 300 s. PCR product was purified using an EXOSAP protocol (Glenn & Schable 2005). All cycle-sequencing reactions were run following the ABI manufacturer's protocols (Applied Biosystems, Inc.). Sequences were obtained using a Applied Biosystems 3730 XL automated DNA sequencer at the Georgia Genomics Facility. Chromatograms were edited using SEQUENCHER 5 (Gene Codes) and aligned manually.

We genotyped 18 microsatellite loci via PCR for 30 individuals per locality except localities 21 and 23 in which only 9 loci were genotyped (if 30 samples were unavailable we genotyped all available samples). We used the published primers for the following loci: *Mf-1*, *Mf-13*, *Gafμ2*, *Gafμ3*, *Gafμ4*, *Gafμ5*, *Gafμ6*, *Gafμ7*, *Gaaf7*, *Gaaf9*, *Gaaf10*, *Gaaf11*, *Gaaf13*, *Gaaf14*, *Gaaf15*, *Gaaf16*, *Gaaf22*, and *Gaaf23* (Purcell *et al.* 2011; Spencer *et al.* 1999; Zane & Nelson 1999). We placed the CAG-tag (5'-CAGTCGGGCGTCATCA-3') on the primer specified in Purcell et al. (2011) and for all other loci we placed it on the forward primer. PCR amplifications had a 12.5 μl volume [10mM Tris-HCl, 50 mM KCl, 0.1% Triton X-100, 100 μg/mL BSA, 0.4 μM unlabeled primer, 0.04 μM tag-labeled primer, 0.36 μM universal dye-labeled primer (FAM or HEX), 4.0 mM $MgCl_2$, 0.8 mM dNTPs, 0.25 units EconoTaq DNA Polymerase (Lucigen), and ~10 ng DNA template] and were carried out on all loci using a touchdown thermal profile: 20 cycles of 96°C for 30 s, highest annealing temperature of 60°C (decreased 0.5°C per cycle) for 30 s, and 72°C for 30 s; and 20 cycles of 96°C for 30 s, 50°C for 30 s, and 72°C for 30 s with a final extension for 300 s. We multiplexed samples by combining PCR product from the following pairs of primer (the first one labeled with FAM and the second labeled with HEX): *Gafμ4-Gaaf7*, *Gaaf15-Gaaf16*, *Mf-1-Gaaf9*, *Mf-13-Gafμ*4, *Gaaf23- Gaaf14*, *Gafμ2-Gafμ3*, *Gafμ7-Gaaf22*, *Gafμ6-Gaaf10*, and *Gaaf11-Gaaf13*. Multiplexed PCR products were run on an Applied Biosystems 3730 XL sequencer and sized with a Naurox size standard (DeWoody *et al.* 2004). Peaks were scored blindly using GENEMARKER version 2.4 (SoftGenetics, State College, PA). We randomly selected ~2% of the individuals and genotyped them again for all loci. Alleles for these individuals were compared with original genotypes to estimate the scoring error rate.

Mitochondrial DNA analyses

We calculated the number of variable sites, number of parsimony informative sites, and nucleotide diversity on the mitochondrial sequences using the software program DNASP v5 (Librado & Rozas 2009). We constructed a minimum-spanning haplotype network of the cyt *b* fragments using statistical parsimony with a 95% probability that no multiple substitutions had occurred with the software program TCS v1.21 (Clement *et al.* 2000; Templeton *et al.* 1992). The network is ideally suited for looking at intraspecific variation allowing us to examine the genealogical relationships of the mitochondrial sequence haplotypes, their frequency in the data, and look for any obvious geographical patterns to their distribution.

Microsatellite analyses

Scored microsatellite alleles were inspected for scoring errors and the presence of null alleles using the software program MICROCHECKER v2.2.3 (Van Oosterhout *et al.* 2004). We used the software program POWSIM v4.1 (Ryman & Palm 2006) to test the statistical power of the microsatellite markers for our tests for genetic homogeneity. We used GENEPOP v4.2 (Raymond & Rousset 1995; Rousset 2008) to detect deviations from Hardy-Weinberg equilibrium and linkage disequilibrium with Bonferroni corrections. We also calculated observed and expected heterozygosity in ARLEQUIN v3.5 (Excoffier & Lischer 2010). We constructed a neighbor-joining tree of the localities using the allele frequencies of the microsatellite genotypes for each locality using the software package PHYLIP (Felsenstein 1989).

In order to specifically test the proposed genetic breaks, we conducted an analysis of molecular variance (AMOVA) on both the cyt *b* fragment and the microsatellites in the software package ARLEQUIN v3.5 (Excoffier & Lischer 2010; Excoffier *et al.* 1992). We grouped the localities into two groups for each proposed break as follows: (A) species boundary, *G. affinis* =

localities 1-24, *G. holbrooki* = localities 25-42; (B) Savannah River, south = localities 25-33, north = localities 34-42; (C) Mississippi River, west = localities 1-17, east = localities 18-24. If the genetic breaks are a barrier for dispersal we would expect most of the variation to be between the two groups resulting in high F_{ST} values.

We used the software program STRUCTURE (Pritchard *et al.* 2000) to estimate the number of clusters for each genetic break and determine how much admixture was occurring across each of the genetic breaks. All 18 microsatellite loci for each locality were analysed under an admixture model, assuming no correlation between alleles and using no prior information about sampling localities. The admixture model allows for mixed ancestry and is a recommended parameter when examining populations with the potential for gene flow. Twenty runs were performed for each K value (from 1 to 15), each beginning with a different random seed, each for 1,000,000 generations with a burn-in of 100,000 generations discarded. We used STRUCTURE HARVESTER to implement the Evanno method for selecting the optimal K value based on delta K values (Earl & VonHoldt 2011). We used CLUMPP to determine the most likely set of cluster membership coefficients for the optimal K value using the Greedy algorithm (Jakobsson & Rosenberg 2007) and the data were visualized in DISTRUCT (Rosenberg 2004).

Results

From the 399 individuals sequenced for the cyt *b* fragment, we found 59 unique haplotypes (Genbank accession numbers KF895041-KF895099, Table 2.2). There were a total of 104 polymorphic sites, 80 of which were parsimony informative. Nucleotide diversity was estimated at 0.02498 and total GC content was 45.3%. The haplotype network resulted in one large network that contained 54 of the haplotypes and three small networks made from the remaining 5 haplotypes (Figure 2.2). The large network is comprised to two major clades that

correspond to the two species aside from a few shared haplotypes. We report all of the haplotypes and their frequency in each locality (Table 2.3). We found a few haplotypes shared across each of the proposed genetic breaks. For the species boundary, haplotypes D1, G, H, J are shared in localities across the boundary. Haplotypes A, B, and G are shared in localities across the Savannah River and haplotypes G, I, and J are shared across the Mississippi River. All shared haplotypes except for D1 are very common and thus, assumed to be ancestral under coalescent theory (Table 2.3).

We found 2.9% of the microsatellite genotypes contained errors when repeated. Scoring errors resulting from data input error were confirmed on original peaks and corrected prior to analysis. Null alleles were detected sporadically and recorded, but no attempt was made to adjust allele frequencies. The power of the microsatellite markers to detect significant differentiation was high suggesting a probability of at least 0.89 to detect a true differentiation of $F_{ST} = 0.001$ under different scenarios of N_e and number of generations (t) with 1000 replications. We detected deviations from Hardy-Weinberg in about 25% of the tests carried out (181 deviations and 705 tests) after Bonferroni corrections, with five of the loci accounting for ~60% of the deviations. Less than 1% of the tests for linkage disequilibrium showed significance after Bonferroni corrections. Mean observed and expected heterozygosity for all *G. holbrooki* localities was 0.4633 and 0.6039, respectively. For the *G. affinis* localities, the heterozygosity values were 0.5507 and 0.6610 (see Table 2.1 for details on each locality).

The neighbor-joining tree of the localities from the microsatellite genotypes yielded a tree largely concordant with the mtDNA haplotype network (Figure 2.3). The 42 localities cluster into two main clades that correspond to the two different species exactly. The tree also shows that the localities north of the Savannah River form their own clade except for localities in South

Carolina (SREL, Lake Marion, and Combahee River) that cluster with localities south of the Savannah River. The localities on the east side of the Mississippi River are scattered throughout the *G. affinis* clade with Pascagoula River being quite dissimilar from the rest.

The AMOVA results are presented in Table 2.4. The species boundary and the Savannah River both showed concordant results between the mtDNA and microsatellite analyses. The largest portion of the variation was explained by within group differences. The Mississippi River showed discordant results between the two marker types, with the mtDNA showing the largest portion of the variation coming from among groups within localities yet the microsatellites showed the largest source of variation from within groups.

The optimal number of clusters for the three genetic breaks was two for each of the potential breaks tested (Figure 2.4). The species boundary showed two distinct clusters, which match the species ID closely. However, there were several localities that showed a fair amount of admixture with the other cluster. In particular, the Pascagoula River (Locality 23) shows ~46% admixture with *G. holbrooki*. The Savannah River also showed two distinct clusters in *G. holbrooki*, however there did not appear a clean break at the location of the Savannah River. Instead we found that several localities in North and South Carolina were admixed with localities south of the Savannah River, even a locality right on the Savannah River (SREL) showed very little signature of the northern localities. Finally, the Mississippi River had two clusters that show a pattern of admixture across the Mississippi River. All localities on either side of the Mississippi River clustered together, while most of the other cluster was made up of localities from drainages outside of the Mississippi River system.

Discussion

We investigated the genetic diversity and population structure of mosquitofish throughout their native range. In particular, we wanted to test for three specific genetic breaks likely to contribute to the population structure. We now evaluate each of these genetic breaks in turn with our results and explore the implications for reconstructing invasion histories.

Species boundary

The results for the localities we collected indicate that the species boundary follows a southwest to northeast direction following the Alabama River and its tributaries (specifically the Coosa and Tallapoosa rivers). Wooten et al. (1988) described the Mobile Bay as an area of demarcation between western and eastern forms of what was then known as *G. affinis*. They established differentiation between the two subspecies and argued for them to be considered two separate species *G. affinis* (west of Mobile Bay) and *G. holbrooki* (east of Mobile Bay). However, we included four localities (localities 21-24) east of the Mobile Bay that were morphologically *G. affinis* that clearly clustered with other *G. affinis* localities in or genetic analyses (Figures 2.3 & 2.4). All other localities east of Mobile Bay cluster with *G. holbrooki* and were morphologically identified as such. Scribner & Avise (1993) observed a similar pattern and argued this region to be a zone of sympatry. However, none of the localities that we sampled showed evidence of both species being present at the same locality based on our morphological examinations.

Hybridization between the two species has been documented (Pyke 2005), however, its prevalence in the wild has not been thoroughly studied. We found very little evidence of gene flow between the two species (Table 2.4, Figure 2.4) suggesting that where they do occur in sympatry, reproductive barriers exist to prevent or limit hybridization. However, we did find

several mitochondrial haplotypes that were shared between the two species (Table 2.3, Figure 2.2). These could indicate introgression between the two species, however, since the shared haplotypes are among the more common and ancestral haplotypes this could indicate ancestral polymorphism. Either way, it does seem clear that gene flow between the two species is very limited.

We note with exception the one locality in the Pascagoula River (locality 18) where we detected admixture between the two species. This *G. affinis* locality showed 46% admixture with *G. holbrooki* (Figure 2.4) and the mtDNA haplotypes from this locality were more closely related to *G. holbrooki* haplotypes than to the rest of *G. affinis* (Table 2.3, Figure 2.2). Others have also found this area to have *G. holbrooki* alleles at a higher frequency (Scribner & Avise 1993). A more fine-scaled study of this area would reveal the prevalence of this admixture between the two species. This area in southern Mississippi could have been a glacial refugia for *G. affinis* and *G. holbrooki*, thus the persistence of alleles from both species (Soltis *et al.* 2006).

Savannah River

Wooten et al. (1988) argued for two distinct types of *G. holbrooki* separated between the Savannah and Altamaha rivers. We found little evidence for a strong genetic break around the Savannah River. Two mitochondrial haplotypes were shared across the Savannah River (Table 2.3, Figure 2.2) and the microsatellite allele frequencies for localities within South Carolina are more similar to localities throughout Georgia (Figure 2.3). The cluster analysis showed a large degree of admixture across the Savannah River going north through South Carolina and into North Carolina (Figure 2.4). It is apparent that the northernmost populations in Virginia and North Carolina are indeed genetically distinct from populations south of the Savannah River, this could be driving the AMOVA results obtained (Table 2.4). However, no clear break is shown by

the remaining results, rather the Savannah River may indicate an area where there is admixture going on from the two groups.

Mississippi River

The Mississippi River is known to be a barrier for dispersal for both terrestrial and aquatic organisms (Soltis *et al.* 2006). This typically results in a clear east-west divide in the population structure. However, *G. affinis* does not seem to follow this pattern based upon our results. We found evidence for admixture across the Mississippi River, in fact, we found that the Mississippi River drainage localities were all quite similar to one another (Figures 2.3 & 2.4). This would suggest there has been continual gene flow throughout this region and perhaps derived from a common refugial population. However, our AMOVA results in this region for each marker type did not agree (Table 2.4). The microsatellites showed the majority of the variation explained within groups (F_{ST}) whereas the cyt *b* AMOVA had most of the variation split almost evenly among groups within localities (F_{SC}) and within groups (F_{ST}).

Our analysis of *G. affinis* revealed that localities from Texas were distinct from the rest of *G. affinis* (Figure 2.4). This could be explained by most of the drainages flowing into the Rio Grande or directly into the Gulf of Mexico, while much of the rest of *G. affinis* seems connected to the Mississippi River drainages. It may also suggest that *G. affinis* had at least two separate refugial populations that gave rise to the current distribution.

The complex phylogeographical patterns found throughout the southeastern United States vary broadly across taxa (Soltis *et al.* 2006). Mosquitofish demonstrate that the events and processes shaping genetic variation in species can be unique to each species. Mosquitofish do not seem to be influenced by many of the proposed biogeographic breaks discussed for co-distributed taxa in the southeastern United States (Bermingham & Avise 1986; Soltis *et al.*

2006). The only area we have observed that shows similarity with previously proposed breaks is the species boundary between *G. affinis* and *G. holbrooki*, which is known as a region influencing many species including freshwater fishes (Bermingham & Avise 1986).

Implications for introduced populations

This study was inspired to help facilitate the reconstruction of invasion routes for mosquitofish by understanding the population structure and genetic diversity of the native range for both of the species introduced around the world. While previous work has characterized some of what we discuss here, their data is not comparable to modern markers and analyses that are common today. We expanded the sampling of native populations beyond any previous work in order to survey a greater portion of the genetic variation that exists throughout the range. While caution should be used in using microsatellite data for different studies, the population structure and genetic diversity can help guide future studies comparing native and introduced populations of mosquitofish. Furthermore, we have demonstrated that there are distinct groups within each species and introduced populations that exhibit similar haplotypes or allele frequencies would limit the geographic range of potential source populations.

The utility of this study can be demonstrated by highlighting a study of introduced *G. holbrooki* populations in Melbourne, Australia (Ayres *et al.* 2010). The authors sampled extensively throughout the city and found all the specimens had the same mitochondrial haplotype. However, without any native specimens they were unable to make any conclusions regarding where in the native range they came from. The haplotype they observed matched with haplotype F reported in this study, which is found only in localities 39 and 42 of this study. Australian mosquitofish were introduced from European populations that were said to have come from Augusta, Georgia (Lloyd & Tomasov 1985). However, our data failed to observe this

haplotype in localities near Augusta. Instead, the occurrence of this Australian haplotype in localities in North Carolina and Virginia supports the growing amount of genetic data demonstrating that European populations were the result of multiple introductions with at least one source somewhere in North Carolina and Virginia (Sanz *et al.* 2013; Vidal *et al.* 2009).

References

Avise J, Arnold J, Ball R*, et al.* (1987) Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual review of ecology and systematics* **18**, 489-522.

Ayres RM, Pettigrove VJ, Hoffmann Aa (2010) Low diversity and high levels of population genetic structuring in introduced eastern mosquitofish (*Gambusia holbrooki*) in the greater Melbourne area, Australia. *Biological Invasions* **12**, 3727-3744.

Beheregaray LB (2008) Twenty years of phylogeography: the state of the field and the challenges for the Southern Hemisphere. *Molecular Ecology* **17**, 3754-3774.

Bermingham E, Avise JC (1986) Molecular zoogeography of freshwater fishes in the southeastern United States. *Genetics* **113**, 939-965.

Clement M, Posada D, Crandall Ka (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology* **9**, 1657-1659.

Coyne JA, Orr HA (2004) *Speciation* Sinauer Associates, Inc Sunderland, MA.

DeWoody JA, Schupp J, Kenefic L*, et al.* (2004) Universal method for producing ROX-labeled size standards suitable for automated genotyping. *BioTechniques* **37**, 348-352.

Earl DA, VonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359-361.

Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data: why, how and so what? *Molecular Ecology* **19**, 4113-4130.

Excoffier L, Lischer HEL (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.

Excoffier L, Smouse P, Quattro J (1992) Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Applications to Human Mitochondrial DNA Restriction Data. *Genetics* **491**, 479-491.

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166.

Glenn TC, Schable Na (2005) Isolating microsatellite DNA loci. *Methods in enzymology* **395**, 202-222.

Hickerson MJ, Carstens BC, Cavender-Bares J*, et al.* (2010) Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution* **54**, 291-301.

Hillis DM, Mable BK, Larson A, Davis SK, Zimmer EA (1996) Nucleic acids IV: sequencing and cloning. In: *Molecular systematics, 2nd edition* eds. Hillis DM, Moritz C, Mable BK), pp. 321-381. Sinauer Associates, Inc, Sunderland, Massachusetts.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Krumholz L (1948) Reproduction in the western mosquitofish, *Gambusia affinis affinis* (Baird & Girard), and its use in mosquito control. *Ecological Monographs* **18**, 1-43.

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

Lloyd L, Tomasov J (1985) Taxonomic status of the mosquitofish, *Gambusia affinis* (Poeciliidae), in Australia. *Australian Journal of Marine and Freshwater Research* **36**, 447-451.

Lowe S, Browne M, Boudjelas S, De Poorter M (2000) 100 of the world's worst invasive species. A selection from the Global Invasive Species Database. The Invasive Species Specialist Group (ISSG) a sepcialist group of the Species Survival Commission (SSC) of the World Conservation Union (IUCN).

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

Purcell KM, Lance SL, Jones KL, Stockwell Ca (2011) Ten novel microsatellite markers for the western mosquitofish *Gambusia affinis*. *Conservation Genetics Resources* **3**, 361-363.

Purcell KM, Ling N, Stockwell Ca (2012) Evaluation of the introduction history and genetic diversity of a serially introduced fish population in New Zealand. *Biological Invasions* **14**, 2057-2065.

Pyke GH (2005) A Review of the Biology of *Gambusia affinis* and *G. holbrooki*. *Reviews in Fish Biology and Fisheries* **15**, 339-365.

Pyke GH (2008) Plague Minnow or Mosquito Fish? A Review of the Biology and Impacts of Introduced *Gambusia* Species. *Annual Review of Ecology, Evolution, and Systematics* **39**, 171-191.

Rauchenberger M (1989) Systematics and Biogeography of the Genus *Gambusia* (Cyprinodontiformes: Poecilidae). *American Museum Novitates* **2951**, 1-76.

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

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**, 137-138.

Rousset F (2008) genepop, '007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.

Ryman N, Palm S (2006) POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Molecular Ecology Notes* **6**, 600-602.

Sanz N, Araguas RM, Vidal O*, et al.* (2013) Genetic characterization of the invasive mosquitofish (*Gambusia* spp.) introduced to Europe: population structure and colonization routes. *Biological Invasions* **15**, 2333-2346.

Scribner K, Avise J (1993) Cytonuclear genetic architecture in mosquitofish populations and the possible roles of introgressive hybridization. *Molecular Ecology* **2**, 139-149.

Soltis DED, Morris ABA, McLachlan JS, Manos PS, Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology* **15**, 4261-4293.

Spencer C, Chlan C, Neigel J (1999) Polymorphic microsatellite markers in the western mosquitofish, *Gambusia affinis*. *Molecular Ecology* **8**, 157-168.

Stockwell CA, Henkanaththegedara SM (2011) Evolutionary conservation biology. In: *Ecology and Evolution of Poeciliid Fishes* eds. Evans JP, Pilastro A, Schlupp I), pp. 128-141. The University of Chicago Press, Chicago.

Taberlet P, Fumagalli L, Wust-Saucy A-G, Cosson J-F (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology* **7**, 453-464.

Templeton A, Crandall K, Sing C (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**, 619-633.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) Micro-Checker: Software for Identifying and Correcting Genotyping Errors in Microsatellite Data. *Molecular Ecology Notes* **4**, 535-538.

Vidal O, García-Berthou E, Tedesco Pa, García-Marín J-L (2009) Origin and genetic diversity of mosquitofish (*Gambusia holbrooki*) introduced to Europe. *Biological Invasions* **12**, 841-851.

Wooten M, Scribner K, Smith M (1988) Genetic Variability and Systematics of *Gambusia* in the Southeastern United States. *Copeia* **1988**, 283-289.

Zane L, Nelson W (1999) Microsatellite assessment of multiple paternity in natural populations of a live-bearing fish, *Gambusia holbrooki*. *Journal of Evolutionary Biology* **12**, 61-69.

Table 2.1 – Sampling localities included in this study. Label and locality names correspond with those in Figure 2.1 and are consistent throughout the text. Number of individuals sequenced/genotyped (N) is provided along with latitude and longitude. Summary statistics for the locality based upon 18 microsatellite loci as calculated in ARLEQUIN: average number of alleles (N_a), observed heterozygosity (H_o), and expected heterozygosity (H_e).

Label	Locality	N	Latitude	Longitude	N_a	H_o	H_e
1	Alamito Creek	10/20	29.52	-104.30	7.2	0.58	0.69
2	San Felipe Creek	10/30	29.37	-100.88	6.4	0.35	0.57
3	Pine Gully	10/30	29.59	-95.00	11.7	0.58	0.79
4	Johnson Creek	10/30	30.15	-99.34	6.9	0.54	0.65
5	South Concho River	10/30	31.21	-100.50	8.4	0.59	0.73
6	North Bosque River	10/30	32.25	-98.23	8.7	0.52	0.73
7	Oakbrook Park	10/30	33.15	-96.81	3.6	0.50	0.51
8	Sanders Creek	10/30	33.87	-95.54	9.2	0.63	0.71
9	Pennington Creek	10/30	34.26	-96.68	8.4	0.53	0.70
10	Red River	10/30	34.86	-99.51	7.6	0.54	0.64
11	Turkey Creek	10/30	35.35	-96.69	7.6	0.53	0.66
12	Pecan Creek	10/30	35.91	-95.12	6.2	0.49	0.64
13	Clarke Bayou	10/30	32.57	-93.49	9.9	0.64	0.73
14	Bayou Macon	10/30	32.45	-91.46	9.2	0.58	0.65
15	Little Missouri River	10/30	34.05	-93.72	7.4	0.61	0.62
16	Brodie Creek	10/30	34.71	-92.38	6.7	0.57	0.60
17	Little Red River	10/30	35.82	-92.55	4.8	0.50	0.54
18	Pascagoula River	10/30	31.34	-89.41	8.1	0.49	0.69
19	Big Black River	10/30	33.38	-89.61	10.0	0.59	0.70
20	Reelfoot Lake	10/30	36.40	-89.34	8.0	0.66	0.65
21	Hillabee Creek	-/15	32.99	-85.86	5.2	0.46	0.59
22	Roebuck Spring Run	10/27	33.58	-86.71	5.2	0.50	0.58
23	James Creek	-/21	33.91	-86.96	4.6	0.55	0.63
24	Conasauga River	10/30	34.68	-84.94	9.1	0.60	0.74
25	Smilies Mill Creek	10/30	31.71	-86.06	3.3	0.31	0.38
26	Canoe Creek	10/30	27.20	-80.30	9.8	0.53	0.73
27	Field Building	10/30	28.59	-81.19	11.4	0.62	0.79
28	Digital Design Wetlands	10/30	29.64	-82.35	10.4	0.60	0.76
29	Altamaha River	10/30	31.67	-81.85	11.6	0.57	0.78
30	Lake Blackshear	10/30	31.85	-83.92	8.7	0.45	0.64
31	Ocmulgee River	9/30	32.00	-83.29	7.7	0.44	0.62
32	Oconee River	10/30	33.13	-83.20	6.4	0.44	0.69
33	Lake Herrick	10/30	33.93	-83.38	4.1	0.35	0.41
34	SREL	10/30	33.34	-81.73	4.4	0.38	0.46
35	Combahee River	10/30	32.71	-80.83	8.0	0.54	0.61
36	Lake Marion	10/30	33.57	-80.44	9.8	0.45	0.68
37	Lumber River	10/30	34.39	-79.00	9.6	0.46	0.71
38	Burnt Mill Creek	10/30	34.23	-77.90	7.6	0.57	0.65
39	Reedy Creek	10/27	36.42	-78.12	3.9	0.35	0.43
40	Herring Creek	10/30	37.33	-77.16	5.1	0.45	0.45
41	Piscatawny Creek	10/30	37.87	-76.85	6.2	0.42	0.56
42	Potomac Creek	10/30	38.36	-77.39	6.0	0.41	0.53

Table 2.2 – A list of the unique haplotypes observed in this study along with their corresponding Genbank accession number. Haplotype labels match those used in Figure 2.2 and throughout the text.

Haplotype	Genbank accesion no.
A	KF895041
A1	KF895042
A2	KF895043
A3	KF895044
A4	KF895045
A5	KF895046
A6	KF895047
A7	KF895048
A8	KF895049
A9	KF895050
B	KF895051
B1	KF895052
B2	KF895053
B3	KF895054
B4	KF895055
B5	KF895056
B6	KF895057
B7	KF895058
C	KF895059
C1	KF895060
C2	KF895061
D	KF895062
D1	KF895063
D2	KF895064
E	KF895065
E1	KF895066
E2	KF895067
E3	KF895068
E4	KF895069
F	KF895070
F1	KF895071
F2	KF895072
F3	KF895073
G	KF895074
G1	KF895075
G2	KF895076
G3	KF895077
G4	KF895078
G5	KF895079
G6	KF895080
G7	KF895081
G8	KF895082
G9	KF895083
H	KF895084
H1	KF895085
H2	KF895086
I	KF895087
I1	KF895088
J	KF895089
J1	KF895090
K	KF895091
K1	KF895092
K2	KF895093
L	KF895094
M	KF895095
M1	KF895096
N	KF895097
O	KF895098
O1	KF895099

Table 2.3 – Haplotype table detailing the number of individuals sequenced for cytochrome *b* at each locality and each haplotype occurring at each locality.

Locality #	A	A1	A2	A3	A4	A5	A6	A7	A8	A9	B	B1	B2	B3	B4	B5	B6	B7	C	C1
1																				
2																				
3																				
4																				
5																				
6																				
7																				
8																				
9																				
10																				
11																				
12																				
13																				
14																				
15																				
16																				
17																				
18																				
19																				
20																				
21																				
22																				
23																				
24																				
25													8							
26											1									
27	1										1					2				
28											8							1		
29	8											1					1			
30	1										2			2	3	2				
31												4		6						
32										7										
33	10																			
34	10																			
35	8	1									1									
36	6																			
37	2	4					2	2												
38		2	2	1	1															
39	1					1														
40																			7	3
41									1										3	2
42																				

Table 2.3 (continued)

Locality #	C2	D	D1	D2	E	E1	E2	E3	E4	F	F1	F2	F3	G	G1	G2	G3	G4	G5
1																			
2														1					
3														6					
4														1					
5														2			1		
6														6					
7																			
8														10					
9																			
10														1	1				
11														1					
12														5		1			
13														4					
14																			
15														5					
16														8					
17														10					
18		10																	
19																			
20														9				1	
21																			
22																			
23																			
24			1																
25			1											1					
26					6		1	1	1										
27					4	1	1												
28					1														
29																			
30																			
31																			
32			1	1															
33																			
34																			
35																			
36												1		1					
37																			
38														1					1
39										7									
40																			
41	4																		
42										5	1		3						

Table 2.3 (continued)

Locality #	G6	G7	G8	G9	H	H1	H2	I	I1	J	J1	K	K1	K2	L	M	M1	N	O	O1
1					7			1		2										
2																		1	7	1
3	3							11							10					
4					2			1		1			4	1						
5																7				
6					2							2								
7					10															
8																				
9								7		3										
10								2		1		5								
11					6							3								
12								4												
13								6												
14								8		2										
15		3	1	1																
16								2												
17																				
18																				
19								9	1											
20																				
21																				
22										9	1									
23																				
24								3		6										
25																				
26																				
27																				
28																				
29																				
30																				
31																				
32																				
33																				
34																				
35																				
36					1												1			
37																				
38							1			1										
39					1															
40																				
41																				
42						1														

Table 2.4 – Analysis of molecular variance (AMOVA) results for each of the three genetic breaks tested. For each source of variation at each marker, we report the percent of variation along with the corresponding F-statistic. All F-statistics were significant ($p<0.001$) except those indicated by an asterisk.

Genetic break	Marker	Among localities (F_{CT})		Among groups within localities (F_{SC})		Within groups (F_{ST})	
Species boundary	cyt b	10.19%	0.10193	41.68%	0.46414	48.12%	0.51876
	usat	15.18%	0.15181	20.78%	0.24494	64.04%	0.35957
Savannah River	cyt b	0.85%	0.00847*	40.87%	0.41224	58.28%	0.41721
	usat	5.81%	0.05813	24.80%	0.26331	69.39%	0.30614
Mississippi River	cyt b	2.12%	0.02117*	49.28%	0.50341	48.61%	0.51392
	usat	1.19%	0.12610*	23.49%	0.23769	75.33%	0.24675

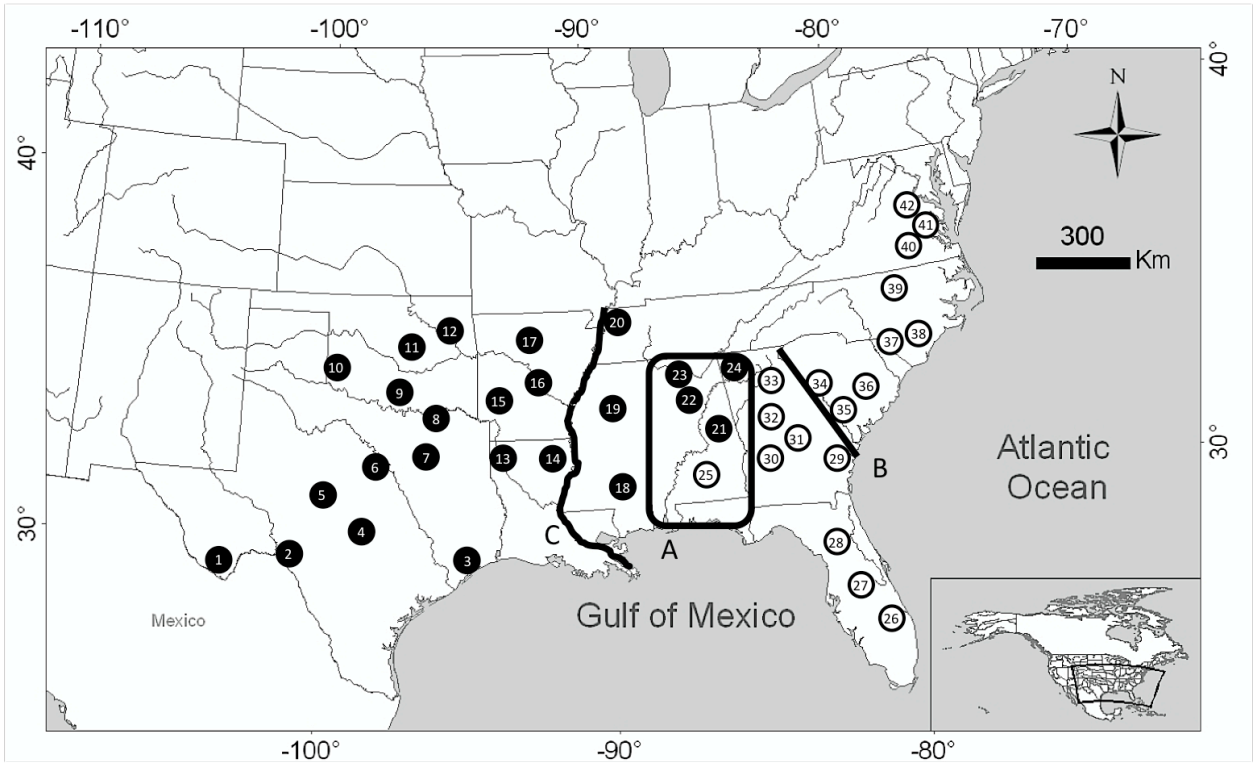

Figure 2.1 – Map of the southeastern United States indicating the location of each of the 42 sampled localities with numbered circles. Black circles indicate localities that were identified as *Gambusia affinis* and white circles indicate *Gambusia holbrooki* localities. The three genetic breaks being tested are also marked on the map with black lines and labeled A, B, and C corresponding to their description in the text.

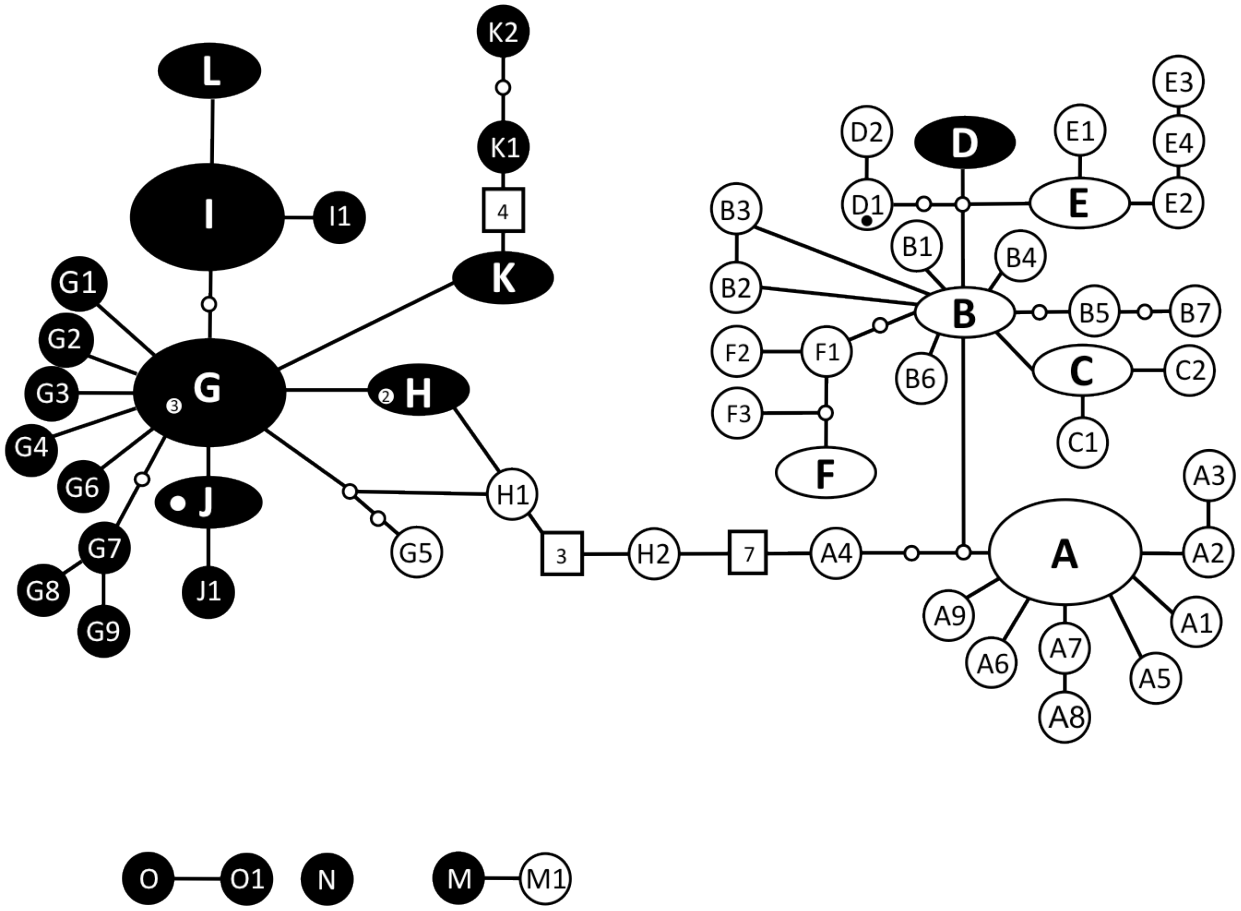

Figure 2.2 – Haplotype network generated from 547-bp sequences of the mitochondrial gene cytochrome b. Table 2.3 follows the same labels and gives specific information on frequency of each haplotypes in each locality. Black ovals/circles indicate *G. affinis* haplotypes and white circles indicate *G. holbrooki* haplotypes. Shared haplotypes between the species is indicated with a small, black or white circle inside the larger oval/circle with a number indicating how many individuals have that haplotype, if no number is present only a single individual shared that haplotype. Size of the oval/circles indicates frequency at which it was found in the data (small circle = 1-9 individuals, medium oval = 10-29 individuals, large ovals = 30 or more individuals).

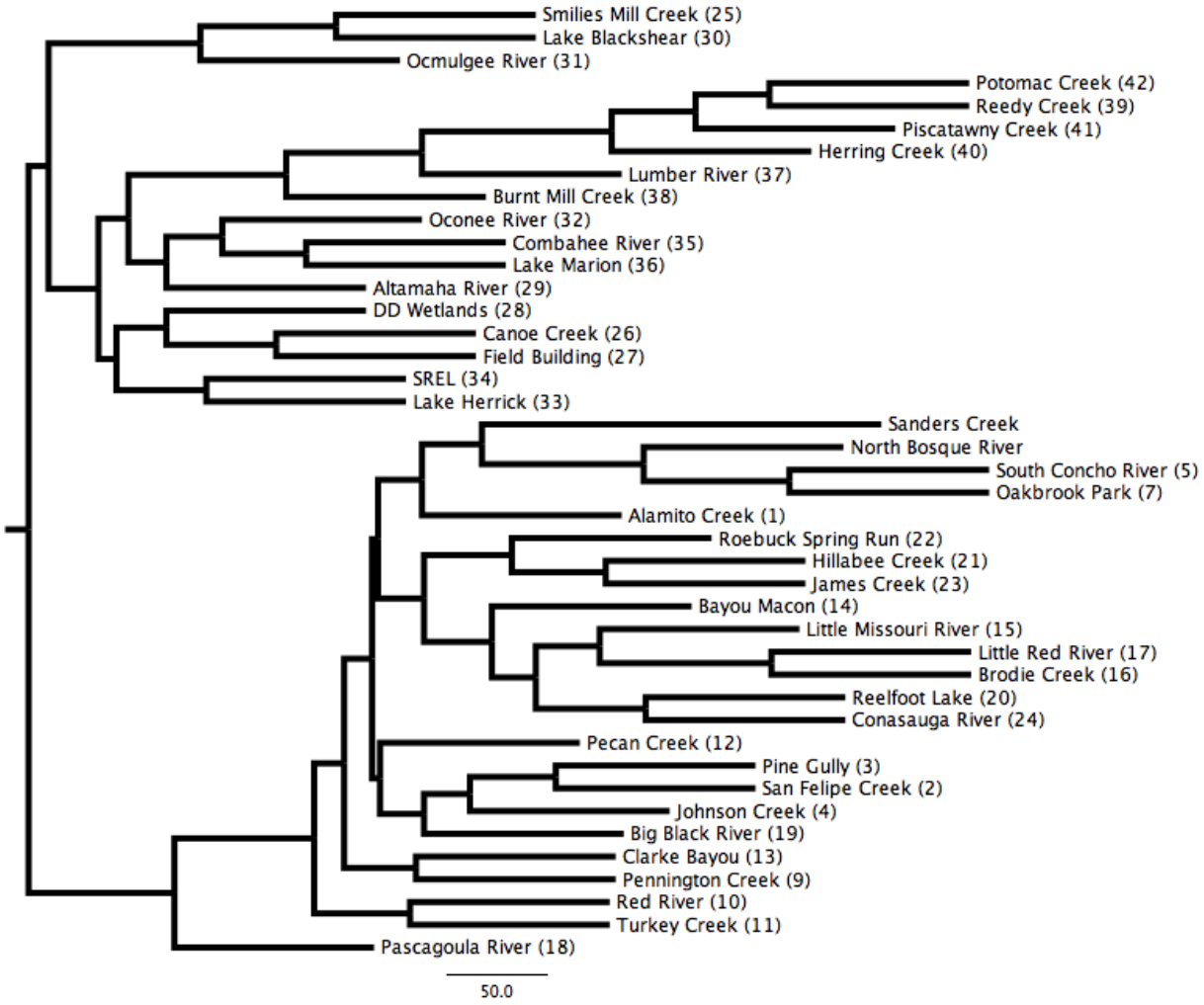

Figure 2.3 – Neighbor-joining tree rooted at the mid-point of the 42 localities based upon allele frequencies of 18 microsatellite markers. Tip labels include name of each locality and the locality number from Figure 2.1.

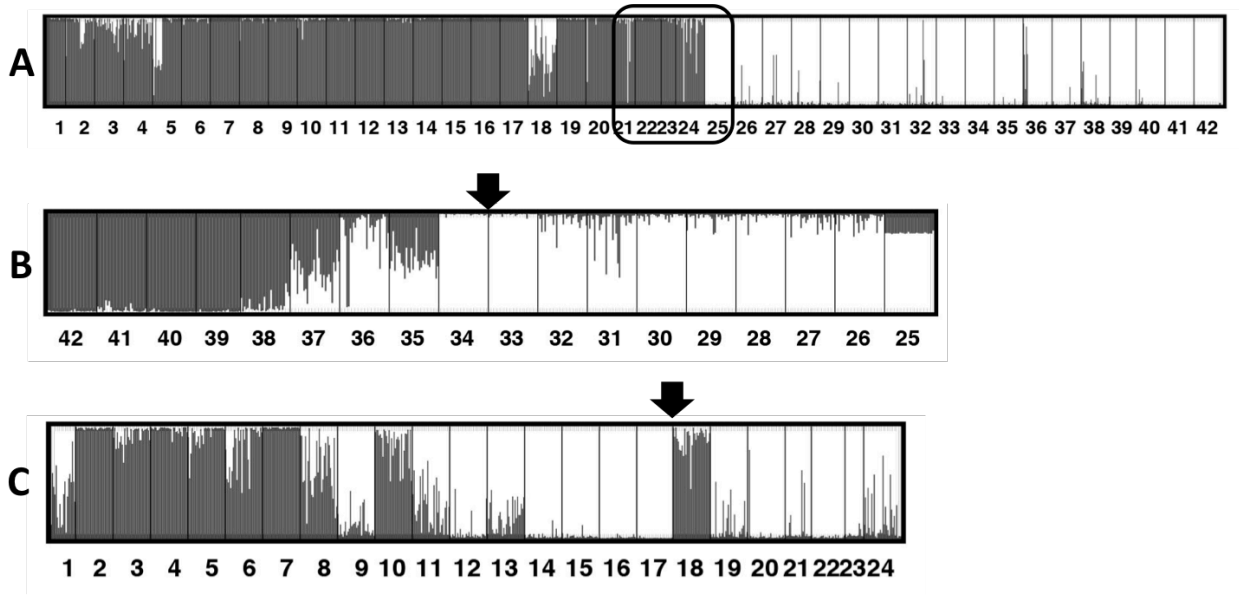

Figure 2.4 – Cluster plots generated by DISTRUCT from 20 runs in the program STRUCTURE for each of the genetic breaks (A= species boundary, B= Savannah River, C= Mississippi River). Numbers below indicate the locality numbers found in Figure 2.1. The box in A corresponds to the localities within the zone of sympatry depicted in Figure 2.1. The arrow above B and C indicate the putative location for the genetic break.

CHAPTER 3: RECONSTRUCTING THE INVASION HISTORY OF *GAMBUSIA AFFINIS* INTO ASIA USING HISTORICAL AND GENETIC DATA[1]

[1] Lee JB and Mauricio R. To be submitted to *Biological Invasions*.

Abstract

Reconstructing the invasion history of an invasive species allows us to understand the route by which they were introduced, estimate the size of their introductions, and identify source populations. Mosquitofish, *Gambusia affinis,* were intentionally introduced into Hawaii as early as 1905 and then spread from there throughout Taiwan, the Philippines, Japan, and China over the next few decades. With this historical backdrop, we reconstruct the invasion history of *G. affinis* using a suite of microsatellite markers and a sequenced fragment of the mitochondria for 20 localities throughout Asia. We found a decrease in the number of haplotypes present and heterozygosity compared to the native range. However, our tests for a recent bottleneck were negative suggesting that the introductions could have been large or have had sufficient time to recover. We assigned 19 of the localities back to a single native population and also found a mitochondrial haplotype unique to that locality that was found in ~73% of the individuals from the introduced range. This native population is the closest sampled locality to the recorded source population. Surprisingly, our results demonstrate that the historical record for mosquitofish introductions to Asia is quite complete and accurate. Mosquitofish introduced to Asia were likely the result of a single introduction event from the recorded source population near Seabrook, Texas. As a popular mosquito control agent in the early 1900s, they were most likely moved around in large numbers allowing them to establish and spread rapidly.

Introduction

An important first step in studying invasions is reconstructing the invasion history of the organism (Estoup & Guillemaud 2010). Invasion histories give us important information regarding the number of introductions, source populations, and the route by which they arrived. With an understanding of the invasion history, studies can be designed that compare native and introduced populations to address mechanisms that make the organism a successful invader (Hierro *et al.* 2005), compare phenotypic shifts in the introduced range from the native range (Brown *et al.* 2007), and develop management strategies for control (Ayala *et al.* 2007). Information from these projects is more robust when the invasion history is well understood and help protect native species threatened from invaders (Allendorf & Lundquist 2003; Sakai *et al.* 2001).

Studies utilize two types of methods used to reconstruct invasion histories, direct and indirect methods (Estoup & Guillemaud 2010). Direct methods typically refer to historical records or other current observations, which can include published accounts, government reports, museum records, harbor/airport inspection records, or other documentation. This information is often available for intentional introductions, where a government or other organized group has managed the introductions. Conversely, accidental introductions will likely have sparse documentation until resource managers or museum field collectors detect the invasive populations. Regardless of how much documentation is available, such records may be unreliable, incomplete or conflicting with other records (Tsutsui & Suarez 2001). Indirect methods use molecular markers from native and introduced populations, which are then analyzed in a statistical framework (Ciosi *et al.* 2008; Facon *et al.* 2003; Lindholm *et al.* 2005). Genetic diversity in both ranges can be directly compared and inferences made regarding the invasion

history (Barun *et al.* 2013; Fitzpatrick *et al.* 2012). Studies using indirect methods have helped establish that, contrary to an earlier paradigm sometimes referred to as a 'genetic paradox' (Allendorf & Lundquist 2003), invasive species actually harbor much of the genetic diversity from the native range as a result of multiple introductions and/or large numbers of founders (Dlugosch & Parker 2008). Thus, indirect methods have added much to our understanding of invasion histories especially for species with little documentation of the introduction.

In the early 20th century, mosquitofish (*Gambusia affinis* and *G. holbrooki*), native to the southeastern United States, were promoted as the solution to mosquito-born diseases (*i.e.,* malaria, yellow fever) and intentionally introduced around the world (Krumholz 1948; Pyke 2008). Mosquitofish established quickly in all areas it was introduced, grew in population size, and expanded their range in the new environments. Its use as a mosquito control agent is debated, but its negative environmental impacts are clearly documented and is sometimes referred to as a 'plague minnow' (Pyke 2008; Stockwell & Henkanaththegedara 2011). Indeed, it has become a pest species throughout its introduced range, which includes all continents except Antarctica, and is considered one of the worst invasive species in the world (Lowe *et al.* 2000). In recent years, several studies have reconstructed the invasion history of *G. holbrooki* into Europe and Australia (Ayres *et al.* 2012; Ayres *et al.* 2010; Sanz *et al.* 2013; Vidal *et al.* 2009; Vidal *et al.* 2012). However, only one study has explored the invasion of *G. affinis* in New Zealand (Purcell *et al.* 2012), leaving other introduced regions unstudied.

In this study, we reconstruct the invasion history of *G. affinis* throughout Asia using both direct and indirect methods. Since introductions of mosquitofish were quite popular in the early 20th century, we expected to find some documentation of their introduction, but also figured many introductions may have gone unrecorded. Our goal was to compare results from both

methods to develop an accurate picture of the invasion history. Specifically, we wanted to address several questions: (1) How much genetic diversity persists in the introduced range compared to the native range? (2) Was the introduction into Asia the result of a single or multiple introduction events? (3) Was there only one source population? (4) Is there evidence for a bottleneck to have occurred during the introductions throughout Asia?

Materials & Methods

Literature review

We sought out historical documentation of the mosquitofish introductions throughout Asia. Our search included scientific journals, government reports, and consultation with researchers in Asia familiar with invasive species. We consulted documents in English, Chinese, and Japanese to piece together any account of the movement of mosquitofish throughout Asia.

Sampling strategy

We collected mosquitofish from introduced localities in Hawaii, Taiwan, the Philippines, Japan, and China resulting in a total of 20 localities from the introduced range. Fish were provided by collaborators or sampled directly by the first author using a dipnet. All fish were preserved in 100% alcohol prior to DNA extraction. We also used the 24 *G. affinis* localities from the native range in Chapter 2, which includes a locality collected as close to the recorded putative source population as can be determined in Seabrook, Texas (Locality 3, Pine Gully). We have kept the labeling of the native localities the same as Chapter 2 for consistency and labeled the introduced localities 25-44 (Table 3.1, Figure 3.1; see also Chapter 2 Figure 2.1 for map of native localities). We identified the species by examining the morphology of the gonopodium on all mature males in a locality (Rauchenberger 1989).

Laboratory protocols

DNA extractions, mitochondrial DNA sequencing, and microsatellite genotyping protocols followed those detailed in Chapter 2 with the following modifications. Since Pine Gully is the putative source population we sequenced an additional 20 individuals in order to get an accurate estimate of the haplotype frequency in this locality. Moreover, Kualoa was the only locality we were able to obtain for Hawaii and since it represents a key intermediate introduction we sequenced an additional 19 individuals.

Mitochondrial DNA analyses

We calculated the number of variable sites, number of parsimony informative sites, and nucleotide diversity on the mitochondrial sequences using the software program DNASP v5 (Librado & Rozas 2009). We constructed a minimum-spanning haplotype network of the cyt *b* fragments for the introduced individuals using statistical parsimony with a 95% probability that no multiple substitutions had occurred with the software program TCS v1.21 (Clement *et al.* 2000; Templeton *et al.* 1992). We compared the haplotypes to those obtained in Chapter 2 to determine how many persisted in the introduced range and if any novel haplotypes were observed.

Microsatellite analyses

Scored microsatellite alleles were inspected for scoring errors and the presence of null alleles using the software program MICROCHECKER v2.2.3 (Van Oosterhout *et al.* 2004). We used the software program POWSIM v4.1 (Ryman & Palm 2006) to test the statistical power of the microsatellite markers for our tests for genetic homogeneity. We used GENEPOP v4.2 (Raymond & Rousset 1995; Rousset 2008) to detect deviations from Hardy-Weinberg equilibrium and linkage disequilibrium with Bonferroni corrections. We also calculated observed and expected

heterozygosity in ARLEQUIN v3.5 (Excoffier & Lischer 2010). We constructed a neighbor-joining tree of the native and introduced localities from the allele frequencies of the microsatellite genotypes for each locality using the software package PHYLIP (Felsenstein 1989).

We used the software program STRUCTURE (Pritchard *et al.* 2000) to estimate the number of clusters in the native and introduced range combined and also in the introduced range alone. All 18 microsatellite loci for each locality were analyzed under an admixture model, assuming no correlation between alleles and using no prior information about sampling localities. Twenty runs were performed for each K value (from 1 to 15), each beginning with a different random seed, each for 1,000,000 generations, and with a burn-in of 100,000 generations discarded. We used STRUCTURE HARVESTER to implement the Evanno method for selecting the optimal K value based on delta K values (Earl & VonHoldt 2011). We used CLUMPP to determine the most likely set of cluster membership coefficients for the optimal K value using the Greedy algorithm (Jakobsson & Rosenberg 2007) and the data were visualized in DISTRUCT (Rosenberg 2004).

We implemented the assignment test in GENECLASS2 (Piry *et al.* 2004) using the microsatellite loci to determine the putative source population for the introduced localities. We used 22 localities from the native range as a baseline to assign each native and introduced locality (localities 21 and 23 were excluded since only 9 loci were available for them). We performed all assignment likelihood tests under the Bayesian criterion (Rannala & Mountain 1997).

Reduced genetic diversity does not always mean a genetic bottleneck has occurred. We tested for a recent bottleneck (within the last $4N_e$ generations) in each of the introduced localities using the program BOTTLENECK v1.2 (Piry *et al.* 1999). Effective population size (N_e) estimates from microsatellite variation in freshwater fishes suggest that this time frame would include the

introductions of the early 20th century (DeWoody & Avise 2000). This program allowed us to implement two measures of founder effects. First, we test for a major change in allele frequencies by testing for deviations from an L-shaped distribution of allele frequencies. Under mutation-drift equilibrium populations are expected to have a large number of low frequency alleles (resulting in the L-shaped distribution). However, a recent founder event will eliminate many of the rare alleles and show more evenly distributed allele frequencies. Second, we tested for heterozygosity excess under all three models of microsatellite mutation [infinite alleles model, IAM; two-phase model, TPM (70% SMM and 30% variance); and step-wise mutation model, SMM]. The TPM and SMM are more suitable mutational models for microsatellites however, it is recommended to use all of them for comparison (Luikart & Cornuet 1998). Statistical significance of the results of each model was tested using a Wilcoxon test.

Results

Historical account

The historical record of the introduction of *G. affinis* throughout Asia details a series of introductions as it made its way through the Pacific and into China. At least 150 mosquitofish were collected in Seabrook, Texas (near Galveston) and transported to Honolulu, Hawaii in 1905 (Jordan 1927; Seale 1905; Seale 1917). All accounts report that the fish thrived in Hawaii and were spread throughout the islands, moreover they became the source for further introductions (Seale 1917). In 1911, mosquitofish from Hawaii were introduced to Taiwan (Jordan 1927; Xie *et al.* 2010; Yan *et al.* 2001). Twenty-four mosquitofish from Hawaii were transported to the Philippine Islands in 1913 and released in the capital city of Manila (Seale 1917), another introduction from Hawaii to Manila is recorded but no date is provided (Jordan 1927). Japan received mosquitofish from Taiwan in 1916 (Koya *et al.* 1998). Finally, two separate sources for

introductions of mosquitofish into China are recorded both lacking in the number of individuals introduced. The first came from Taiwan in 1924 and has no record of the location they were introduced (Yan *et al.* 2001). Another source describes introductions from the Philippines to Shanghai in 1927 and into Guangzhou in the 1960s (Pan *et al.* 1980). While not absolutely complete, this historical record will provide a useful comparison with the results from molecular markers.

Mitochondial DNA

From the 219 introduced individuals sequenced for the cyt *b* fragment, we found 6 unique haplotypes (Genbank accession no. KF895074, KF895087, KF895094, KF895100-KF895102). There were a total of 6 polymorphic sites, 5 of which were parsimony informative. Nucleotide diversity was estimated at 0.00229 and total GC content was 44.8%. Of the six haplotypes observed, three were identical to haplotypes G, I, and L from Chapter 2. The haplotypes produced one network that showed the same relationships of G, I, and L in Chapter 2 (Figure 3.2 and Chapter 2 Figure 2.2 for comparison). We use the same labels for simplicity and further labeled the other three haplotypes G10, G11, and G12 since they connected to haplotype G. The striking feature of these haplotypes is that 159 individuals (72.6%) throughout the introduced range had haplotype L and was found in all introduced regions sampled (Hawaii, Taiwan, Japan, the Philippines, and China) but not in all localities (Table 3.2). Haplotype L was found in only one locality in the native range, the putative source locality Pine Gully (Table 3.3).

Microsatellite results

We found 2.9% of the microsatellite genotypes contained errors when repeated. The software program MICROCHECKER detected scoring errors resulting from data input error and were confirmed on original peaks. Null alleles were detected sporadically and recorded, but no

attempt was made to adjust allele frequencies. The power of the microsatellite markers to detect significant differentiation was high suggesting a probability of at least 0.95 to detect a true differentiation of $F_{ST} = 0.001$ under different scenarios of N_e and number of generations (*t*) with 1000 replications. We detected deviations from Hardy-Weinberg in about 23% of the tests carried out (75 deviations and 323 tests) after Bonferroni corrections, with five of the loci accounting for ~60% of the deviations. Less than 1% of the tests for linkage disequilibrium showed significance after Bonferroni corrections. Mean observed and expected heterozygosity for all introduced localities was 0.4187 and 0.5385, respectively (Table 3.1).

The neighbor-joining tree of the localities revealed two distinct clades. The native localities showed the same relationships as previously observed (Chapter 2), however, the introduced localities all clustered together in one clade with two of the native localities from Texas (San Felipe Creek and Pine Gully, Figure 3.3).

Two distinct clusters (k=2) were determined to be optimal when both ranges were analyzed and for just the introduced range (Figure 3.4). For the combined dataset, the two clusters are broadly divided into the native and introduced range. However, the same two native populations described in the neighbor-joining tree above showed the most admixture with the introduced range. The introduced dataset showed two clusters with five of the localities (26, 37, 38, 39, and 44) belonging to one group and significant admixture detected in two other localities (36 & 40) while the remaining localities clustered into the other group (Figure 3.4B).

Using 22 native localities as a baseline, GENECLASS2 assigned all of the native localities correctly to their native source. Of the twenty introduced localities, 19 were assigned to Pine Gully and one was assigned to Alamito Creek with a probability score of at least 99.9% (Table 3.1).

We observed differences in the results for heterozygosity excess depending on the underlying mutational model (Table 3.1). Under the IAM, 8 native localities showed significant heterozygosity excess while 11 introduced localities showed significant heterozygosity excess. Under the TPM, only two localities (one native and one introduced) showed significant results. Finally, under the SMM none of the localities showed significant results. The IAM is argued to detect bottlenecks better, but has shown Type I errors with microsatellite data (Luikart & Cornuet 1998). Furthermore, all populations showed a normal L-distribution in allele frequency suggesting no evidence for a bottleneck except for the introduced locality from Guilin. Only 10 individuals were available for this locality, which is low for microsatellite markers (Hale *et al.* 2012) and the smallest sample size in this study, thus the allele frequency shift may be the result of the inadequate sampling in this locality.

Discussion

We proposed to better understand the invasion history of *G. affinis* in Asia by comparing native and introduced populations with both historical records and molecular markers. Our sampling included localities from across the entire native range of *G. affinis* and throughout the major regions in Asia where mosquitofish are established. This large dataset of both native and introduced samples allows us to compare the results and reconstruct the invasion history. We address each of the main goals of this project in turn.

Genetic diversity in the introduced range compared to the native range

The introduced localities throughout Asia overall show reduced genetic diversity from that in the native range. Only three of the 24 haplotypes found in the native range persist in Asia. Haplotypes G and I are quite common throughout the native range, however, haplotype L was sampled only in Pine Gully, which is also the putative source for mosquitofish introductions to

Asia based on historical records (Seale 1905; Seale 1917). This haplotype occurred at a frequency of 33% in Pine Gully, furthermore, Pine Gully also contained haplotypes I and G at frequencies of 37% and 20% respectively (Table 3.2). The three remaining haplotypes detected in the introduced range were not sampled in the native range. The new haplotypes are only 1-2 mutational steps away from haplotype G with two occurring at a low frequency (G11 and G12), suggesting they could be the result of new mutations having arisen after the introduction. However, more exhaustive sampling in the native range would be the only way to confirm this hypothesis. Moreover, the introduced localities showed reduced genetic diversity for the microsatellites as shown by average number of alleles, observed heterozygosity, and expected heterozygosity. Indeed, Pine Gully showed some of the highest genetic diversity for all of the native localities, however the introduced localities showed reductions not only from Pine Gully, but from the averages for the native localities as a whole. Our results are consistent with a scenario of serial introductions where the initial introduction to Hawaii shows some of the most diversity and subsequent introductions show decreasing amounts of genetic diversity with the lowest being the furthest introduction (locality 44) from the putative source.

Source and number of introductions

Our results identified a source population corresponding to Pine Gully, which we sampled as our putative source population based on the historical record. Pine Gully consistently showed evidence of being a source population across analyses. Pine Gully clustered with all of the introduced populations by similarity (Figure 3.3) and showed the most admixture (50.9%) with the introduced range in our Bayesian clustering analysis (Figure 3.4A). Moreover, the assignment test had 19 of the 20 localities assigned to Pine Gully with a high degree of confidence (>99.9%).

Some evidence suggested two other potential source populations. San Felipe Creek also clustered with the introduced populations (Figure 3.3) and showed admixture with the introduced range (47.3%, Figure 3.4A). However, this locality has consistently shown some distinctness from the rest of *G. affinis* (Chapter 2), suggesting it is a more divergent population perhaps more related to other *G. affinis* populations that extend into Mexico, an area not sampled. San Felipe Creek could have clustered with other introduced populations because it is simply more divergent than the rest of *G. affinis*. One locality from Taiwan was assigned to Alamito Creek in southwest Texas. Alamito Creek seems an unlikely source since it is a remote location by today's standards, let alone sometime in the past. Furthermore, given that mosquitofish are often introduced locally after initially being brought to an area, one would expect that nearby localities have similar assignments, which they do not seem to show in Taiwan.

The number of introductions is a challenging matter. It seems clear that one locality provided most (if not all) of the individuals that later were introduced around Asia. However, the mtDNA haplotype G is clustered in three localities in eastern China (Table 3.2, Figure 3.1). While this haplotype is present in the native range, it occurs only in the localities mentioned above and Hawaii. None of the intermediate locations contained this haplotype as would be expected according to the historical record. Unlike haplotype L, which can be found all over the introduced range, haplotype G may suggest a second introduction from Hawaii to China. Similarly, haplotype I is restricted to Taiwan in the introduced range, but common in the native range, including Pine Gully (Table 3.2). However, the sampling of more localities throughout the Hawaiian Islands may reveal the presence of this haplotype. Otherwise, it would suggest a second introduction from the United States and possibly a second source population.

Lack of evidence for a recent bottleneck

Intentional introductions typically involve large numbers of individuals, the initial recorded introduction of mosquitofish to Hawaii involved at least 150 individuals. We tested for evidence of a recent bottleneck in both the native and introduced localities as evidence for more introductions than those in the historical record, which may have involved smaller numbers of individuals. Despite a reduction in genetic diversity discussed above, we found no evidence for a recent bottleneck except for locality 44, which had recently been introduced to the pond we sampled from (personal observation). The lack of evidence for a bottleneck concurs with the historical record and large numbers of individuals being used to found new areas. Even a small number of individuals introduced to an area, like is documented for the Philippines (Seale 1917), could still contain large effective population sizes since females are capable of storing sperm from multiple individuals for long periods of time (Evans *et al.* 2011), though this introduction was supplemented by another introduction to the Philippines (Jordan 1927).

Overall, we found that the historical record and the molecular markers corroborated one another nicely. This may be due to the fact that the introduction of mosquitofish was sponsored by government agencies and thus required reports on the completion and follow-up studies (Seale 1905). Mosquitofish are known to be successful colonizers, establishing large populations quickly (Pyke 2005), thus, further introductions may have been unnecessary since they established so well and grew to such densities that they could be moved around locally.

A recent paper on the introduction of *G. holbrooki* throughout Europe, also found a very limited area for the source population in the United States (Sanz *et al.* 2013). They found most of their introduced samples could be traced back to a sampling locality within Virginia. Combined with our results, this would suggest that mosquitofish introduced around the world may have just

a few source localities from the native range, severely limiting the amount of genetic diversity throughout the introduced range. Mosquitofish would therefore be excellent systems for studying the impacts of reduced genetic variation on a very successful species, an area of research that would have implications for invasion biology and the conservation of small populations.

References

Allendorf F, Lundquist L (2003) Introduction : Population Biology, Evolution, and Control of invasive Species. *Conservation Biology* **17**, 24-30.

Ayala JR, Rader RB, Belk MC, Schaalje GB (2007) Ground-truthing the impact of invasive species: spatio-temporal overlap between native least chub and introduced western mosquitofish. *Biological Invasions* **9**, 857-869.

Ayres R, Pettigrove V, Hoffmann A (2012) Genetic structure and diversity of introduced eastern mosquitofish (*Gambusia holbrooki)* in south-eastern Australia. *Marine and Freshwater Research* **63**, 1206-1214.

Ayres RM, Pettigrove VJ, Hoffmann Aa (2010) Low diversity and high levels of population genetic structuring in introduced eastern mosquitofish (*Gambusia holbrooki*) in the greater Melbourne area, Australia. *Biological Invasions* **12**, 3727-3744.

Barun A, Niemiller ML, Fitzpatrick BM, Fordyce Ja, Simberloff D (2013) Can genetic data confirm or refute historical records? The island invasion of the small Indian mongoose (*Herpestes auropunctatus*). *Biological Invasions* **15**, 2243-2251.

Brown GP, Shilton C, Phillips BL, Shine R (2007) Invasion, stress, and spinal arthritis in cane toads. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 17698-17700.

Ciosi M, Miller NJ, Kim KS*, et al.* (2008) Invasion of Europe by the western corn rootworm, *Diabrotica virgifera virgifera*: multiple transatlantic introductions with various reductions of genetic diversity. *Molecular Ecology* **17**, 3614-3627.

Clement M, Posada D, Crandall Ka (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology* **9**, 1657-1659.

DeWoody J, Avise J (2000) Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. *Journal of Fish Biology* **56**, 461-473.

Dlugosch KM, Parker IM (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology* **17**, 431-449.

Earl DA, VonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359-361.

Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data: why, how and so what? *Molecular Ecology* **19**, 4113-4130.

Evans JP, Pilastro A, Schlupp I (2011) *Ecology and evolution of poeciliid fishes* The University of Chicago Press, Chicago.

Excoffier L, Lischer HEL (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.

Facon B, Pointier J-P, Glaubrecht M*, et al.* (2003) A molecular phylogeography approach to biological invasions of the New World by parthenogenetic Thiarid snails. *Molecular Ecology* **12**, 3027-3039.

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166.

Fitzpatrick BM, Fordyce Ja, Niemiller ML, Reynolds RG (2012) What can DNA tell us about biological invasions? *Biological Invasions* **14**, 245-253.

Hale ML, Burg TM, Steeves TE (2012) Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PloS one* **7**, e45170.

Hierro J, Maron J, Callaway R (2005) A biogeographical approach to plant invasions: the importance of studying exotics in their introduced and native range. *Journal of Ecology* **93**, 5-15.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Jordan D (1927) The mosquitofish (*Gambusia*) and its relation to malaria. *Report of the Board of Regents of the Smithsonian Institution 1926* **1926**, 361-368.

Koya Y, Itazu T, Inoue M (1998) Annual reproductive cycle based on histological changes in the ovary of the female mosquitofish, *Gambusia affinis*, in central Japan. *Ichthyological Research* **45**, 241-248.

Krumholz L (1948) Reproduction in the western mosquitofish, *Gambusia affinis affinis* (Baird & Girard), and its use in mosquito control. *Ecological Monographs* **18**, 1-43.

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

Lindholm AK, Breden F, Alexander HJ*, et al.* (2005) Invasion success and genetic diversity of introduced populations of guppies *Poecilia reticulata* in Australia. *Molecular Ecology* **14**, 3671-3682.

Lowe S, Browne M, Boudjelas S, De Poorter M (2000) 100 of the world's worst invasive species. A selection from the Global Invasive Species Database. The Invasive Species

Specialist Group (ISSG) a sepcialist group of the Species Survival Commission (SSC) of the World Conservation Union (IUCN).

Luikart G, Cornuet J (1998) Empirical Evaluation of a Test for Identifying Recently Bottlenecked Populations from Allele Frequency Data. *Conservation Biology* **12**, 228-237.

Pan J, Su B, Zheng W (1980) Biological characteristics of *Gambusia affinis* and the prospects for its use in for mosquito control. *Journal of South China Normal University (Natural Science)*, 117-138.

Piry S, Alapetite a, Cornuet J-M*, et al.* (2004) GENECLASS2: a software for genetic assignment and first-generation migrant detection. *Journal of Heredity* **95**, 536-539.

Piry S, Luikart G, Cornuet J-M (1999) BOTTLENECK: A computer program for detecting recent reductions in the effective population size using allele frequency data. *Journal of Heredity* **90**, 502-503.

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

Purcell KM, Ling N, Stockwell Ca (2012) Evaluation of the introduction history and genetic diversity of a serially introduced fish population in New Zealand. *Biological Invasions* **14**, 2057-2065.

Pyke GH (2005) A Review of the Biology of *Gambusia affinis* and *G. holbrooki*. *Reviews in Fish Biology and Fisheries* **15**, 339-365.

Pyke GH (2008) Plague Minnow or Mosquito Fish? A Review of the Biology and Impacts of Introduced *Gambusia* Species. *Annual Review of Ecology, Evolution, and Systematics* **39**, 171-191.

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 9197-9201.

Rauchenberger M (1989) Systematics and Biogeography of the Genus *Gambusia* (Cyprinodontiformes: Poecilidae). *American Museum Novitates* **2951**, 1-76.

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

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**, 137-138.

Rousset F (2008) genepop, '007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.

Ryman N, Palm S (2006) POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Molecular Ecology Notes* **6**, 600-602.

Sakai A, Allendorf F, Holt J (2001) The population biology of invasive species. *Annual Review of Ecology and Systematics* **32**, 305-332.

Sanz N, Araguas RM, Vidal O*, et al.* (2013) Genetic characterization of the invasive mosquitofish (*Gambusia* spp.) introduced to Europe: population structure and colonization routes. *Biological Invasions* **15**, 2333-2346.

Seale A (1905) Report of Mr. Alvin Seale of the United States Fish Commission, on the introduction of top-minnows to Hawaii from Galveston, Texas. *The Hawaiian Forester and Agriculturalist* **2**, 364-367.

Seale A (1917) The mosquitofish, *Gambusia affinis* (Baird and Girard), in the Philippine Islands. *Philippine Journal of Science*.

Stockwell CA, Henkanaththegedara SM (2011) Evolutionary conservation biology. In: *Ecology and Evolution of Poeciliid Fishes* eds. Evans JP, Pilastro A, Schlupp I), pp. 128-141. The University of Chicago Press, Chicago.

Templeton A, Crandall K, Sing C (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**, 619-633.

Tsutsui N, Suarez A (2001) Relationships among native and introduced populations of the Argentine ant (*Linepithema humile*) and the source of introduced populations. *Molecular Ecology* **10**, 2151-2161.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) Micro-Checker: Software for Identifying and Correcting Genotyping Errors in Microsatellite Data. *Molecular Ecology Notes* **4**, 535-538.

Vidal O, García-Berthou E, Tedesco Pa, García-Marín J-L (2009) Origin and genetic diversity of mosquitofish (*Gambusia holbrooki*) introduced to Europe. *Biological Invasions* **12**, 841-851.

Vidal O, Sanz N, Araguas R-M*, et al.* (2012) SNP diversity in introduced populations of the invasive *Gambusia holbrooki*. *Ecology of Freshwater Fish* **21**, 100-108.

Xie Y-P, Fang Z-Q, Hou L-P, Ying G-G (2010) Altered development and reproduction in western mosquitofish (*Gambusia affinis*) found in the Hanxi River, southern China. *Environmental toxicology and chemistry* **29**, 2607-2615.

Yan X, Zhenyu L, Gregg W, Dianmo L (2001) Invasive species in China—an overview. *Biodiversity & Conservation* **10**, 1317-1341.

Table 3.1 – List of sampling localities used in the study. The labels and names are consistent with the figures. Region (N = native range (mainland United States), HI = Hawaii, TW = Taiwan, PH = Philippines, JP = Japan, and CH = China), number of individuals per locality used (N), and locality coordinates used are provided. Genetic diversity estimates (average number of alleles, observed heterozygosity and expected heterozygosity) for each locality are reported. Assignment test results are displayed as the baseline population each locality was assigned back to with at least 99.9% confidence. A significant value for excess heterozygosity under two different mutation models (IAM and TPM) is listed.

Label	Locality Name	Region	N	Lat.	Long.	N_a	H_o	H_e	Assignment	IAM	TPM
1	Alamito Creek	N	20	29.52	-104.30	7.2	0.58	0.69	1	0.049	NS
2	San Felipe Creek	N	30	29.37	-100.88	6.4	0.35	0.57	2	NS	NS
3	Pine Gully	N	30	29.59	-95.00	11.7	0.58	0.79	3	NS	NS
4	Johnson Creek	N	30	30.15	-99.34	6.9	0.54	0.65	4	0.003	NS
5	South Concho River	N	30	31.21	-100.50	8.4	0.59	0.73	5	0.010	NS
6	North Bosque River	N	30	32.25	-98.23	8.7	0.52	0.73	6	0.001	NS
7	Oakbrook Park	N	30	33.15	-96.81	3.6	0.50	0.51	7	0.006	0.018
8	Sanders Creek	N	30	33.87	-95.54	9.2	0.63	0.71	8	NS	NS
9	Pennington Creek	N	30	34.26	-96.68	8.4	0.53	0.70	9	NS	NS
10	Red River	N	30	34.86	-99.51	7.6	0.54	0.64	10	NS	NS
11	Turkey Creek	N	30	35.35	-96.69	7.6	0.53	0.66	11	NS	NS
12	Pecan Creek	N	30	35.91	-95.12	6.2	0.49	0.64	12	0.004	NS
13	Clarke Bayou	N	30	32.57	-93.49	9.9	0.64	0.73	13	NS	NS
14	Bayou Macon	N	30	32.45	-91.46	9.2	0.58	0.65	14	NS	NS
15	Little Missouri River	N	30	34.05	-93.72	7.4	0.61	0.62	15	NS	NS
16	Brodie Creek	N	30	34.71	-92.38	6.7	0.57	0.60	16	NS	NS
17	Little Red River	N	30	35.82	-92.55	4.8	0.50	0.54	17	0.019	NS
18	Pascagoula River	N	30	31.34	-89.41	8.1	0.49	0.69	18	NS	NS
19	Big Black River	N	30	33.38	-89.61	10.0	0.59	0.70	19	NS	NS
20	Reelfoot Lake	N	30	36.40	-89.34	8.0	0.66	0.65	20	NS	NS
21	Hillabee Creek	N	21	32.99	-85.86	5.2	0.46	0.59	21	NS	NS
22	Roebuck Spring Run	N	27	33.58	-86.71	5.2	0.50	0.58	22	NS	NS
23	James Creek	N	15	33.91	-86.96	4.6	0.55	0.63	23	NS	NS

24	Conasauga River	N	30	34.68	-84.94	9.1	0.60	0.74	24	0.002	NS
25	Kualoa	HI	30	21.51	-157.84	6.8	0.54	0.64	3	0.027	NS
26	SuAo	TW	30	24.57	121.85	4.1	0.43	0.43	3	NS	NS
27	Yilan University	TW	26	24.75	121.74	5.7	0.28	0.60	3	0.010	NS
28	Sanxia	TW	30	24.88	121.42	5.7	0.43	0.57	3	NS	NS
29	Gangziliao	TW	20	25.13	121.78	4.3	0.43	0.50	1	NS	NS
30	Jiji	TW	21	23.83	120.80	4.8	0.49	0.61	3	0.004	NS
31	Guagua	PH	30	14.96	120.64	6.6	0.51	0.65	3	NS	NS
32	Apalit	PH	30	14.93	120.76	6.2	0.55	0.66	3	0.003	NS
33	Guiguinto	PH	30	14.83	120.88	6.2	0.55	0.64	3	0.033	NS
34	Barrio Muron	PH	30	14.67	120.98	6.0	0.53	0.64	3	0.004	NS
35	Midori River	JP	21	32.75	130.70	5.2	0.48	0.58	3	NS	NS
36	Zuibaiji River	JP	23	33.59	130.25	5.0	0.32	0.59	3	0.013	NS
37	SHOU[1]	CH	30	30.88	121.90	5.0	0.34	0.45	3	NS	NS
38	AHNU[2]	CH	20	31.33	118.37	3.3	0.37	0.41	3	0.047	NS
39	AHNU South	CH	17	31.29	118.38	3.4	0.33	0.42	3	NS	NS
40	East Lake	CH	30	30.54	114.39	4.8	0.35	0.48	3	NS	NS
41	South Lake	CH	30	30.47	114.38	6.3	0.41	0.61	3	NS	NS
42	Lover's Lake	CH	30	23.14	113.35	4.8	0.39	0.54	3	0.013	NS
43	Guilin	CH	10	25.27	110.29	3.7	0.41	0.51	3	0.024	NS
44	XTBG[3]	CH	28	21.93	101.26	2.1	0.22	0.23	3	0.007	0.014

[1] Shanghai Ocean University
[2] Anhui Normal University
[3] Xishuangbanna Tropical Botanical Garden

Table 3.2 – Haplotype table detailing the number of individuals for each cytochrome *b* haplotype found in the introduced range at the putative source locality (3) and each introduced locality (25-44).

Label	Locality Name	Haplotype					
		G	G10	G11	G12	I	L
3	Pine Gully	6				11	10
25	Kualoa	6	11	2	1		9
26	SuAo					10	
27	Yilan University					1	9
28	Sanxia						10
29	Gangziliao						10
30	Jiji						10
31	Guagua						10
32	Apalit						10
33	Guiguinto						10
34	Barrio Muron						10
35	Midori River						10
36	Zuibaiji River						10
37	SHOU	9					1
38	AHNU	10					
39	AHNU South	10					
40	East Lake						10
41	South Lake						10
42	Lover's Lake						10
43	Guilin						10
44	XTBG						10

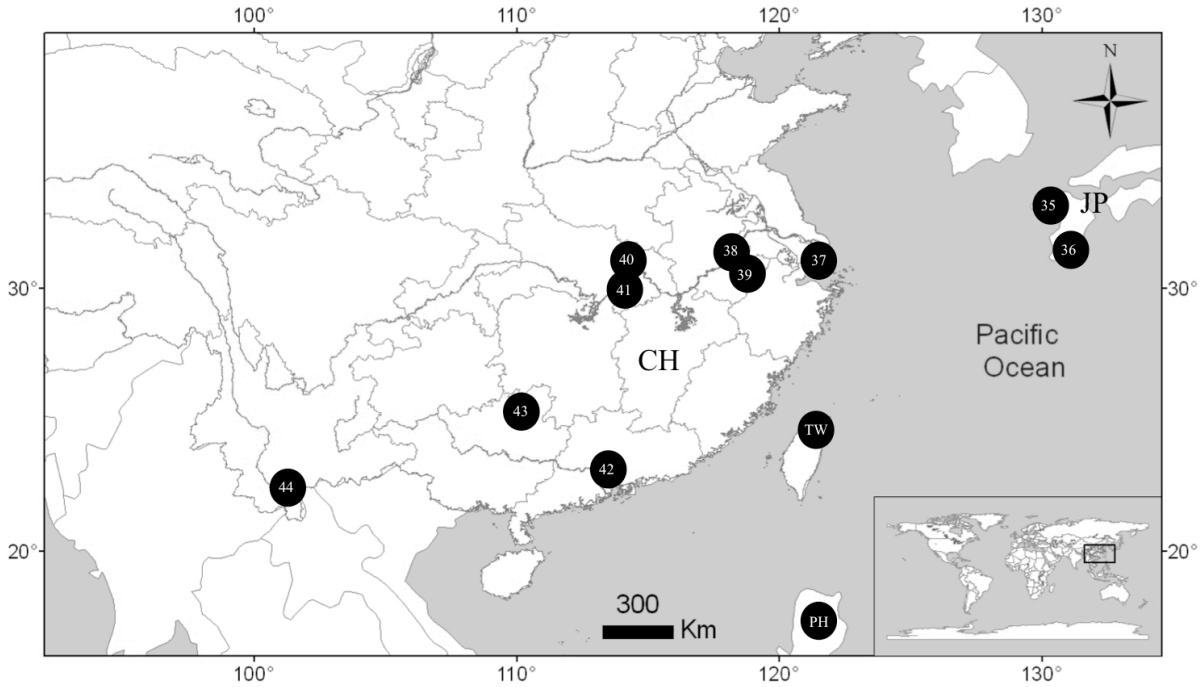

Figure 3.1 – Map of introduced localities in Taiwan, the Philippines, Japan and China. Black circle indicates location (see Table 3.1). China (CH) and Japan (JP) are labeled. Multiple localities in Taiwan (TW=26-30) and the Philippines (PH=31-34) are represented by a single circle. Locality 25 from Hawaii not shown.

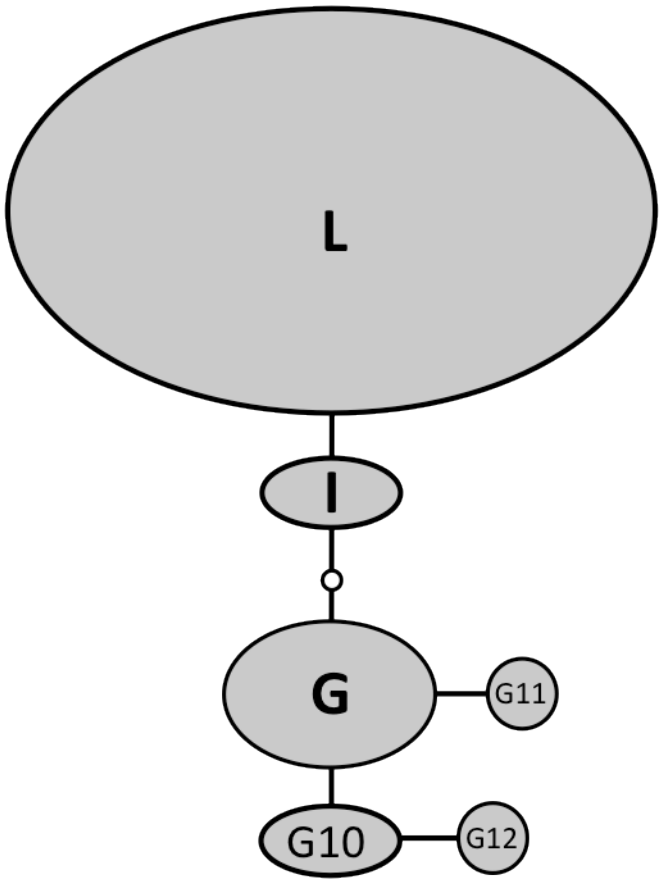

Figure 3.2 – Genealogical relationships of the six mitochondrial haplotypes found throughout the introduced range of *G. affinis*. Size of the circle indicates the frequency at which the haplotype occurred in the dataset. Each circle indicates one mutational step along the line away from other haplotypes. The empty circle indicates a hypothesized haplotype that has gone unsampled.

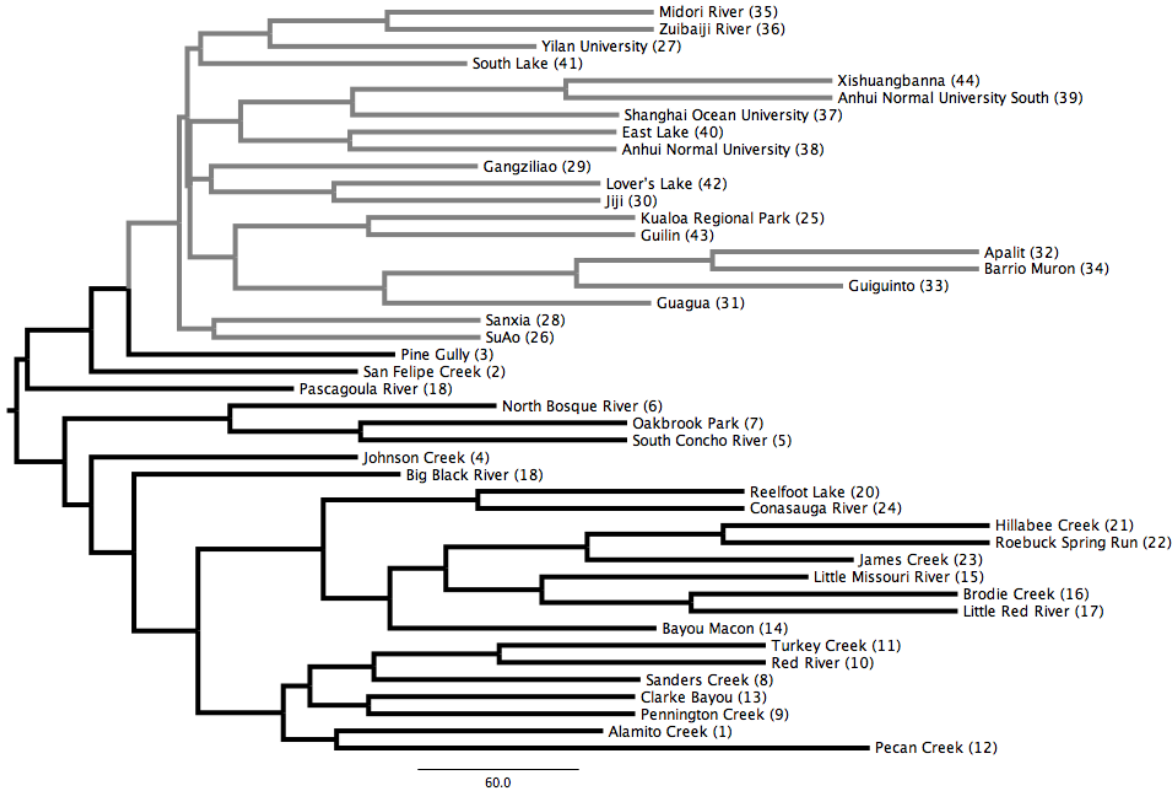

Figure 3.3 – Neighbor-joining population tree of the native (black) and introduced (gray) localities of G. affinis based on the allele frequencies of 18 microsatellite markers. Locality names follow those listed in Table 3.1.

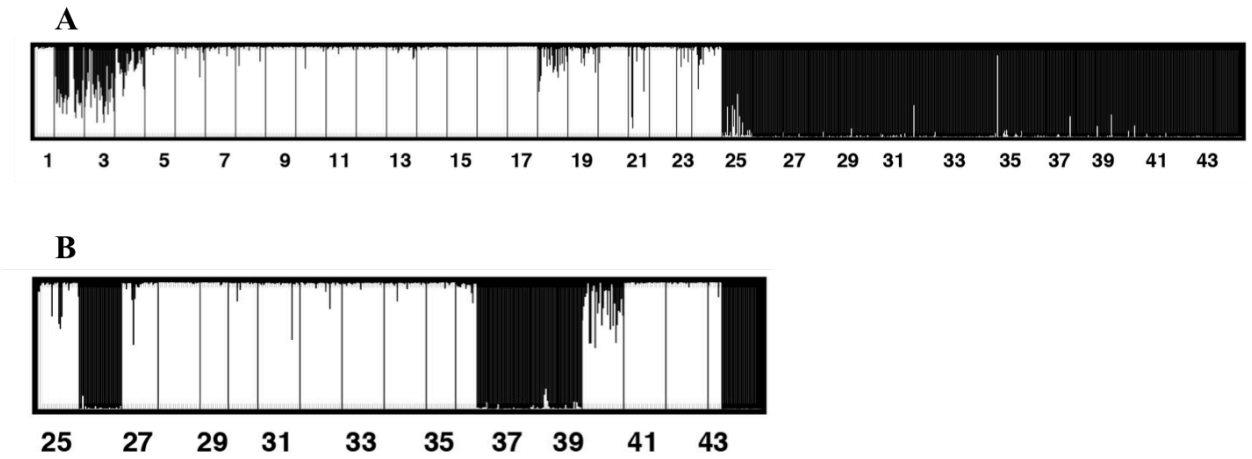

Figure 3.4 – Plots of the optimal clusters found for *G. affinis* (K=2), the native and introduced localities combined (A) and the introduced localities alone (B). Labels follow Table 3.1 with only the odd labels. Each column is an individual showing the percent membership of each group with localities divided by dark lines.

CHAPTER 4: IMPACT OF MISSING DATA ON POPULATION GENETIC INFERENCES OF INVASION SCENARIOS FROM SIMULATED RADSEQ DATA[1]

[1] Lee JB and Mauricio R. To be submitted to *PLoS Computational Biology*.

Abstract

The use of next-generation sequencing (NGS) technology is drastically changing the scale at which we can sample the genome. However, despite the rapid advances in NGS technology, missing data can still be present and potentially impact the results. We investigate the impact of missing data in restriction-site associated DNA sequencing (RADseq) datasets by simulating data under six scenarios of an invasion. We simulate increasing amounts of missing data in these datasets and also examine how filtering the datasets compares with random subsamples. We estimated pairwise FST for the simulated populations in all datasets and performed an assignment test for each dataset. We observed no real difference in F_{ST} estimates and probability of correct assignment in the number of loci used without any missing data. The missing data simulated in the datasets had little impact upon the estimates of FST. However, probability of correct assignment began to decline at 50% missing data for scenarios with high migration. Scenarios of low and moderate declined only slightly at 90% missing data. The filtered datasets showed no difference from random subsets in FST estimates, but improved the assignment probabilities. We discuss the results in light of the robustness of the datasets with missing data, how the filtering process helps, and other implications for invasion biology.

Introduction

Population genetics focuses on describing patterns and testing hypotheses of evolutionary processes within and between populations (Hartl & Clark 1997). Historically, researchers have sampled large numbers of individuals in several populations, scored them for a number of genetic markers, and estimated parameters based on allele frequencies. One major criticism of this approach has focused on the low number of markers that researchers have used arguing that they represent a small percentage of the genome (Rokas & Abbot 2009). Indeed, evolutionary genetics has constantly strived to increase the number of markers used in studies in an effort to more thoroughly sample the genome and thus obtain more accurate estimates for the population and species. Next-generation sequencing (NGS) technology has alleviated this challenge by introducing methods that allow researchers to sample thousands of markers from many individuals at the same time, especially in non-model organisms (Allendorf *et al.* 2010; Ellegren 2008). Thus, researchers are now able to obtain large datasets (thousands of markers, many individuals, multiple populations) for the organism they are using to investigate evolutionary processes in nature (Davey & Blaxter 2010; Faircloth *et al.* 2012; Hohenlohe *et al.* 2010; Lemmon & Lemmon 2013; McCormack *et al.* 2013).

One NGS method that has gained popularity is restriction-site associated DNA sequencing, or RADseq (Baird *et al.* 2008). This method employs a genome reduction approach by digesting genomic DNA with restriction enzymes, adding platform specific adapters, and selecting size fragments within a certain distribution. Protocols for RADseq vary mostly at the number of restriction enzymes used and the size selection method incorporated (Elshire *et al.* 2011; Peterson *et al.* 2012). The resulting sequenced reads from this library are then assembled using a reference genome or *de novo* (Willing *et al.* 2011) and polymorphic single nucleotide

polymorphisms (SNPs) are scored for each individual (Bradbury *et al.* 2007; Catchen *et al.* 2011). It is important to point out that the steps described above can be outsourced completely or partially. The result is a large matrix of scored SNPs for the individuals that a researcher then uses as raw data for analyses. Population geneticists have eagerly adopted RADseq as a method to obtain genome-wide data to address a variety of questions (Narum *et al.* 2013).

Missing data can be introduced at various stages of the RADseq protocol. Poor sample quality could lead to systemic missing data for an entire individual. A mutation at the restriction cut site may prevent the cutting into smaller fragments, resulting in a larger fragment that may not be selected for sequencing. Poor efficiency in ligating adapters and tags to the digested fragments could lead to a loss of fragments for some individuals. Low coverage may exclude loci for certain individuals since coverage is not uniform across sequenced reads. The missing data is represented by an 'N' at a particular datapoint, instead of a called SNP represented by a nucleotide or one of its ambiguity codes for two alternate bases (representing the heterozygote). In sum, RADseq datasets will have missing data, some correlated to a single locus or individual and others more randomly distributed.

However, unlike more traditional Sanger sequencing methods, data cannot be obtained for markers that are missing for individuals due to the nature of the library preparation and sequencing method. Researchers have to make decisions regarding how to analyze the data regardless of the amount of missing data. Many researchers choose to filter the datasets prior to analysis in order to obtain the SNPs of the highest quality. This can reduce a raw dataset from ten of thousands of SNPs to a few thousand or hundred depending on how the researcher chooses to filter the SNPs. What would be helpful is an understanding of how missing data in these large datasets impacts analyses and, by extension, the inferences made.

The goal of this study is to simulate RADseq datasets with increasing amounts of missing data and examine how the missing data affects the results of common population genetic analyses. We do this under several scenarios of an introduced species because of our own research interests in this area and because we feel that conservation genetics has much to gain from these large RADseq datasets. We address four main questions to achieve this goal: (1) How many SNPs are needed to obtain correct estimates? (2) How do increasing amounts of missing data impact the estimates? (3) How do varying the number of SNPs and the amount of missing data impact estimates? (4) Do estimates improve when a filtering approach is used? These questions are ones commonly asked by researchers and we hope the results presented here will provide assistance in making decisions and spark more interest in understanding the generation and analysis of NGS data.

Methods

Data simulation and scenarios

We began by simulating 10 datasets with 5000 called SNPs for each of six simple scenarios that sample 30 individuals for each of three populations (two native and one introduced, Figure 4.1). We used a Python script (https://github.com/mgharvey/mps-sim, last accessed March 21, 2014) that relies upon ms (Hudson 2002), seq-gen (Rambaut & Grassly 1997), and BioPython (Cock *et al.* 2009) to simulate RADseq datasets similar to those produced by the genotyping-by-sequencing method (Elshire *et al.* 2011). We emphasize that our simulations do not address sequencing depth, quality scores, or the actual source of missing data. Rather, our simulations produced complete datasets of called SNPs, which we manipulate to include missing data.

We developed simple demographic scenarios by varying two parameters: the number of introductions (m_1) and migration rate in the native range (m_2, Figure 4.1). A single introduction occurs when a group of individuals is introduced to a new region and establishes with no more immigrants from the native range. We simulated a single introduction in ms (Hudson 2002) by forcing the introduced population to diverge recently (tau_1) from the actual source and setting the migration rate to zero. A multiple introduction will follow the same pattern except there is ongoing migration from the native range. Migration can come from the same source population or from multiple source populations. In order to simplify the scenario, we chose the former to simulate multiple introductions by setting a moderate, asymmetric migration rate from the actual source population to the introduced population. We simulated population structure in the native range by forcing the native populations to have a deep divergence from one another (tau_2) and varied the migration to represent low, moderate, and high rates that we selected after a survey of several published studies. While the divergence of populations in a native range may vary, we chose a deep divergence time to allow us to look at the impact of migration alone. The pairwise combination of two introduction parameters and three migration parameters created six scenarios. We use these parameters throughout the text to refer to a specific scenario or a subset of the scenarios (Table 4.1). The 10 datasets simulated for each of these scenarios contained no missing data, in other words, they were perfect datasets in that every SNP for every individual was called. The specific ms command values for the parameters described above are provided in Table 4.1. For all scenarios, we selected a theta value of 0.4 for the mutation rate parameter and used 0.001 as the theta/site value for gene tree scaling. For each dataset, the script simulated alignments of 64 bp and selected only alignments containing a single biallelic polymorphic site (SNP) until we obtained 5000 alignments. Each alignment used was saved in a separate nexus

file and we generated a HapMap file of all the SNPs, which was used for all downstream manipulations and analyses conducted. Configuration files for the generation of these simulated datasets are available upon request.

Number of loci

As a baseline for downstream analyses, we randomly sampled 2500, 1000, 500 and 100 SNPs from each of the 60 datasets creating random subsamples of perfect datasets from the full 5000 SNPs for each scenario. The analysis of these randomly subsampled 'perfect' datasets allowed us to explore how estimated values varied with decreasing number of loci. We expected these randomly subsampled datasets to have similar averages to those of the full datasets but as the number of SNPs decreased the standard error for the estimates would increase.

Impact of missing data

In order to test the impact of missing data, we simulated missing data in each 5000 SNP dataset using a custom Python script (Appendix 1), which takes each individual and randomly substitutes a number of called SNPs with an 'N' from a normal distribution. The mean for the normal distribution was calculated by multiplying the desired amount of missing data by the number of SNPs in the dataset (in this case, 5000). We chose to scale the standard deviation for the normal distribution at 3% of the mean. The scaling of the standard deviation was an arbitrary decision as no information on how this occurs in empirical datasets is available. The script simulated missing data in 10% increments from 0-90%, effectively creating 10 treatments with the perfect datasets described above acting as the control (0% missing data). This allowed us to compare the estimated values on increasing amounts of missing data and we expected datasets with larger amounts of missing data to have lower average values with a large standard error, which could lead to inaccurate inferences made.

Number of loci and missing data

In order to examine the interaction between the number of loci and missing data, we randomly subsampled the datasets treated with all amounts of missing data for 2500, 1000, 500, and 100 SNPs using a custom Perl script. The same random individuals were selected for each treatment in order to compare across treatments. We expected the estimated values for these datasets to decrease with increasing standard error with lower amounts of missing data as compared to those with the full datasets.

Filtering of missing data

One method to minimize the impact of missing data is to filter out loci based upon a threshold of missing data determined by the researcher. For example, a researcher can determine they only want to analyze loci with 20% or less missing data. Since we already simulated the amount of missing data, we chose to filter down to approximately 2500, 1000, 500, and 100 SNPs in the software program TASSEL v3.0 (Bradbury *et al.* 2007) so as to compare with the randomly sampled datasets. This required us to vary the filtering parameters for each of the treatments and for each of the number of loci targeted. For example, in order to filter down to ~2500 SNPs in datasets with 10% simulated missing data, we set the filter to accept loci with at least 80 called SNPs (Table 4.2). However, in order for datasets with the same amount of missing data to be filtered to lower amounts of SNPs, we increased the minimum count required to be included. Table 4.2 provides the exact values used to filter and the average number of SNPs per dataset. Thus, the filtered datasets contain not just a subsample of the full datasets, but the 'best' subsample as opposed to the random subsample. We compared the estimated values of the filtered datasets with those randomly selected with the expectation that the filtered datasets would provide better average values as missing data increased and have smaller standard errors.

Analyses

We selected two population genetic values to estimate for all of the datasets described above and calculated them in the R statistical software package (R Development Core Team 2012). First, we calculate pairwise F_{ST} for all datasets as a measure of differentiation between populations. We selected pairwise F_{ST} since it is broadly accepted and understood as a standard measurement for population differentiation. We calculated pairwise F_{ST} for all populations using the R-package *hierfstat* (Goudet 2005) and report the mean pairwise F_{ST} and standard error for all replicates in each dataset. The second value estimated was the probability of correct assignment of the introduced population to its actual source. Assignment tests are a common and powerful method used in identifying source populations for introduced species and a wide range of other questions. We performed assignment tests using the R-package *PSMix* (Wu *et al.* 2006). Since there were only two possible source populations, we set K=2 and used the default settings for the analyses. Since we knew the correct source population, we were able to assess whether the introduced individuals were correctly assigned. We calculated the mean assignment probability for each population to each group. We report the mean probability of each introduced population assigned to the group with the highest mean assignment probability for the actual source population along with its standard error. Thus, with the datasets described above we can assess how these two values (pairwise F_{ST} and probability of correct assignment) changes by decreasing the number of loci sampled, increasing the amount of missing data, increasing the amount of missing data as loci are decreased, and by filtering for the best loci.

Results

Number of loci

In order to explore our first question of how many loci are needed to obtain correct results, we compared the results for the 5000 SNPs to those obtained by a random sample of 2500, 1000, 500, and 100 SNPs without any missing data. Estimated pairwise F_{ST} values for all datasets were consistent across all scenarios (Figure 4.2). The standard error was also very small for all average values and only noticeably increased when only 100 SNPs were randomly sampled. The probability of correct assignment of the introduced population also remained consistent across the varying number of SNPs (Figure 4.3). For datasets containing 500-5000 SNPs, probability of correct assignment was high (>0.98) across all scenarios. Datasets with 100 SNPs showed a decrease in probabilities for scenarios with high migration (>0.85). For scenarios with moderate and low migration, the decrease in probability was observable but still remained above 0.95. We observed no difference in the results due to the invasion parameters simulated for the F_{ST} estimates or the probability of correct assignment.

Impact of missing data

The results presented for the datasets without missing data provide a baseline comparison as we examine how missing data impacts the estimates of F_{ST} and probability of correct assignment. We found that pairwise F_{ST} remained consistent as missing data increased throughout the datasets and across all of the scenarios (Figure 4.4). At levels of 90% missing data, average pairwise F_{ST} dropped slightly, but no more than 0.03. The standard error did increase as missing data increased, however, we note that they remained relatively small. The average probability of correct assignment showed a similar pattern for both invasion scenarios (Figure 4.5). Probability of correct assignment remained high (>0.98) for all scenarios up to 50%

missing data. Scenarios with low and moderate migration continue to have such high probabilities of assignment up to 90% missing data where moderate scenarios decline to probabilities of 0.89 or greater. For scenarios with high migration, probability of correct assignment begins to decline at 60% missing data and shows sharper drops in probability at 80% missing data. Under a single introduction scenario high migration remained above 0.5, while the multiple introductions with high migration scenario actually dropped to 0.496. With only two populations to potentially be assigned to this means that assignment was close to random.

Number of loci and missing data

We randomly sampled the 5000 SNPs for 2500, 1000, 500, and 100 SNPs to determine how our estimated values changed by decreasing the number of loci in the datasets with missing data. Since all F_{ST} estimates performed similarly we report only the F_{ST} value between the two native populations (Figures 4.6 and 4.7). F_{ST} estimates remained consistent as the number of loci decreased, however as expected, we saw an increase in the standard error as the amount of missing data grew for all numbers of loci. The probability of correct assignment was high for all datasets in all scenarios at low amounts of missing data (Figure 4.8 and 4.9). Datasets with 100 loci were consistently lower than those from 500-5000 and had larger standard errors. The probability of correct assignment began to decrease as missing data increased with sharp declines at 70% and 50% for scenarios with moderate and high migration, respectively. Standard errors showed much more variability than previously seen for all datasets and scenarios.

Filtering of missing data

We filtered the datasets to approximate numbers of loci comparable to the random sample. This allowed us to compare how filtering out the 'worst' loci can improve overall estimates. We observed that F_{ST} values remained consistent for filtered datasets and showed very

little difference from the full dataset of 5000 SNPs or from those sampled randomly (Figures 4.6 and 4.7). We note that for scenarios with low and moderate migration the 100 SNP datasets vary widely in their mean averages with large standard error bars. The assignment tests of filtered datasets showed higher probabilities of correct assignment at larger amounts of missing data compared to random datasets (Figures 4.8 and 4.9). Filtered datasets improved assignment for scenarios of high migration particularly at the highest amounts of missing data. Both filtered and random datasets of 500-5000 SNPs performed similar to one another while datasets with only 100 SNP loci consistently had lower probabilities of correct assignment, especially for scenarios of moderate and high migration. We also note that the standard error for filtered datasets was smaller for all scenarios and number of SNPs when compared to randomly sampled SNPs.

Discussion

Next generation sequencing technology will have a profound impact on evolutionary biology over the next several years by providing genome-wide markers and datasets enabling researchers to address a wide range of question in greater depth (Allendorf *et al.* 2010; Ellegren 2008; McCormack *et al.* 2013). This study was motivated by an attempt to explore the robustness of one kind of NGS method by simulating missing data in RADseq-like datasets. We first discuss some of the limitations of our simulations before addressing the robustness of the analyses to missing data and how improvements were made through filtering. We then conclude with a brief comment on some applications for invasion biology.

Limitations

As with any modeling and simulation study, we made several simplifying assumptions in order to address our question of interest. We also assert that it is better to construct simple models to begin with and then increase complexity in order to understand what aspects of the

model are impacting the outcome. We choose to address some of the simplifications we made here in an effort to ensure our results are interpreted in the proper framework.

First, we simulated a demographic scenario with only two native populations making the assignment tests a 50/50 choice. In reality, assignment tests for introduced populations rarely only have two native populations, for example, in Chapter 3 we used 22 native populations. Thus, it would be informative to include larger numbers of native populations that would perhaps make the assignment more challenging depending on the level of migration used.

Second, we only simulated 5000 SNPs for the full dataset while most RADseq methods produce raw SNP calls orders of magnitude larger (Hamlin & Arnold 2014; Hohenlohe *et al.* 2010; McCormack *et al.* 2012). We chose 5000 SNPs for two reasons, one empirical and another practical. A study of simulated RADseq datasets specifically looking at how many loci are needed for accurate estimates of phylogenetic and demographic parameters concluded that datasets larger than 5000 SNPs improved very little in accuracy (Harvey *et al.* 2013). We also note that the disk drive space and analysis time required for larger datasets could be prohibitive.

Third, we sampled 30 individuals per population, which is actually high compared to published studies (Hamlin & Arnold 2014; Harvey *et al.* 2013; Hohenlohe *et al.* 2010). We chose a high sample size to ensure we had accurate allele frequencies for each population so that the analyses would be robust for the control datasets. Lower sample sizes in empirical datasets and/or uneven sample sizes could vastly impact the allele frequencies used for analysis. Thus, we feel our sample size is robust.

Fourth, we introduce a novel method for simulating missing data randomly in RADseq datasets. We acknowledge that not all missing data in these datasets is random. For example, individual samples could have a high amount of missing data due to poor DNA template quality.

Thus, we hope that future studies will improve on our initial attempt to simulate missing data. Published empirical datasets and modeling sequencing error are two sources that could provide information on how to model this better.

Finally, we filtered the datasets down to a specific number of loci, which is not what is commonly done in practice nor does it reflect the range of decisions that go into filtering. We chose to filter this way because we wanted the number of loci comparable to the datasets of randomly sampled loci. However, often researchers will select the amount of missing data they are comfortable with and filter to that amount and then run their analyses with the remaining loci. Furthermore, data can also be filtered based on poor performing samples and the frequency of minor alleles. For example, Hamlin & Arnold (2014) chose to filter out samples that performed poorly, loci with more than 20% missing data, and loci with a minor allele frequency of less than 1%. We did not have to deal with poor samples and our question focused on the impact of missing data and not minor allele frequencies.

Robustness of analyses

We found that both pairwise F_{ST} estimates and assignment tests were robust to missing data. Indeed, we found that F_{ST} estimates overall were consistent regardless of the amount of missing data or the number of SNPs used. The assignment tests accurately assigned the introduced population to its source with a probability of 0.98 or greater with up to 50% missing data. While at higher amounts of missing data the probability decreases, particularly for scenarios with high migration. However, while the average probability of some scenarios at 90% missing data decreased, only the scenario with multiple introductions and high migration resulted in an average probability that was random (0.496). Thus, all the other scenarios resulted in probabilities that favored correct assignment.

Filtering for better results

The filtering of RADseq datasets is a common practice and our results confirm its ability provide better results (Figures 4.8 and 4.9). The filtering process allows the researcher to proceed through their analyses with higher quality data that will provide more accurate estimates. By nature it will result in a smaller number of loci used for analysis, however we have shown that results are robust when smaller numbers of loci are used both with and without missing data. We found that the filtering did not differ much from the randomly sampled datasets in our F_{ST} estimates, however, the filtered datasets consistently had high probabilities of correct assignment, especially for datasets from 500-5000 SNPs.

Applications for invasion biology

We simulated invasion scenarios to reflect our own research interests and further emphasize the broad range of questions NGS datasets are used to address. One of the results that we did not anticipate was the lack of difference in the invasion scenarios we constructed. We found the main driver in the differences on how missing data impacted the results was due to the migration rate in the native range of our scenarios. The population structure and demography of the native range is an important aspect in reconstructing the invasion history of any species (Hierro *et al.* 2005; Sakai *et al.* 2001).

There are some recommendations that we suggest for researchers using RADseq datasets. First, the amount of missing data in should be reported in some way. The simplest way would be to count the number of 'N's in the entire dataset and present it as a percentage by dividing it by the total number of possible datapoints. A more elaborate report may also include observed patterns of missing data by certain samples or loci. Second, we recommend that researchers use at least 500 SNPs for their studies. While our simulations showed that datasets with 100 loci

gave accurate results we note that those datasets had the largest amount of variation in the results, thus any one dataset could give very different results and lead to wrong interpretations. Given that RADseq datasets typically produce raw SNPs on the order of tens or hundreds of thousands, we fell this will not be a problem even if stringent filtering is used. Finally, we suggest F_{ST} as a measure of how robust any dataset will be to missing data. Our analyses showed that F_{ST} was accurately obtained for all levels of missing data in all scenarios and for all the different numbers of loci examined. Thus, if the researcher knows how much missing data is in the raw dataset and they also have estimated F_{ST} they can make an informed decision on filtering. For example, a high F_{ST} might indicate that an assignment test will be robust to large amounts of missing data, whereas a low F_{ST} would indicate that such levels of missing data could lead to lower probabilities of correct assignment. In such cases, researchers would be wise to filter the dataset to obtain more reliable results.

While RADseq datasets have gained popularity for a wide range of issues in evolutionary biology (Davey & Blaxter 2010; Harvey *et al.* 2013; Hohenlohe *et al.* 2010; McCormack *et al.* 2013), invasion biology and conservation genetics studies utilizing such resources seem to be fewer despite the benefits (Allendorf *et al.* 2010). Yet conservation genetics will often deal with samples that may be more prone to missing data (*i.e.,* scat samples, museum tissues). We hope that continued simulation studies will provide accurate insights into how to best utilize the NGS technology for use in both evolutionary and conservation studies.

Missing data will always be an issue with any dataset. The ability to decrease and eliminate sources of missing data in NGS datasets will likely improve as library preparation methods are refined, new sequencing chemistries are advanced, and new technology becomes available. However, we will likely never be able to visualize the perfect dataset that we have

used in this study, but as we have shown, we may not have to in order to make correct and accurate inferences regarding population histories.

References

Allendorf FW, Hohenlohe Pa, Luikart G (2010) Genomics and the future of conservation genetics. *Nature reviews. Genetics* **11**, 697-709.

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

Bradbury PJ, Zhang Z, Kroon DE*, et al.* (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633-2635.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda, Md.)* **1**, 171-182.

Cock PJa, Antao T, Chang JT*, et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423.

Davey JW, Blaxter ML (2010) RADSeq: next-generation population genetics. *Briefings in functional genomics* **9**, 416-423.

Ellegren H (2008) Comparative genomics and the study of evolution by natural selection. *Molecular Ecology* **17**, 4586-4596.

Elshire RJ, Glaubitz JC, Sun Q*, et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* **6**, e19379.

Faircloth BC, McCormack JE, Crawford NG*, et al.* (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* **61**, 717-726.

Goudet J (2005) HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **2**, 184-186.

Hamlin JAP, Arnold ML (2014) Determining population structure and hybridization for two iris species. *Ecology and Evolution* **4**, 743-755.

Hartl D, Clark A (1997) *Principles of population genetics* Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts.

Harvey M, Smith B, Glenn T (2013) Sequence Capture versus Restriction Site Associated DNA Sequencing for Phylogeography. *arXiv:1312.6439 [q-bio.GN]*, 1-53.

Hierro J, Maron J, Callaway R (2005) A biogeographical approach to plant invasions: the importance of studying exotics in their introduced and native range. *Journal of Ecology* **93**, 5-15.

Hohenlohe Pa, Bassham S, Etter PD*, et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS genetics* **6**, e1000862.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-338.

Lemmon EM, Lemmon AR (2013) High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* **44**, 99-121.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* **66**, 526-538.

McCormack JE, Maley JM, Hird SM*, et al.* (2012) Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Molecular Phylogenetics and Evolution* **62**, 397-406.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe Pa (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology* **22**, 2841-2847.

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

R Development Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut a, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences : CABIOS* **13**, 235-238.

Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. *Trends in Ecology & Evolution* **24**, 192-200.

Sakai A, Allendorf F, Holt J (2001) The population biology of invasive species. *Annual Review of Ecology and Systematics* **32**, 305-332.

Willing E-M, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics* **27**, 2187-2193.

Wu B, Liu N, Zhao H (2006) PSMIX: an R package for population structure inference via maximum likelihood method. *BMC Bioinformatics* **7**, 317.

Appendix 1
Python script that adds missing data ('N') randomly to a HapMap formatted files in a given directory and outputs them to a subdirectory.

```
#!/usr/bin/env python
import os
import sys
import random

percent = 0.1
nignore = 11
dest = sys.argv[1]
os.mkdir(dest)

for file in os.listdir(os.getcwd()):
	if file.endswith(".txt"):
		f = open(file, 'r')
		header = f.readline().split()
		nindiv = len(header)-nignore
		data = [line.split() for line in f]
		f.close()
		nloci = len(data)
		mu = percent*nloci
		sigma = mu*0.03

		for i in range(nindiv):
			for j in random.sample(range(0,nloci),int(random.gauss(mu,sigma))):
				data[j][i+nignore] = 'N'
		outfile=open('%s/%s' % (dest,file),'w')
		outfile.write('\t'.join(header))
		outfile.write('\n')
		for line in data:
			outfile.write('\t'.join(line))
			outfile.write('\n')
		outfile.close()
```

Table 4.1 – Population parameters used to simulate the data for the six scenarios. For each scenario, we specify the divergence time (tau) and migration rates (m) used. Labels correspond with Figure 4.1 and the scenario names are consistent throughout the text.

Scenario	**tau_1**	**tau_2**	**m_1**	**m_2**
Single, Low	0.01	0.5	0	0.2
Single, Moderate	0.01	0.5	0	1.2
Single, High	0.01	0.5	0	6
Multiple, Low	0.01	0.5	1.2	0.2
Multiple, Moderate	0.01	0.5	1.2	1.2
Multiple, High	0.01	0.5	1.2	6

Table 4.2 – Details of the filtering process of simulated datasets. For each target number of SNPs in the first column, the minimum number of correctly called SNPs (minCount command in TASSEL) required for the locus to be included (maximum of 90) is given for each treatment with the average number of SNPs that resulted from the filter given below.

Target no.	10%	20%	30%	40%	50%	60%	70%	80%	90%	Overall
2500	80	71	61	51	44	35	25	16	5	-
	2676	2409	2532	2602	2534	2542	2669	2461	2461	2543
1000	83	75	66	57	48	40	30	20	10	-
	1184	951	1023	1021	1161	941	1072	1170	1189	1079
500	85	77	68	60	50	42	32	22	12	-
	395	423	541	411	612	503	578	621	568	517
100	87	79	72	63	54	46	35	25	15	-
	63	143	83	113	122	92	173	171	113	119

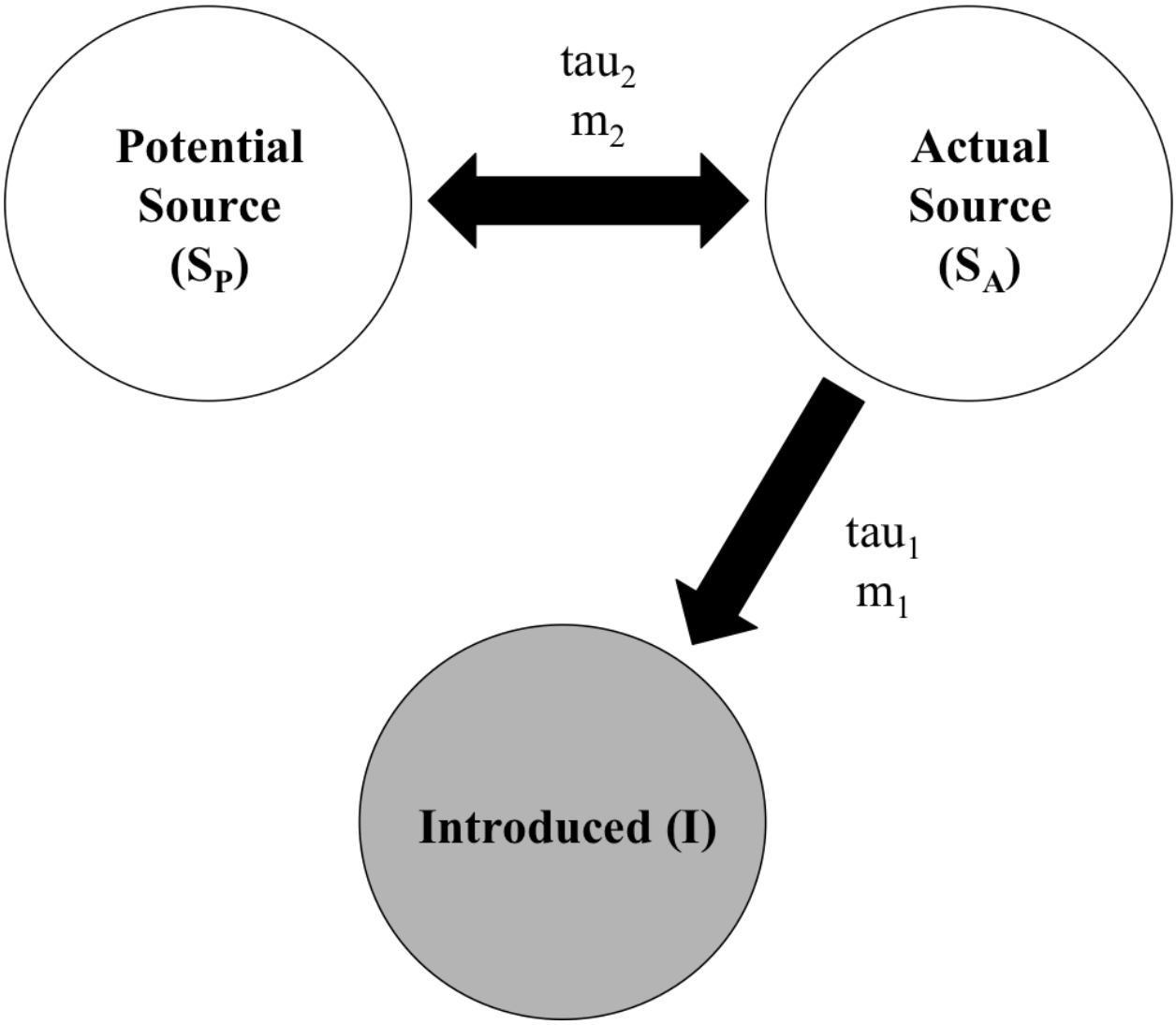

Figure 4.1 – Depiction of the overall scenario under which the datasets were simulated as described in the text.

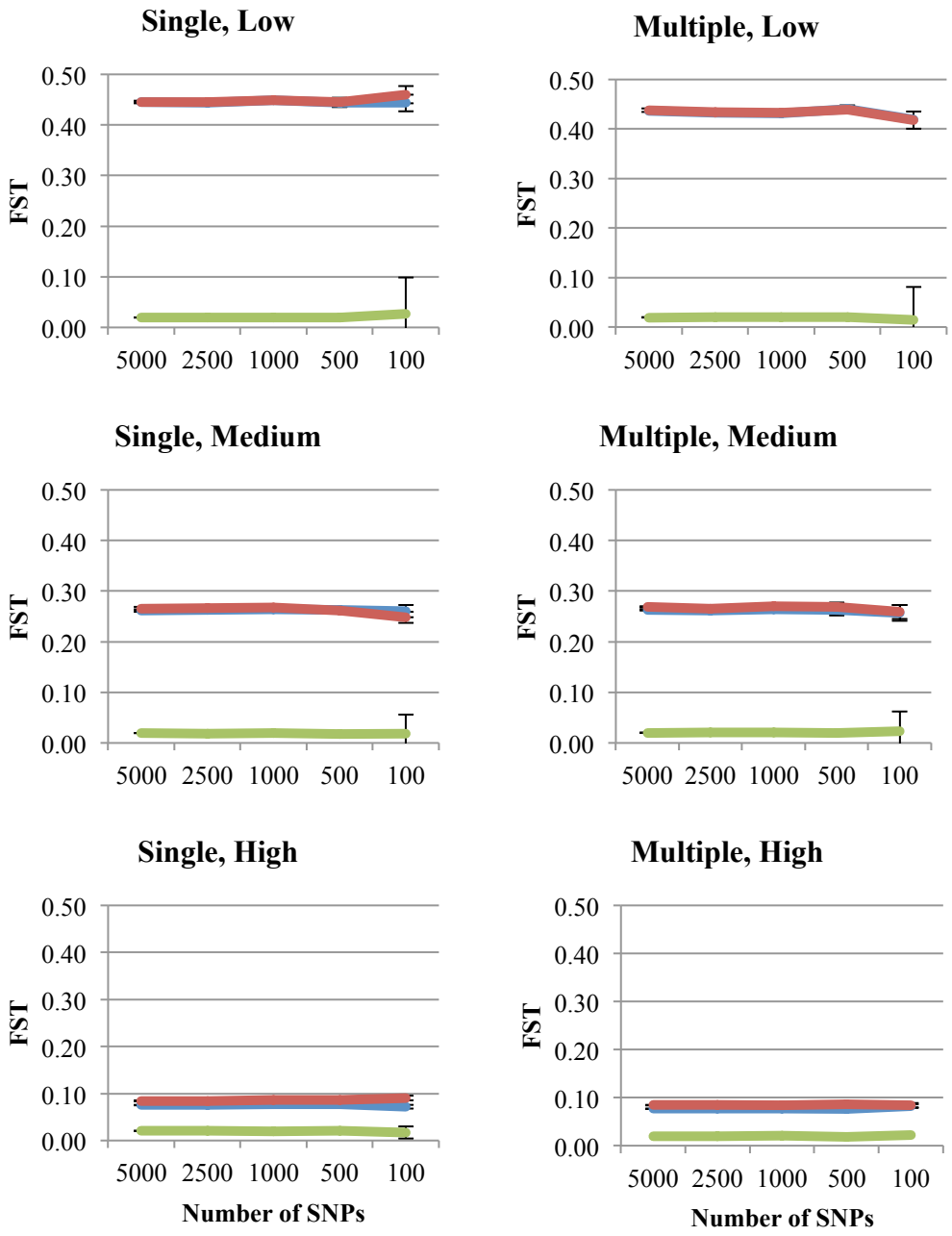

Figure 4.2 – Average pairwise F_{ST} estimates with standard error bars between the three populations in each of the simulated datasets without missing data for each scenario titled above each chart (Figure 4.1). Estimates are given for the full dataset of 5000 SNPs and a random sample of 2500, 1000, 500, and 100 SNPs. S_p v. S_a (blue), S_p v. I (red), and S_a v. I (green).

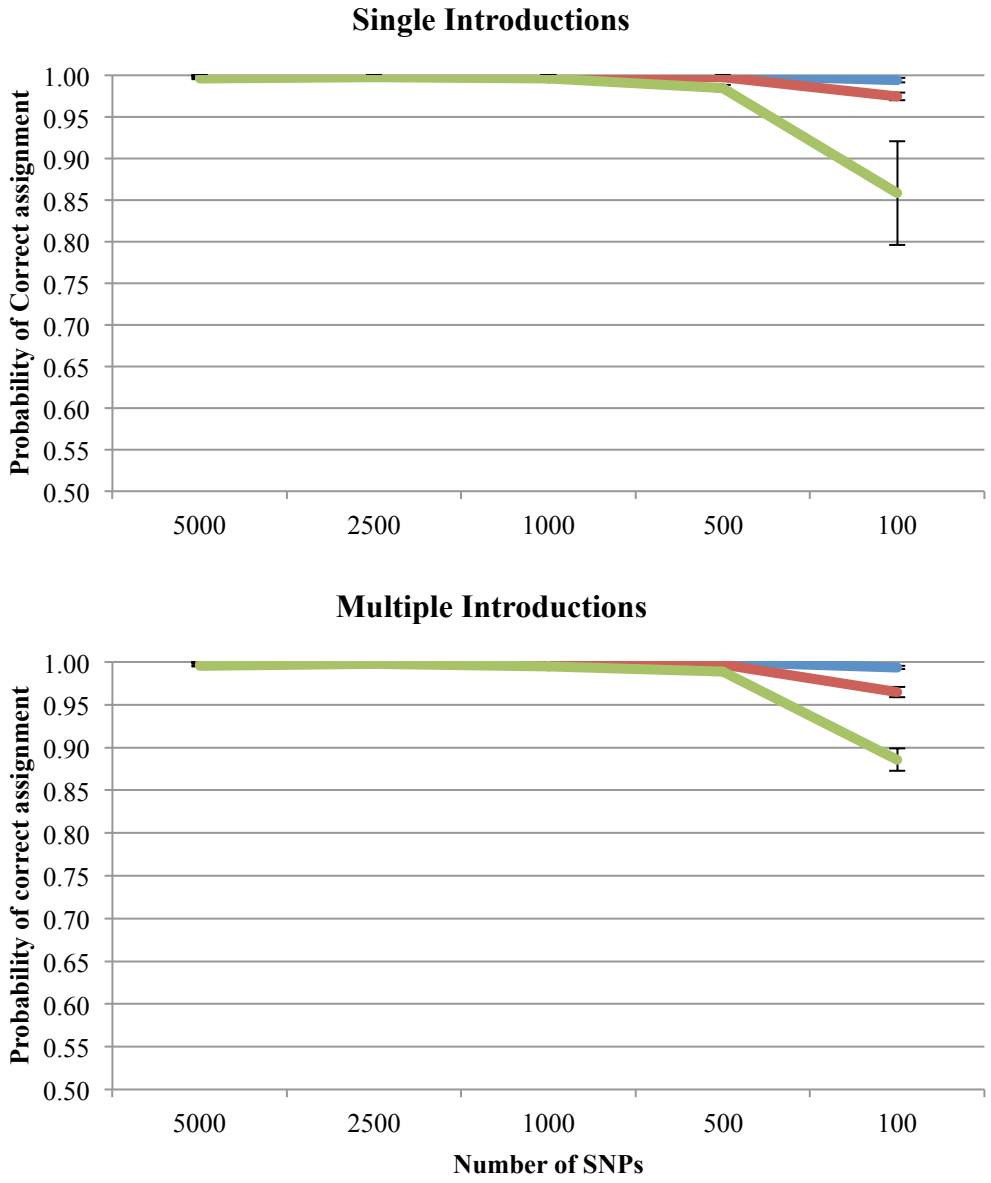

Figure 4.3 – Average probability of correct assignment of the introduced population for scenarios of simulated SNPs without missing data. Upper panel represent the single introduction scenarios with low (blue line), moderate (red line), and high (green line) migration in the native range as depicted in Figure 4.1. The lower panel depicts multiple introductions with the same color scheme for migration parameters. Probability is estimated for the full dataset of 5000 SNPs and a random sample of 2500, 1000, 500, and 100 SNPs with standard error bars.

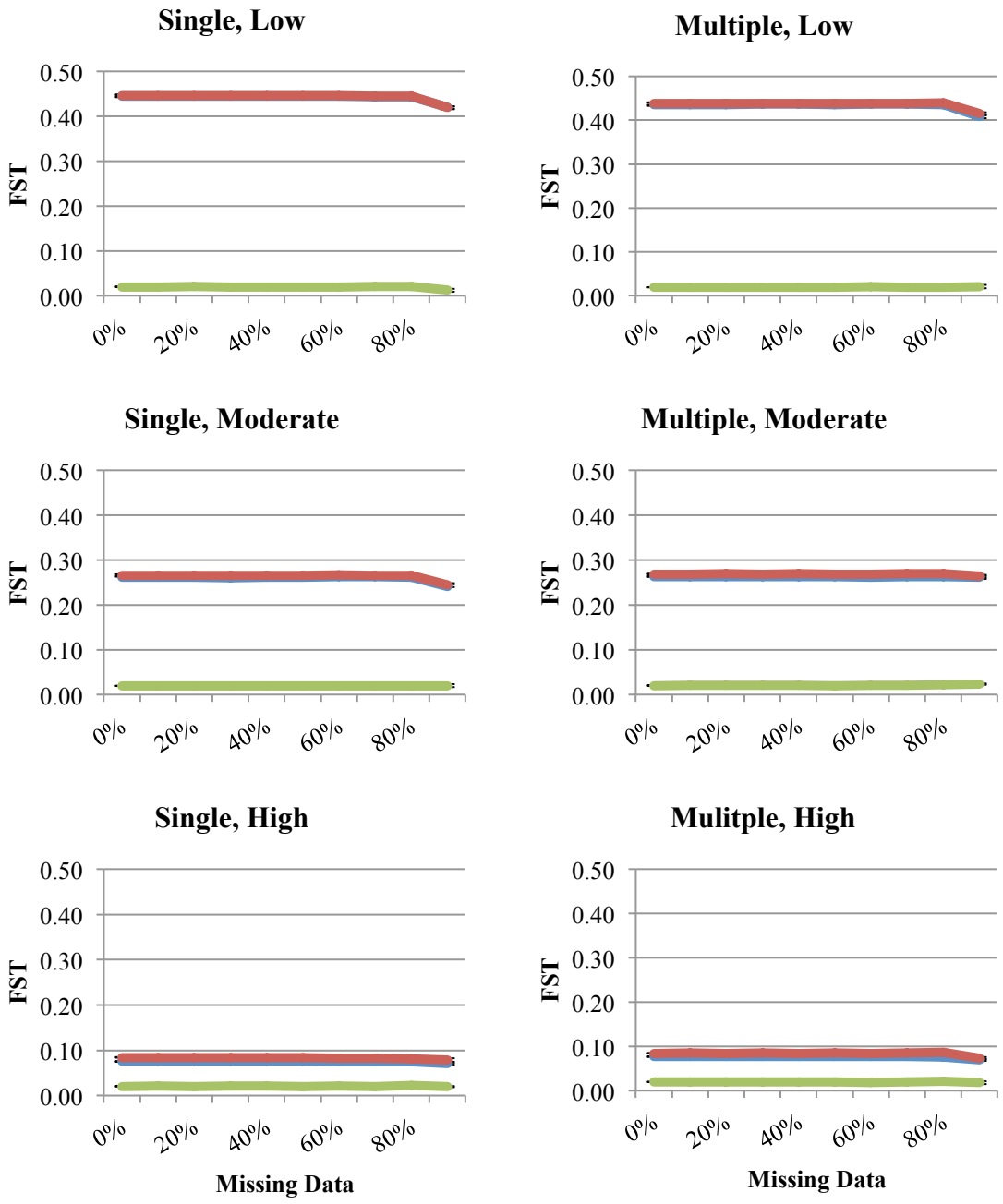

Figure 4.4 – Average pairwise FST for the six scenarios with increasing amounts of missing data in the simulated datasets. S_p v. S_a (blue), S_p v. I (red), and S_a v. I (green).

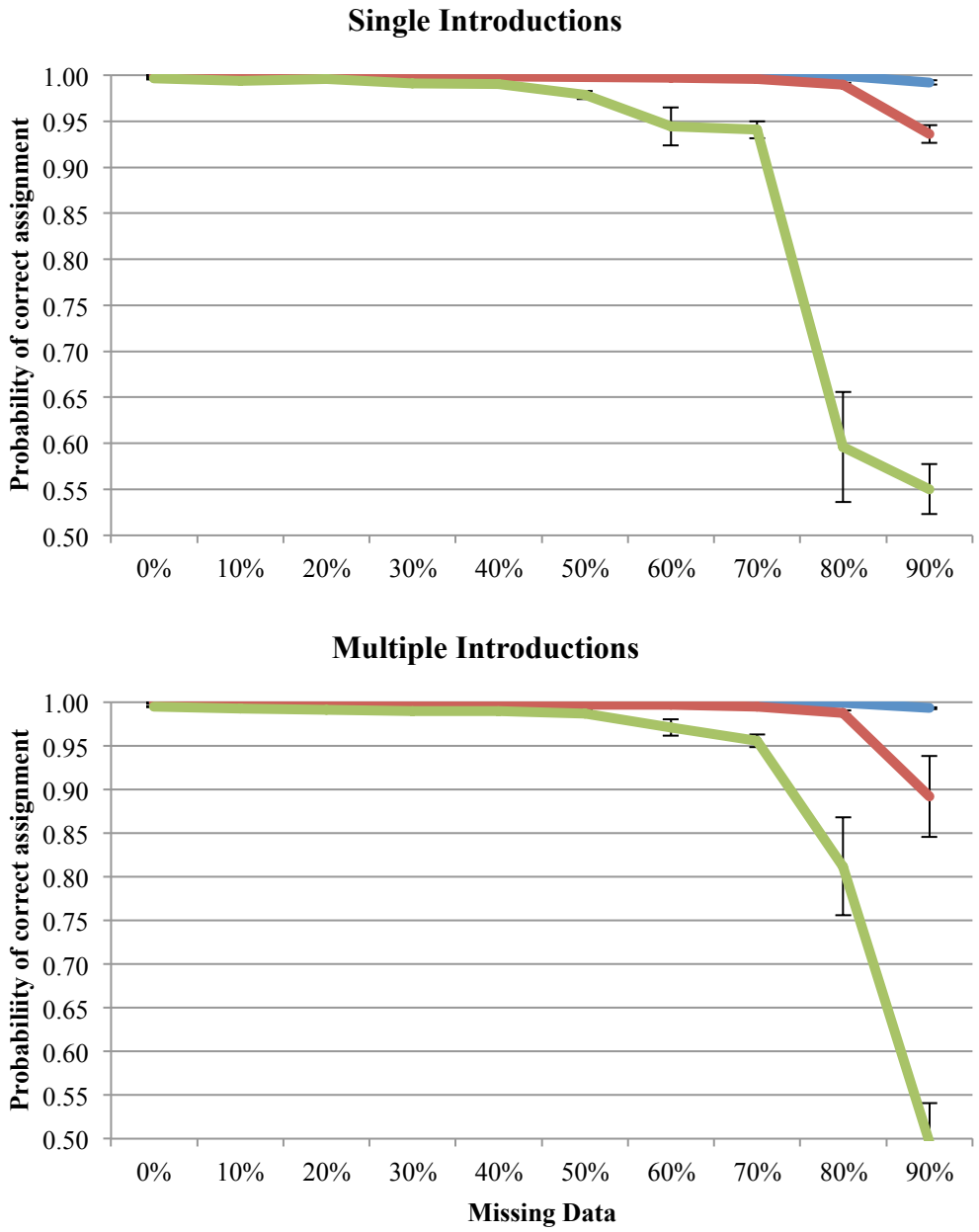

Figure 4.5 – Average probability of correct assignment for all six scenarios of 5000 simulated SNPs with increasing amounts missing data. Upper panel represents the single introductions with low (blue line), moderate (red line), and high (green line) migration in the native range as depicted in Figure 4.1. Lower panel represent the multiple introduction with the same color scheme.

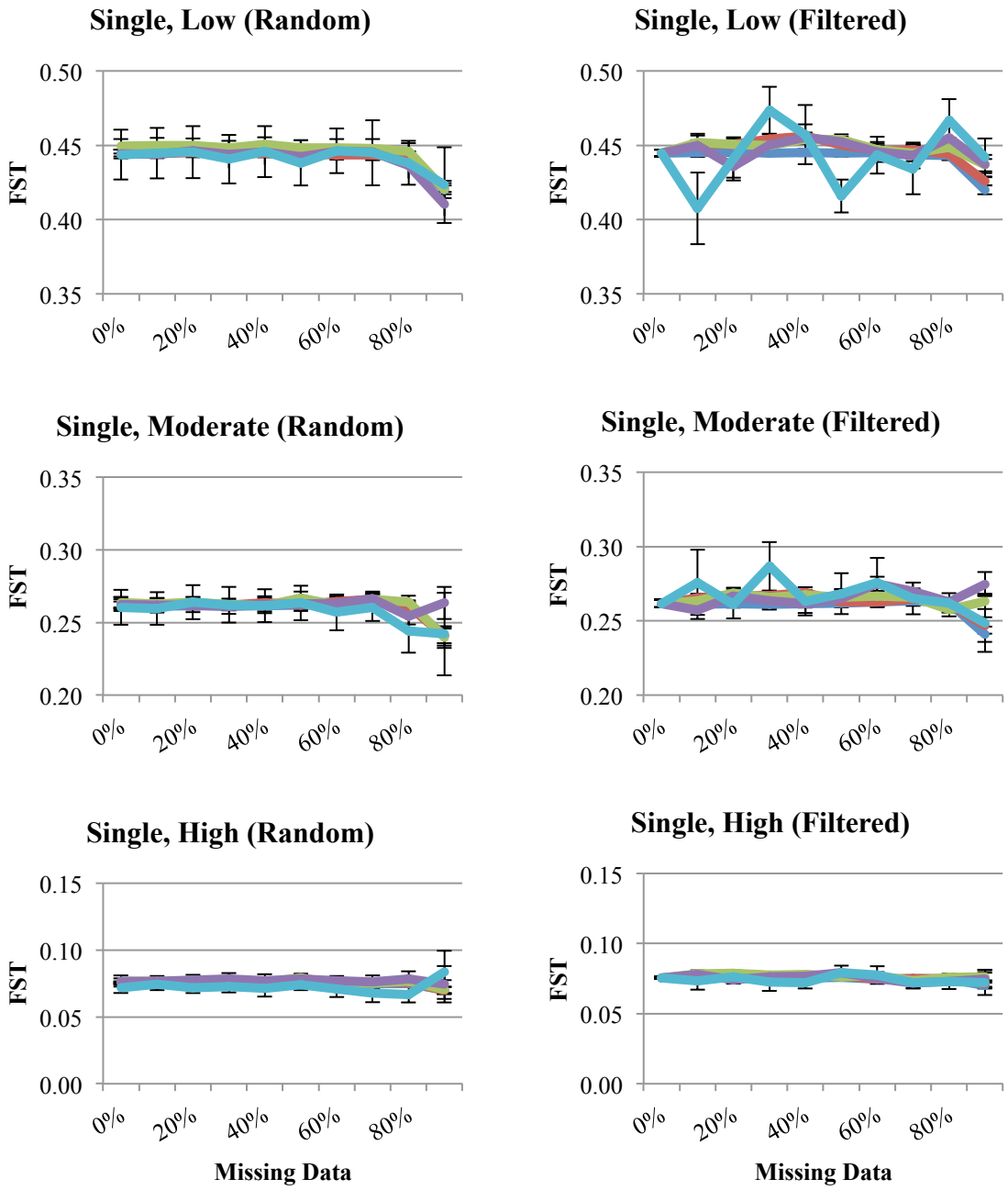

Figure 4.6 – Average F_{ST} values for the simulated SNPs from the two native populations (S_p v. S_a) under single introduction scenarios. On the left are the average values for 5000 (blue lines), 2500 (red lines), 1000 (green lines), 500 (purple lines), and 100 (turquoise lines) SNPs randomly selected with standard error bars. On the right are the average values for a similar number of SNPs filtered in TASSEL.

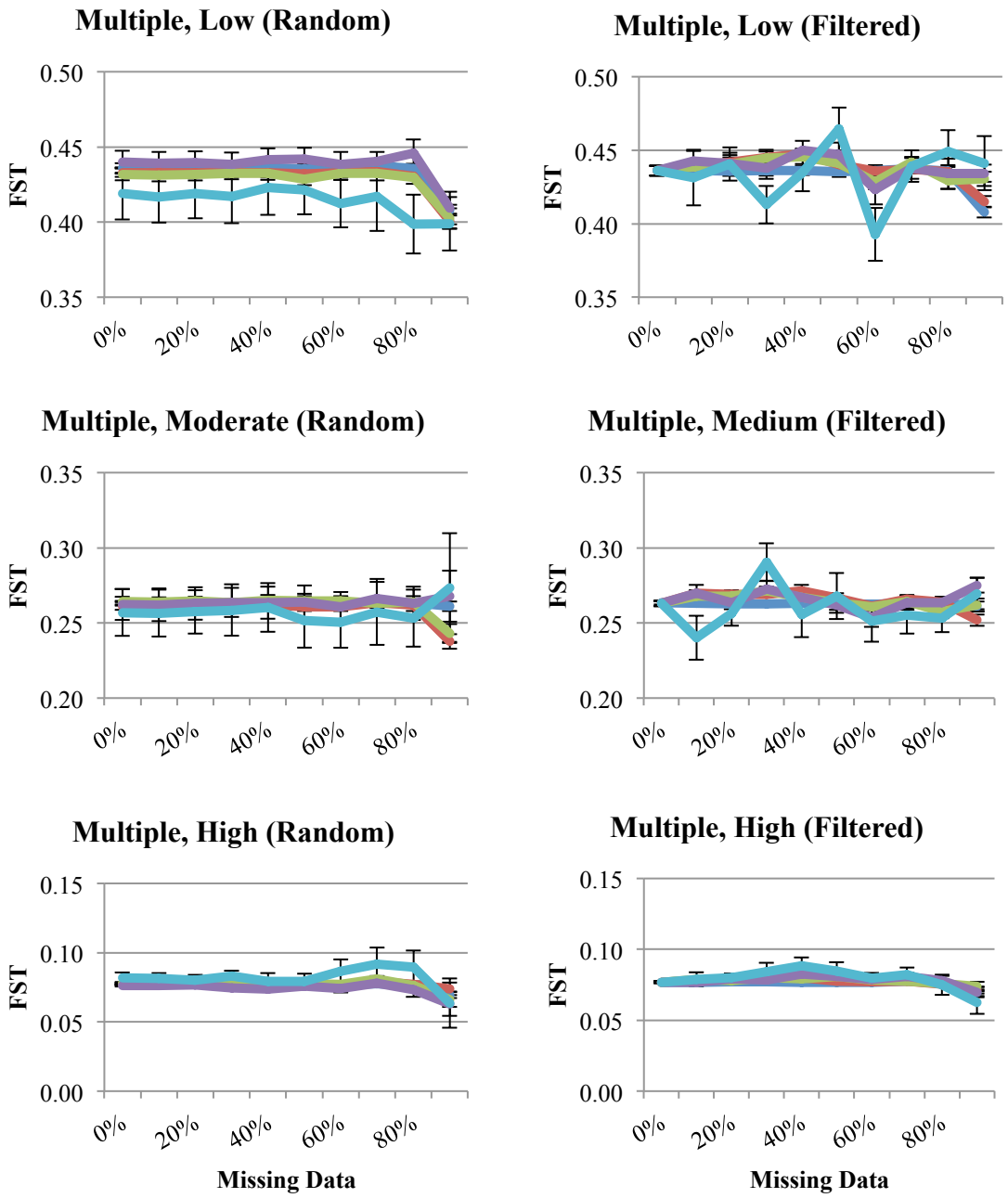

Figure 4.7 – Average F_{ST} values for the simulated SNPs from the two native populations (S_p v. S_a) under multiple introduction scenarios. On the left are the average values for 5000 (blue lines), 2500 (red lines), 1000 (green lines), 500 (purple lines), and 100 (turquoise lines) SNPs randomly selected with standard error bars. On the right are the average values for a similar number of SNPs filtered in TASSEL.

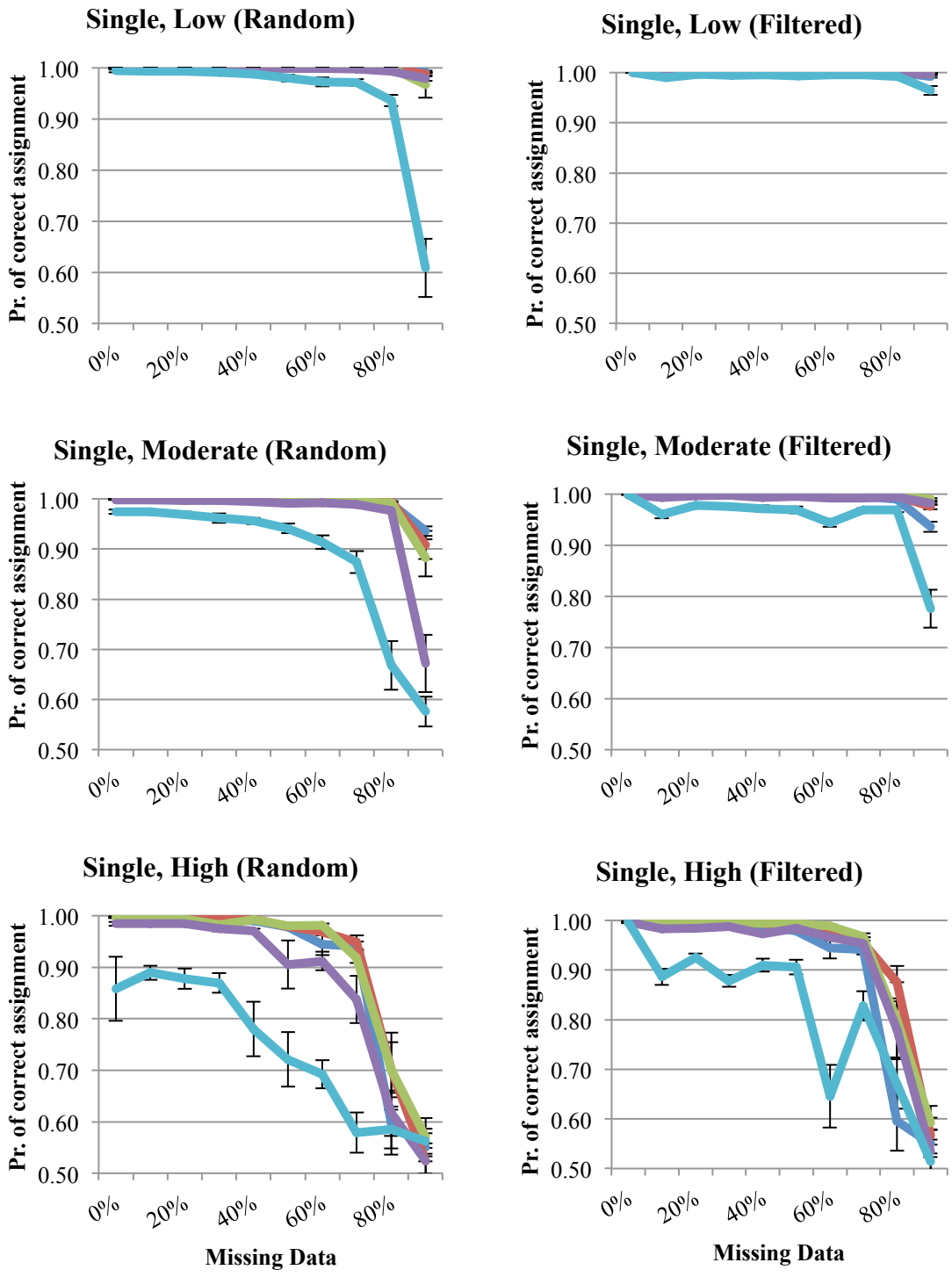

Figure 4.8 – Average probability of correct assignment of the introduced population for the simulated SNPs under single introduction scenarios with increasing amounts of missing data. The left panel are the average values for 5000 (blue lines), 2500 (red lines), 1000 (green lines), 500 (purple lines), and 100 (turquoise lines) SNPs randomly selected with standard error bars. On the right are the average values for a similar number of SNPs filtered in TASSEL.

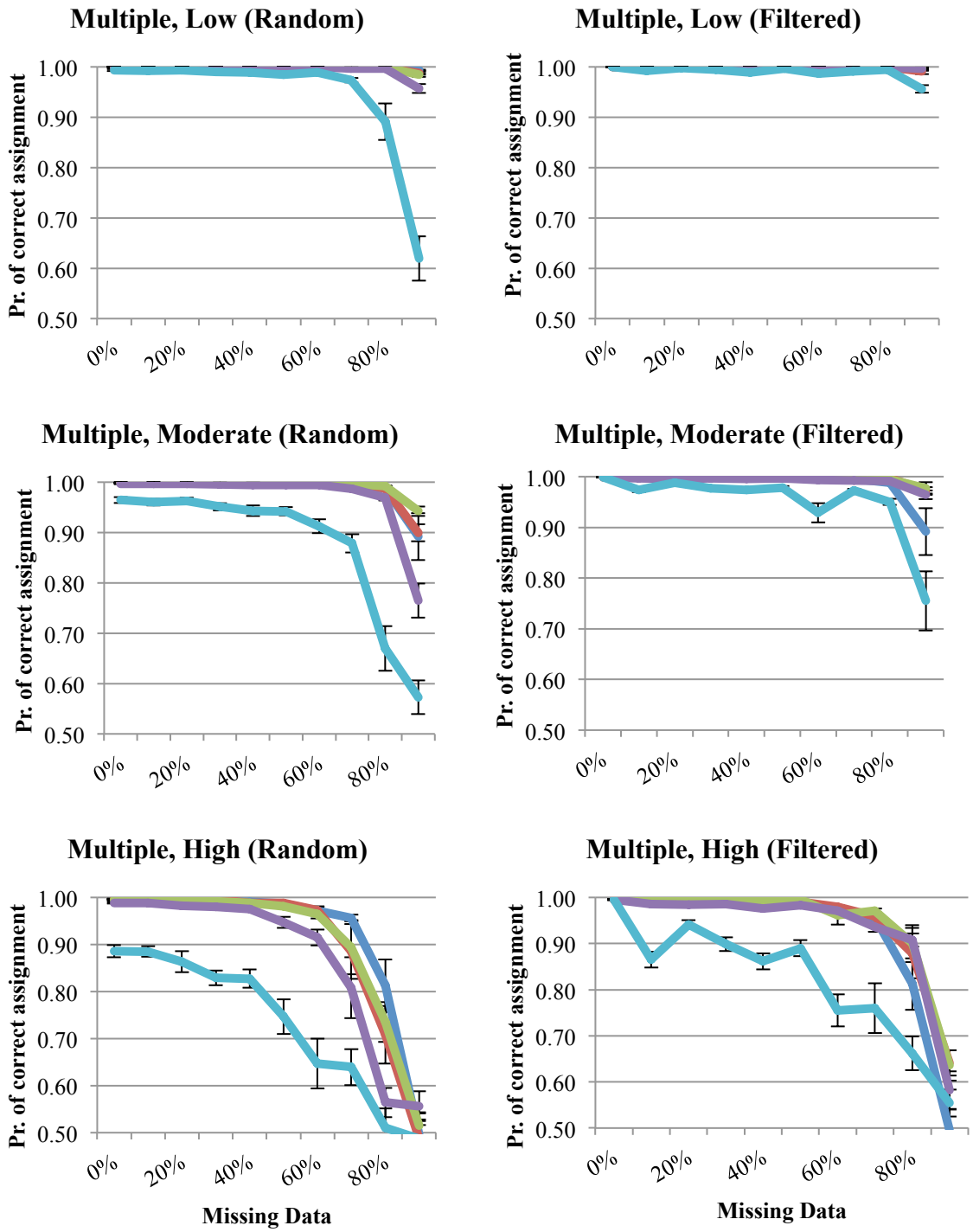

Figure 4.9 – Average probability of correct assignment of the introduced population for the simulated SNPs under multiple introduction scenarios with increasing amounts of missing data. The left panel are the average values for 5000 (blue lines), 2500 (red lines), 1000 (green lines), 500 (purple lines), and 100 (turquoise lines) SNPs randomly selected with standard error bars. On the right are the average values for a similar number of SNPs filtered in TASSEL.

CHAPTER 5: CONCLUSIONS

Biological invasions are a major threat to biodiversity and global change could potentially increase their impact on the environment (Bradley *et al.* 2010; Lodge 1993; Rahel & Olden 2008; Vitousek *et al.* 1996). In order to better prevent and manage invasive species, we must understand their invasion history, which can lead to better management strategies (Sakai *et al.* 2001). In this dissertation, I traced the invasion history of *Gambusia affinis* in Asia using a suite of microsatellite markers, a fragment of the mitochondrial genome, and historical records. I also explored the impact of missing data on large RADseq datasets and their ability to properly assign introduced populations to their correct source using simulated data. The common theme throughout this research is the importance of understanding the genetic variation and population structure of the native range. Patterns from the native range can help identify the source population(s), determine how genetic diversity has changed, and develop hypotheses on introduction routes taken. I demonstrated this by examining sampling localities from the native range of *G. affinis* throughout the southeastern United States and from the introduced range including Hawaii, Taiwan, the Philippines, Japan, and China. I further simulated large RADseq datasets with increasing levels of missing data under six invasion scenarios that included native and introduced populations.

In chapter 2, I sequenced a fragment of the mitochondrial gene cytochrome *b* and genotyped 18 microsatellites for 42 localities spanning the distribution of *G. affinis* and *G. holbrooki* throughout the southeastern United States. I tested three specific breaks that were previously described as barriers for gene flow (Soltis *et al.* 2006; Wooten *et al.* 1988). The species boundary between the two species show little admixture, suggesting that while they may

occur in sympatry there appears to be very little hybridization going on in natural populations. I show that the Savannah River is not a strong barrier to gene flow isolating localities north and south of the river in *G. holbrooki*. Localities throughout South Carolina and parts of North Carolina showed significant admixture with localities south of the Savannah River indicating that this region is an area of admixture between the two groups. The Mississippi River also does not serve as a barrier to gene flow within *G. affinis*. Instead, localities within the Mississippi River system all cluster together and are actually distinct from localities collected farther west in Texas and Oklahoma. One challenge not discussed previously of this study is that mosquitofish are transported by humans within the native range as well, creating the potential for population structure to be broken down and obscure patterns. For example, the lack of a clear East-West split at the Mississippi River could have two likely explanations. First, mosquitofish within the drainage system have been able to move around historically due to their high population density and colonization ability (Pyke 2008). Second, mosquitofish introductions within the native range could have broken down population structure around the Mississippi River within the last century. However, distinguishing between these two scenarios was not the scope of this study but is worth considering as a mechanism for the current population structure.

In chapter 3, I conducted a search for historical documentation of mosquitofish introductions to Asia and also gathered genetic data (as described in Chapter 2) for 20 introduced localities from Hawaii, Taiwan, the Philippines, Japan, and China in an attempt to reconstruct an accurate invasion history. I found several records detailing the introduction of mosquitofish from Seabrook, Texas to Hawaii and from Hawaii to Taiwan and the Philippines. Mosquitofish were taken from Taiwan to Japan, while China received mosquitofish from both Taiwan and the Philippines. I found a mitochondrial haplotype that occurred in ~72% of introduced individuals

sequenced occurred in only one native locality, the putative source population near Seabrook, Texas. Furthermore, 19 introduced localities were assigned to that same native locality using all 18 microsatellite markers. While genetic diversity was reduced across the introduce range, very little evidence for a genetic bottleneck was detected. These results corroborate the historical record and suggest that mosquitofish introductions were carried out with large numbers of individuals throughout Asia.

Chapter 3 provides valuable results for management implications and future research on the evolution of invasive species. Mosquitofish are bred in large numbers and supplied as mosquito control agents (Ghosh & Dash 2007). However, if we are to reduce their impact on the environment one strategy should be to educate the public regarding the impact of mosquitofish. Outreach efforts that help the public understand the detrimental impact of mosquitofish could curb their continued spread. Furthermore, agencies responsible for controlling mosquito-borne disease should also be included in outreach efforts, especially if a native species can be substituted for mosquitofish. Stopping future introductions and slowing their spread will help, but further action has to be taken. I identified a specific geographic location in Texas that gave rise to most, if not all, Asian mosquitofish. Given that mosquitofish are widely distributed, the search for a biological control agent in that source population could provide an efficient method of controlling and decreasing mosquitofish populations in Asia. Another theoretical approach that has been modeled in mosquitofish is the use of Trojan sex chromosome individuals that when introduced only produce male offspring that can hypothetically lead to the collapse of the population (Senior *et al.* 2013). Thus, with the identity of the source population for Asia there is potential for strategies to control and reduce the impact of mosquitofish.

In a broader context, by tracing the invasion of mosquitofish further studies can be conducted on the evolution of invasiveness. For example, life history traits are often targets of natural selection and the introduced range may exhibit life history traits different from the native range (Barrett *et al.* 2008; Gonçalves da Silva *et al.* 2010). Behavioral traits are increasingly being considered as components that aid invasion success (Light 2005; Pintor *et al.* 2009; Rehage & Sih 2004). By knowing the source population, we can compare traits between the native source and the introduced range. Furthermore, we can compare the native source with the rest of the native range to look for any local adaptation that may be unique to the source.

In chapter 4, I simulated RADseq datasets for six invasion scenarios and simulated increasing amounts of missing data. I calculated pairwise F_{ST} for all of the datasets and performed assignment tests for introduced populations. All F_{ST} estimates were consistent across all treatments of missing data, all scenarios, and for all numbers of loci sampled. Assignment tests were robust for scenarios with low and moderate migration up to 90% missing data. For scenarios with high migration probabilities of correct assignment began declining after 50% missing data. Filtering of the data improved results for the assignment tests significantly. I found that the simulation of multiple and single introduction had very little influence on the results. The results obtained provide helpful information for researchers making decisions regarding the generation and analysis of large RADseq SNP datasets. These large datasets will become increasingly common over the next several years and understanding how missing data impacts the tracing of an invasion or other population genetic analyses will be important.

In conclusion, the study of biological invasions gives us the opportunity to address fundamental questions in ecology and evolutionary biology, while also addressing an important issue threatening biodiversity. The native and introduced ranges can often present challenges, in

resources and time, to sampling and conducting experiments. However, developing collaboration with colleagues can help alleviate this challenge. I would note that this is the major goal of the funding which supported the entirety of this research and made the extensive sampling in Asia possible. Thus, the use of native and introduced populations combined with genome-wide sequencing technology in studies on invasive species will provide great hope in ultimately preserving biological diversity around the world.

References

Barrett SCH, Colautti RI, Eckert CG (2008) Plant reproductive systems and evolution during biological invasion. *Molecular Ecology* **17**, 373-383.

Bradley Ba, Blumenthal DM, Wilcove DS, Ziska LH (2010) Predicting plant invasions in an era of global change. *Trends in Ecology & Evolution* **25**, 310-318.

Ghosh S, Dash A (2007) Larvivorous fish against malaria vectors: a new outlook. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **101**, 1063-1064.

Gonçalves da Silva A, Eberhard JR, Wright TF*, et al.* (2010) Genetic evidence for high propagule pressure and long-distance dispersal in monk parakeet (Myiopsitta monachus) invasive populations. *Molecular Ecology* **19**, 3336-3350.

Light T (2005) Behavioral effects of invaders: alien crayfish and native sculpin in a California stream. *Biological Invasions* **7**, 353-367.

Lodge D (1993) Biological Invasions: Lessons for Ecology. *Trends in Ecology & Evolution* **8**, 133-137.

Pintor L, Sih A, Kerby J (2009) Behavioral correlations provide a mechanism for explaining high invader densities and increased impacts on native prey. *Ecology* **90**, 581-587.

Pyke GH (2008) Plague Minnow or Mosquito Fish? A Review of the Biology and Impacts of Introduced *Gambusia* Species. *Annual Review of Ecology, Evolution, and Systematics* **39**, 171-191.

Rahel FJ, Olden JD (2008) Assessing the effects of climate change on aquatic invasive species. *Conservation Biology* **22**, 521-533.

Rehage J, Sih A (2004) Dispersal behavior, boldness, and the link to invasiveness: a comparison of four Gambusia species. *Biological Invasions*, 379-391.

Sakai A, Allendorf F, Holt J (2001) The population biology of invasive species. *Annual Review of Ecology and Systematics* **32**, 305-332.

Senior AM, Krkosek M, Nakagawa S (2013) The practicality of Trojan sex chromosomes as a biological control: an agent based model of two highly invasive *Gambusia* species. *Biological Invasions* **15**, 1765-1782.

Soltis DED, Morris ABA, McLachlan JS, Manos PS, Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology* **15**, 4261-4293.

Vitousek PM, D'Antonio CM, Loope LL, Westbrooks R (1996) Biological invasions as global environmental change. *American Scientist* **84**, 468-478.

Wooten M, Scribner K, Smith M (1988) Genetic Variability and Systematics of *Gambusia* in the Southeastern United States. *Copeia* **1988**, 283-289.